ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

School of Science
Master's degree in Mathematics

A Berry-Esseen Theorem for Wide Quantum Neural Networks

Degree Thesis in Quantum Machine Learning

Supervisor:
Prof.
Giacomo De Palma

Candidate:
Alanna Benn

Academic year 2023/2024

In loving memory of Brian Benn, Marilyn Manson, and Audrey Benn.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, professor Giacomo De Palma, for the opportunity to work on this topic, as well as for his mentorship, expertise, detailed instruction, and invaluable feedback. I would also like to thank Anderson Melchor Hernandez for his assistance and availability throughout the course of this research. To the Università di Bologna, and in particular the mathematics department, thank you for the privilege of studying at such a hallowed institution.

I am grateful to every one of my Italian friends and classmates for their endless patience, encouragement, and assistance as I continue to practice and improve my Italian. I am grateful also to my fellow international students for their friendship and camaraderie.

To all those in Canada that crossed an ocean to visit, your presence means the world to me. And to my family in Germany, thank you for the warm welcome each holiday season these past years.

I would especially like to thank my parents, Allan Benn and Mafalda Di Iorio, for their endless love and support in all my endeavors, academic and otherwise.

And finally, thank you Marco for embarking on this journey with me and for being by my side every step of the way.

Abstract

Quantum neural networks are the quantum counterpart of deep neural networks and generate model functions given by the expectation value of a quantum observable measured on the state generated by a parametric quantum circuit. Parametric quantum circuits are made by the composition of elementary parametric quantum operations (gates) and are considered prime candidates for practical applications of quantum computing with the noisy intermediate-scale quantum devices that will be available in the forthcoming years. Quantum neural networks have wide applications both in machine learning problems, such as supervised learning, and in optimization problems. The recent work [Girardi et al., arXiv:2402.08726] has proven that the law of the model function generated by an untrained quantum neural network with random parameters converges in distribution to a Gaussian process in the limit of infinite width of the circuit. In this thesis we establish a quantitative version of this result. We consider randomly initialized quantum neural networks of finite width and a single input, and we establish an upper bound on the Kolmogorov distance between the law of the random variable generated by the network and the Gaussian law with the same mean and variance. Our proof is based on the method of cumulants to derive an upper bound on the absolute difference between the characteristic functions of the two random variables.

Contents

Chapter 1

Introduction

The fields of quantum computing and artificial intelligence (AI) are rapidly evolving [1, 2, 3]. On their own, each has the potential to revolutionize entire industries. At their intersection lies the burgeoning field of quantum machine learning.

Classical computers remain inadequate for simulating many-body physical systems due to the exponential space and time requirements of recording their states and dynamics. In a lecture given in 1981, the contents of which were published in [4], Richard Feynman proposed to build a computer capable of simulating quantum mechanical systems using resources, in terms of both time and space, proportional to the size of the system. This computer, he posited, must itself be quantum mechanical. In 1985, David Deutsch gave rigor to this idea by introducing a theoretical universal model of quantum computation [5]. In 1994 Peter Shor introduced the ground-breaking quantum algorithms for efficient integer factorization and for computing discrete logarithms [6]. This was a massive revelation since the classical hardness of integer factorization and the discrete logarithm problem form the basis of the widely used RSA cryptographic protocol. Such algorithms gave birth to the field of quantum computing and provided motivation to search for further quantum algorithms and to physically build a quantum computer.

Since then, quantum algorithms have been proposed to search an unstructured database [7], solve systems of linear equations [8], and more, bolstering the belief that quantum computing is more powerful than classical computing. In parallel, many technological advances were made in the quest to physically build a quantum computer. In 2001 Shor's algorithm was used to factor the number 15 using a quantum processor with 7 qubits [9]. Nowadays, companies such as IBM, Google, Microsoft and Amazon have unveiled quantum processors with a number of qubits ranging from tens to a thousand and offer cloud-based quantum computing platforms. However, despite their proliferation, quantum computers have not yet been able to solve a computational problem of practical relevance faster than any existing classical computer.

AI refers to the ability of programmable machines to exhibit human-like intelligence. Image recognition, game-playing agents, and natural language processing are all examples of this. Machine learning is a sub-field of AI concerned with algorithms and techniques that allow machines to detect and "learn" patterns in large data sets in order to make predictions or decisions without explicitly being programmed to do so. Deep learning is a further sub-field of machine learning concerned with a class of model functions called deep

neural networks.

Artificial neural networks are mathematical models designed to mimic the way in which biological neurons store and transmit information. They consist of a network of interconnected nodes, called artificial neurons, which are typically aggregated into input, hidden, and output layers, with deep neural networks having many hidden layers. They are parametrized functions in which information flows from the input layer to the output layer via weighted connections and non-linearities in the hidden layers. By tweaking these weights – a process called training – the model can be made to fit a corpus of data that is representative of the patterns the model aims to learn. The ability of the model to learn the overall pattern rather than memorizing specific examples is referred to as the generalization power of the model.

Deep neural networks have achieved extraordinary performances on several machine-learning tasks. An example is the ImageNet Large Scale Visual Recognition Challenge, a now discontinued annual event in which software programs compete to classify images into categories. During the 2012 iteration of the competition a deep neural network named AlexNet achieved a top-5 error rate [1] of 15.3%, which was more than 10% lower than the runner-up, and a considerable improvement over the previous state-of-the-art [10]. For comparison, a human expert annotator later classified 1500 images from the same data set with a top-5 error rate of 5.1% after practicing on 500 images, and a second annotator classified 258 images with an error rate of 12% after practicing on 100 images [11]. In 2016 a deep neural network called AlphaGo, which was developed by DeepMind to play the board game Go, defeated the reigning human European champion in 5 out of a series of 5 games [12]. This was considered a remarkable feat due to the complexity of Go and, in particular, its immense number (approximately 2.1×10^{170} [13]) of legal game states. In 2022 OpenAI released ChatGPT, an interactive large language model designed for conversational use [14]. ChatGPT quickly gained attention due to its general purpose utility and ability to respond in a human-like way.

AlexNet, AlphaGo, and ChatGPT are all products of deep learning. These milestones illustrate the extraordinary generalization power of deep neural networks and the success they have achieved across a variety of AI applications. In particular, they are known to perform exceedingly well when modeling high-dimensional data. Their empirical success has motivated the development of a deeper understanding of their theoretical underpinnings and some major breakthroughs have been made in this regard. In [15] the authors establish that feedforward networks can approximate any continuous function on a compact subset of the real numbers to any desired precision given enough hidden neurons. In [16] the equivalence between a Gaussian process and a neural network with one infinitely wide hidden layer and random weights is shown. This result was generalized to the case of deep neural networks in [17].

The motivation for studying models with random weights can be summarized as follows. The optimal network weights are obtained by minimizing a loss function which quantifies the total error in the predictions of the model on the training data. This optimization is carried out via an iterative procedure, such as gradient descent, in which the weights are initially randomly sampled and then updated such that the loss function decreases at each

[1] Top-5 error rate refers to the percentage of images in which the program did not identify the correct label among the five most likely.

time step. Understanding a model with random weights therefore provides insight into the dynamics of the optimization.

In particular, [18] proves that in the limit of infinite width the variation of any single parameter during training by gradient descent is negligible, yet, since there are infinitely many parameters, the combined variation results in a perceptible difference in the generated function. Furthermore, the probability distribution of the generated function converges in distribution to a Gaussian process that perfectly fits the training data exponentially fast in the training time, and whose mean and covariance can be analytically evaluated given said data and the network architecture. This revelation is central to explaining why overparametrized neural networks (i.e. neural networks with more parameters than training examples) are always trainable without suffering from bad local minima and are able to generalize well without suffering from overfitting.

The desire to process ever-larger data sets and model increasingly complex patterns has naturally led many to wonder whether quantum computing can be leveraged to improve machine learning models. Quantum-enhanced machine learning is a sub-discipline of quantum machine learning in which quantum computers assist in the learning of classical data [19]. One approach to this task is to define a model function using a parametric quantum circuit to prepare a system of qubits in a state that encodes both data input and trainable parameters. Such a circuit is made by the composition of elementary parametric quantum gates. The model function is given by the expected value of a measurement, described by some observable, on the state of the qubits prepared by the circuit. These models are known as quantum neural networks. Besides their potential utility in machine learning, they have gained attention for their perceived suitability to near-term quantum devices [20]. Outside of machine learning variational circuits can be applied to a plethora of applications such as quantum chemistry and simulation, solving systems of linear equations, combinatorial optimization, and more [21].

Like in the classical setting, quantum neural networks are trained by minimizing a cost function that is dependent on the circuit architecture. It has been observed that for many architectures the function exhibits vanishing gradients far from its minimum, impeding gradient-based optimization techniques and thus preventing successful training of the network. This phenomenon is known as barren-plateaus [22, 23] and is one of the main challenges facing parametric quantum circuits. It has been speculated that any circuit which provably avoids barren-plateaus can be efficiently simulated with a classical computer [24].

Similar to the classical setting, various works have aimed to develop a more rigorous mathematical framework regarding quantum neural networks. In [25] the author extends to the quantum setting the result that infinitely wide random neural networks are equivalent to Gaussian processes. More specifically, it is proved that under certain hypotheses the function defined by a quantum neural network converges in distribution to a Gaussian process as the number of qubits of the model tends to infinity. It is further proved that the function defined by the trained quantum neural network converges in distribution to a Gaussian process which perfectly fits its training set exponentially fast in the training time. Essential to these results holding is the use of a circuit that does not induce barren-plateaus, and [25] provides examples of such circuits which are not classically simulable with brute-force algorithms. These results rigorously prove that wide quantum neural networks are efficiently trainable and do not suffer from bad local minima of the cost function.

In practice, of course, quantum neural networks have finitely many qubits and therefore it is desirable to know how fast this convergence takes place or, in other words, how well the function generated by a finite model is approximated by a Gaussian process. This is achieved in [26], in which the authors find an upper bound on the Wasserstein distance of order 1 between the multivariate normal law and the law of the function generated by a given finite quantum neural network.

1.1 Our contribution

In this thesis we provide an alternative quantitative version of the results of [25]: we consider untrained quantum neural networks with a single input and random parameters, and we establish an upper bound on the Kolmogorov distance between the law of the output of the network and the Gaussian law with the same mean and variance. If we consider a sequence of randomly initialized quantum neural networks with increasingly many qubits, then our bound provides the rate of convergence of the sequence of their outputs to a Gaussian random variable. While in [25, 26] the measured observable is constrained to be the sum of single-qubit observables, our result is valid for any measurement.

To prove this result we first use the method of cumulants to derive an upper bound on the absolute difference between the characteristic functions of the two random variables in a neighborhood of zero. This involves expressing the model function as a sum of random variables and quantifying the maximum degree of dependency any one variable has on the others. We then translate this into a bound on the absolute difference between the cumulative distribution functions of the two variables and then solve an optimization problem to find the tightest bound possible.

The thesis is outlined as follows. In chapter 2 we begin with a brief overview of the postulates of quantum mechanics followed by an introduction to the circuit model of quantum computation. We then discuss the potential benefit afforded by quantum computers, the main challenges they face, and the pursuit of a quantum advantage. Finally, we introduce variational quantum algorithms (VQAs), which are a class of algorithms that are based on variational circuits.

In chapter 3 we fix our assumptions on the architecture of the variational circuit and we rigorously define the random variable generated by its output. We also define a number of quantities, interpreted as properties of the random variable, that are determined by the circuit architecture and the observable describing the measurement of the qubits. We then discuss the classical simulability of the circuit and the potential to obtain a quantum advantage.

The original work of this thesis is contained in chapter 4. We begin this chapter with an introduction to convergence in distribution, distances between probability measures, and the Berry-Esseen theorem [2] on which our result is based. We then state and prove Theorem 5, our main result: an upper bound on the Kolmogorov distance between the law of the output of a randomly initialized variational quantum circuit and the Gaussian law with the same mean and variance. This is followed by a discussion of the asymptotic behavior of the bound

[2]The Berry-Esseen theorem quantifies the rate of convergence that takes place in the Central Limit Theorem.

for certain architecture choices. To conclude this chapter we compare our bound with the one given in [26].

In chapter 5 we give our closing remarks including a summary of our findings, their implications and limitations, and avenues for further work.

Chapter 2

Quantum computing and variational quantum algorithms

Quantum computing is a model of computation that leverages the principles of quantum mechanics. As a theoretical model it is believed to be more powerful than its classical counterpart. Indeed, quantum computers have the potential to efficiently perform tasks, such as factoring large composite integers or simulating quantum systems, that are intractable to modern classical computers. However, today's quantum computers suffer from significant drawbacks such as noise, decoherence, and a lack of error correction that have prevented them from realizing this potential.

VQAs are hybrid quantum-classical algorithms that have been envisaged for a wide variety of applications. They have garnered much attention due to their perceived suitability to the noisy intermediate-scale quantum (NISQ) devices that are expected to be available in the near future [27]. In fact, they have emerged as a leading contender in the pursuit of a quantum advantage which is demonstrated when a quantum computer executes an algorithm that solves a practical problem more efficiently than any classical computer is capable of.

This chapter serves as an introduction to quantum computing and VQAs. We begin in section 2.1 with an overview of the mathematical foundations of quantum mechanics. In 2.2 we provide an introduction to the theory of quantum computing, highlighting the properties that set it apart from classical computing and providing an intuition as to why it is believed to be superior. We will then discuss the challenges currently facing the field of quantum computing. These sections will lay the groundwork for understanding both the motivation behind, and the concept of, a VQA, which we present in section 2.3 along with two specific examples of its applications.

2.1 Foundation of quantum mechanics

There are four key postulates describing quantum systems, their states, how they evolve, and how they are measured. These postulates are presented in the language of linear algebra and rely heavily on the theory of complex Hilbert spaces and linear operators. For a more in-depth introduction to this topic we refer the reader to [28, Chapter 2]. This text provides, along with the necessary mathematical prerequisites, a comprehensive introduction to quantum

mechanics with an emphasis on computational applications.

2.1.1 States

The first postulate deals with the mathematical representation of a physical system and its state.

Postulate 1. Associated to every isolated physical system is a Hilbert space $\mathcal{H}$ called the *state space* of the system. The state of the system is described by a linear operator ρ acting on $\mathcal{H}$ that is positive semidefinite and has trace equal to 1.

$$\rho \in \mathcal{L}(\mathcal{H}), \quad \rho \geq 0, \quad \mathrm{Tr}\rho = 1. \tag{2.1}$$

ρ is called a *density operator*.

Let $|\psi\rangle \in \mathcal{H}$ be a unit vector and suppose that ρ is the orthogonal projector onto the subspace of $\mathcal{H}$ spanned by $|\psi\rangle$:

$$\rho = |\psi\rangle\langle\psi|, \quad \langle\psi|\psi\rangle = 1. \tag{2.2}$$

ρ is a valid quantum state. Such a state is called *pure* and it is completely described by $|\psi\rangle$ which we call a *state vector*. If a quantum state is not pure it is called *mixed*.

Let $|\psi\rangle, |\phi\rangle \in \mathcal{H}$. $|\psi\rangle$ and $|\phi\rangle$ describe the same state if and only if they differ by a *global phase*:

$$|\psi\rangle = \lambda|\phi\rangle, \quad \lambda \in \mathbb{C}, \quad |\lambda| = 1. \tag{2.3}$$

In this case, we say that $|\psi\rangle$ and $|\phi\rangle$ are *proportional*.

2.1.2 Evolution

The second postulate specifies how undisturbed quantum systems evolve over time.

Postulate 2. The evolution of an isolated quantum system is described by a unitary operator acting on the state space of the system. Denote by ρ_t the state of the system at time t. For all $t_1, t_2 \in \mathbb{R}$ with $t_1 \leq t_2$

$$\rho_{t_2} = U_{t_1 \to t_2} \rho_{t_1} U_{t_1 \to t_2}^{\dagger} \tag{2.4}$$

for some unitary operator $U_{t_1 \to t_2} \in \mathcal{L}(\mathcal{H})$.

Pure states evolve into pure states. If $\rho_{t_1} = |\psi_{t_1}\rangle\langle\psi_{t_1}|$, then $\rho_{t_2} = |\psi_{t_2}\rangle\langle\psi_{t_2}|$ where $|\psi_{t_2}\rangle = U_{t_1 \to t_2}|\psi_{t_1}\rangle$. An alternative formulation of Postulate 2 is the following.

Postulate 2. Associated to every isolated quantum system is a Hermitian operator $H \in \mathcal{L}(\mathcal{H})$ called a *Hamiltonian*. The Hamiltonian governs the dynamics of the system via the equation

$$i\hbar \frac{d\rho_t}{dt} = [H, \rho_t] \tag{2.5}$$

where $\hbar$ is a constant known as *reduced Planck's constant*. If ρ_t describes a pure state with state vector $|\psi_t\rangle$, then the evolution is given by *Schrödinger's equation*:

$$i\hbar \frac{d|\psi_t\rangle}{dt} = H|\psi_t\rangle. \tag{2.6}$$

The connection between the two formulations of Postulate 2 is this. Write $|\psi_t\rangle$ as the evolution of the state $|\psi_{t_0}\rangle$.

$$|\psi_t\rangle = U_{t_0\to t}|\psi_{t_0}\rangle. \tag{2.7}$$

Then according to Schrödinger's equation the operator $U_{t_0\to t}$, as a function of t, must satisfy the differential equation

$$i\hbar\frac{d}{dt}U_{t_0\to t} = HU_{t_0\to t}. \tag{2.8}$$

If H is time invariant, then the solution is given by

$$U_{t_0\to t} = \exp\left[\frac{-iH(t-t_0)}{\hbar}\right]. \tag{2.9}$$

Any operator of this form with H Hermitian is unitary. Furthermore, for any unitary $U \in \mathcal{L}(\mathcal{H})$ it is possible to find $K \in \mathcal{L}(\mathcal{H})$ Hermitian such that $U = \exp(iK)$.

2.1.3 Measurement

It is impossible to know the state of an isolated quantum system without first measuring it. Furthermore, the act of measuring the system disturbs it. The third postulate deals with measurements, their possible outcomes, and their effect on the system.

Postulate 3. A *projective measurement* with outcome set A is described by a collection of measurement operators

$$\{P_a : a \in A\} \subset \mathcal{L}(\mathcal{H}) \quad \text{such that} \quad \mathcal{H} = \bigoplus_{a\in A}^{\perp} \text{supp}P_a. \tag{2.10}$$

Each $a \in A$ refers to a possible outcome of the measurement. If, before the measurement, the state of the system is ρ, then the probability of the measurement described by $\{P_a : a \in A\}$ resulting in outcome a is given by

$$\mathbb{P}(a|\rho) = \text{Tr}[P_a\rho]. \tag{2.11}$$

If the outcome of the measurement is a, the state of the system after the measurement becomes

$$\rho'_a = \frac{P_a\rho P_a}{\text{Tr}[P_a\rho]}. \tag{2.12}$$

In this case we say that ρ *collapses* to the state ρ'_a.

Remark 1. *Projective measurements are just a subset of a broader class of measurements that are allowed on quantum systems. However, in the chapters to come we deal exclusively with projective measurements and thus will not delve into the more general case.*

If the pre-measurement state is pure, i.e. $\rho = |\psi\rangle\langle\psi|$, then

$$\mathbb{P}(a|\rho) = \text{Tr}[P_a|\psi\rangle\langle\psi|] = \langle\psi|P_a|\psi\rangle \tag{2.13}$$

and the post-measurement state is $\rho'_a = |\psi'_a\rangle\langle\psi'_a|$ where

$$|\psi'_a\rangle = \frac{P_a|\psi\rangle}{\sqrt{\langle\psi|P_a|\psi\rangle}}. \tag{2.14}$$

Definition 1. An *observable* of a quantum system described by $\mathcal{H}$ is a Hermitian operator $\mathcal{O} \in \mathcal{L}(\mathcal{H})$.

There is a one-to-one correspondence between projective measurements with outcomes in the real numbers and observables. To illustrate this, let $\mathcal{O}$ be an observable and let A be the spectrum of $\mathcal{O}$. By the spectral theorem, $A \subset \mathbb{R}$ and

$$\mathcal{O} = \sum_{a \in A} a P_a \tag{2.15}$$

where P_a is the orthogonal projector onto the eigenspace of $\mathcal{O}$ corresponding to the eigenvalue a. One can show that the collection $\{P_a : a \in A\}$ meets the definition of a projective measurement. Conversely, if $\{P_a : a \in A\}$ is a projective measurement, then $\mathcal{O}$ as defined in (2.15) is Hermitian and therefore an observable. Given this correspondence we can say that a projective measurement is described by an observable.

Observables are useful when computing statistics on the outcome of a measurement. For example, let X be the random variable associated with the outcome of the measurement described by the observable $\mathcal{O}$. If the pre-measurement state is $\rho = |\psi\rangle\langle\psi|$, then

$$\begin{aligned} \mathbb{E}[X] &= \sum_{a \in A} a \mathbb{P}(a|\rho) \\ &= \sum_{a \in A} a \langle\psi| P_a |\psi\rangle \\ &= \langle\psi| \left(\sum_{a \in A} a P_a \right) |\psi\rangle \\ &= \langle\psi| \mathcal{O} |\psi\rangle. \end{aligned} \tag{2.16}$$

2.1.4 Composite systems

The fourth postulate defines how we interpret a composite system made up of two or more distinct physical systems.

Postulate 4. Let $\mathcal{H}_A$ and $\mathcal{H}_B$ be the state spaces of the physical systems A and B. The state space of the composite physical system composed of A and B is the tensor product of $\mathcal{H}_A$ and $\mathcal{H}_B$.

$$\mathcal{H}_{AB} = \mathcal{H}_A \otimes \mathcal{H}_B. \tag{2.17}$$

Let $|\psi_A\rangle \in \mathcal{H}_A$ be the state of A and $|\psi_B\rangle \in \mathcal{H}_B$ the state of B, then

$$|\Psi_{AB}\rangle = |\psi_A\rangle \otimes |\psi_B\rangle \tag{2.18}$$

is the state of the composite system and $|\Psi_{AB}\rangle$ is called a *product* vector. Conversely, if $|\Psi_{AB}\rangle \in \mathcal{H}_{AB}$ and there are no $|\psi_A\rangle \in \mathcal{H}_A$ and $|\psi_B\rangle \in \mathcal{H}_B$ such that (2.18) is true, then $|\Psi_{AB}\rangle$ is an *entangled* vector.

One may perform a measurement of the whole composite system or only part of it.

Definition 2. A *partial measurement* of a composite system is any measurement in which there is at least one component of the system on which each of the measurement operators acts as the identity.

For example, if $\{P_a : a \in A\} \subset \mathcal{L}(\mathcal{H}_A)$ is a projective measurement of system A, then

$$\{P_a \otimes I_{\mathcal{H}_B} : a \in A\} \subset \mathcal{L}(\mathcal{H}_A \otimes \mathcal{H}_B). \tag{2.19}$$

is a partial measurement of the system AB. A partial projective measurement is described by a *local observable*.

Definition 3. A *local observable* is an observable of a quantum system which acts as the identity on at least one component of a composite system.

For example, if $\mathcal{O} \in \mathcal{L}(\mathcal{H}_A)$ is the observable associated with the projective measurement $\{P_a : a \in A\}$, then the local observable describing the partial measurement $\{P_a \otimes I_{\mathcal{H}_B} : a \in A\}$ is

$$\mathcal{O} \otimes I_{\mathcal{H}_B} \in \mathcal{L}(\mathcal{H}_A \otimes \mathcal{H}_B). \tag{2.20}$$

2.2 Overview of quantum computing

We are now ready to introduce the theoretical framework under which quantum computers operate. Within this framework the basic unit of information is a physical system called a *qubit*. This is the quantum equivalent of a classical bit. When processing information, a quantum computer manipulates composite systems of qubits via series of controlled unitary transformations called *quantum gates*. The composition of these gates is called a *quantum circuit*. Gates and circuits are the quantum analogue of classic logic gates and binary circuits. This section serves as a primer on these topics, which are fundamental to understanding how quantum computers work. For a more exhaustive introduction we refer the reader to [28, chapter 4].

In section 2.2.1 we give a presentation of qubits and some projective measurements that are performed on them. In section 2.2.2 we describe quantum gates and circuits, and the typical procedure for executing an algorithm on a quantum computer. In section 2.2.3 we discuss why quantum computing is believed to be a more powerful model of computation than classical computing and some of the practical drawbacks that it currently suffers from.

2.2.1 Qubits and qubit measurements

Definition 4. A *qubit* is a quantum system described by the Hilbert space $\mathbb{C}^2$ equipped with the standard inner product.

Let $\{|0\rangle, |1\rangle\}$ be an orthonormal basis of $\mathbb{C}^2$ such as e.g. the canonical or *computational* basis:

$$|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad |1\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{2.21}$$

Postulate 1 tells us that the pure states of a qubit are described by the state vectors of the form

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \tag{2.22}$$

for some $\alpha, \beta \in \mathbb{C}$ such that $|\alpha|^2+|\beta|^2= 1$. If $\alpha, \beta > 0$, then the state of the qubit is neither $|0\rangle$ nor $|1\rangle$ but rather a *coherent superposition* of the two. As illustrated by Example 1 we can interpret this as $|\psi\rangle$ being $|0\rangle$ with probability $|\alpha|^2$ and $|1\rangle$ with probability $|\beta|^2$. This is in stark contrast with the state of a classical bit which is determined at any given time and must necessarily be one of $\{0, 1\}$.

Example 1. *Consider the observable*

$$\sigma_Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{2.23}$$

Its spectral decomposition is

$$\sigma_Z = |0\rangle\langle 0|-|1\rangle\langle 1| \tag{2.24}$$

therefore this observable describes the projective measurement

$$\{P_1, P_{-1}\}, \quad P_1 = |0\rangle\langle 0|, \quad P_{-1} = |1\rangle\langle 1|. \tag{2.25}$$

Suppose we perform this measurement on a qubit in the state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$. Postulate 3 tells us that the probability of each outcome is

$$\mathbb{P}(1) = \langle\psi|0\rangle\langle 0|\psi\rangle = |\langle 0|\psi\rangle|^2= |\alpha\langle 0|0\rangle + \beta\langle 0|1\rangle|^2= |\alpha|^2$$

and

$$\mathbb{P}(-1) = \langle\psi|1\rangle\langle 1|\psi\rangle = |\langle 1|\psi\rangle|^2= |\alpha\langle 1|0\rangle + \beta\langle 1|1\rangle|^2= |\beta|^2$$

with expected outcome

$$\langle\psi|\sigma_Z|\psi\rangle = \langle\psi|0\rangle\langle 0|\psi\rangle - \langle\psi|1\rangle\langle 1|\psi\rangle = |\alpha|^2-|\beta|^2. \tag{2.26}$$

The corresponding post measurement states are

$$|\psi_1'\rangle = \frac{1}{|\alpha|}|0\rangle\langle 0|\psi\rangle = \frac{\alpha}{|\alpha|}|0\rangle \tag{2.27}$$

and

$$|\psi_{-1}'\rangle = \frac{1}{|\beta|}|1\rangle\langle 1|\psi\rangle = \frac{\beta}{|\beta|}|1\rangle. \tag{2.28}$$

These vectors are proportional to $|0\rangle$ and $|1\rangle$, therefore the state collapses to $|0\rangle$ with probability $|\alpha|^2$ and $|1\rangle$ with probability $|\beta|^2$. Taking this measurement is called measuring a qubit in the computational basis.

The observable σ_Z of Example 1 is one of three special operators known as the *Pauli matrices*.

Definition 5. The *Pauli matrices* are defined as

$$\sigma_X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{2.29}$$

The Pauli matrices are traceless, Hermitian, unitary, and involutory:

$$\mathrm{Tr}\sigma_i = 0, \quad \sigma_i = \sigma_i^\dagger = \sigma_i^{-1}, \quad \sigma_i^2 = I \quad \text{for all } i \in \{X, Y, Z\}. \tag{2.30}$$

They satisfy the relations

$$\sigma_X\sigma_Y = i\sigma_Z, \quad \sigma_Y\sigma_Z = i\sigma_X, \quad \sigma_Z\sigma_X = i\sigma_Y. \tag{2.31}$$

They each have spectrum $\{1, -1\}$ and corresponding eigenvectors

$$\sigma_X : \{|+\rangle, |-\rangle\}, \quad \sigma_Y : \{|i\rangle, |-i\rangle\}, \quad \sigma_Z : \{|0\rangle, |1\rangle\} \tag{2.32}$$

where

$$|+\rangle = \frac{|0\rangle + |1\rangle}{\sqrt{2}}, \quad |-\rangle = \frac{|0\rangle - |1\rangle}{\sqrt{2}} \tag{2.33}$$

is the *Hadamard* basis and

$$|i\rangle = \frac{|0\rangle + i|1\rangle}{\sqrt{2}}, \quad |-i\rangle = \frac{|0\rangle - i|1\rangle}{\sqrt{2}} \tag{2.34}$$

is the *imaginary* basis. Both are alternative orthonormal bases of $\mathbb{C}^2$ to the computational one. Performing the projective measurement described by the Pauli-X and Y observables means to measure a qubit in these bases.

The identity on $\mathbb{C}^2$ is sometimes referred to as the fourth Pauli matrix: $\sigma_I = \mathbb{1}_{\mathbb{C}^2}$. This is because $\{\sigma_I, \sigma_X, \sigma_Y, \sigma_Z\}$ is a basis for the Hermitian operators in $\mathcal{L}(\mathbb{C}^2)$. We can express any observable $\mathcal{O} \in \mathcal{L}(\mathbb{C}^2)$ as a linear combination of these terms:

$$\mathcal{O} = v_0\sigma_I + v_1\sigma_X + v_2\sigma_Y + v_3\sigma_Z, \quad (v_0, v_1, v_2, v_3) \in \mathbb{R}^4. \tag{2.35}$$

We now consider the quantum analog of a length-n bit string: n *qubits*.

Definition 6. n *qubits* is a composite quantum system described by the 2^n dimensional Hilbert space

$$(\mathbb{C}^2)^{\otimes n} = \underbrace{\mathbb{C}^2 \otimes \cdots \otimes \mathbb{C}^2}_{n \text{ times}}. \tag{2.36}$$

Once again, let $\{|0\rangle, |1\rangle\}$ be an orthonormal basis of $\mathbb{C}^2$. An orthonormal basis of $(\mathbb{C}^2)^{\otimes n}$ is

$$\{|x_1 \cdots x_n\rangle : x_i \in \{0, 1\}\} \tag{2.37}$$

where we have used the shorthand notation $|x_1 \cdots x_n\rangle$ to refer to $|x_1\rangle \otimes \cdots \otimes |x_n\rangle$. Let $x_1 \cdots x_n$ be the binary representation of the natural number $x \in \{0, ..., 2^n - 1\}$. For any state vector $|\psi\rangle \in (\mathbb{C}^2)^{\otimes n}$ we can write

$$|\psi\rangle = \sum_{x=0}^{2^n-1} \alpha_x |x_1 \cdots x_n\rangle \tag{2.38}$$

for some $\alpha \in \mathbb{C}^{2^n}$ such that

$$\sum_{x=0}^{2^n-1} |\alpha_x|^2 = 1. \tag{2.39}$$

$|\psi\rangle$ is in a superposition of the basis states. To measure $|\psi\rangle$ in the computational basis means to perform the measurement described by the observable $(\sigma_Z)^{\otimes n}$. For all $x \in \{0, ..., 2^n - 1\}$ this measurement will collapse $|\psi\rangle$ to the basis state $|x_1 \cdots x_n\rangle$ with probability $|\alpha_x|^2$.

Just as in the single qubit case we can construct general n-qubit observables out of the Pauli matrices. $\{\sigma_I, \sigma_X, \sigma_Y, \sigma_Z\}^{\otimes n}$ is a basis for the Hermitian operators in $\mathcal{L}((\mathbb{C}^2)^{\otimes n})$, therefore any n-qubit observable can be expressed as

$$\mathcal{O} = \sum_{\mathcal{O}_j \in \{\sigma_I, \sigma_X, \sigma_Y, \sigma_Z\}^{\otimes n}} w_j \mathcal{O}_j \tag{2.40}$$

for some $w \in \mathbb{R}^{4^n}$. The $\mathcal{O}_j$ are called *Pauli strings*.

The physical system composed of n qubits is a composite system and therefore may be *entangled*. If the state of two qubits are entangled, then it is impossible to describe the state of either qubit independently of the other.

Example 2. *The state $|\Psi^+\rangle \in (\mathbb{C}^2)^{\otimes 2}$ defined by*

$$|\Psi^+\rangle = \frac{|00\rangle + |11\rangle}{\sqrt{2}} \tag{2.41}$$

is entangled. Indeed, if $|\Psi^+\rangle$ were not entangled, then by definition there would exist state vectors $|\psi_1\rangle, |\psi_2\rangle \in \mathbb{C}^2$ such that $|\Psi^+\rangle = |\psi_1\rangle \otimes |\psi_2\rangle$. However, letting

$$|\psi_1\rangle = \alpha_1|0\rangle + \beta_1|1\rangle \quad \text{and} \quad |\psi_2\rangle = \alpha_2|0\rangle + \beta_2|1\rangle, \tag{2.42}$$

we see this is impossible since

$$\begin{aligned} \frac{|00\rangle + |11\rangle}{\sqrt{2}} &= (\alpha_1|0\rangle + \beta_1|1\rangle) \otimes (\alpha_2|0\rangle + \beta_2|1\rangle) \\ &= \alpha_1\alpha_2|00\rangle + \alpha_1\beta_2|01\rangle + \beta_1\alpha_2|10\rangle + \beta_1\beta_2|11\rangle \end{aligned} \tag{2.43}$$

implies that $\alpha_1\alpha_2 = \beta_1\beta_2 = \frac{1}{\sqrt{2}}$ and $\alpha_1\beta_2 = \beta_1\alpha_2 = 0$ which is a contradiction.

When two qubits are entangled the measurement of one can instantaneously affect the state of the other no matter how far apart they are. For example, the state $|\Psi^+\rangle$ as in Example 2 is one of four special two-qubit states known as the Bell states. If a two-qubit system were in this state and we were to perform a partial measurement of only the first qubit in the computational basis we may find it in the state $|0\rangle$ or $|1\rangle$ with equal probability. However, if we were to measure the second qubit directly after, we know with certainty that we would find it in the same state as the first. Two classical bits, on the other hand, can always be described independently of one another and manipulation or observation of one bit does not affect the other.

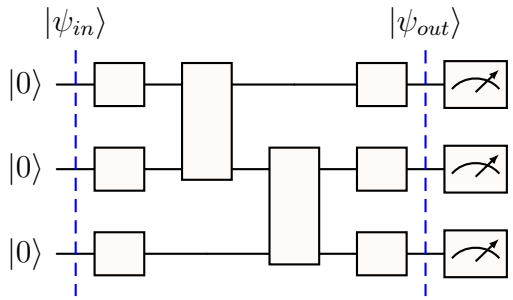

Figure 2.1: An example of a circuit diagram with width 3 and depth 4.

2.2.2 Quantum gates and circuits

Definition 7. An *n-qubit quantum gate* is a unitary operator $U \in \mathcal{L}((\mathbb{C}^2)^{\otimes n})$. An *elementary quantum gate* is a one or two-qubit gate.

Intuitively, a gate is just a unitary evolution of a system of qubits. One achieves a desired transformation by enforcing a specific Hamiltonian on the system for a controlled duration.

Remark 2. *When applying a gate $U \in \mathcal{L}((\mathbb{C}^2)^{\otimes n})$ to a system of $m \geq n$ qubits, it is implicit that we are, in fact, applying the gate $U' \in \mathcal{L}((\mathbb{C}^2)^{\otimes m})$ to the system where U' is U tensored with the identity on the Hilbert space of the $m - n$ qubits on which U does not act. If U is unitary, then so too is U'.*

Definition 8. A *quantum circuit* is a finite composition of quantum gates.

A composition of unitary operators is itself unitary, therefore a quantum circuit on a system of m qubits is itself a single m-qubit quantum gate. On the other hand, any m-qubit gate can be approximated to arbitrary precision in the operator norm by a finite composition of elementary gates. We can therefore interpret a quantum circuit as a finite composition of elementary gates. The *depth* of a circuit is the number of time steps necessary to apply all the gates of the circuit. The *width* of the circuit is its number of qubits.

The typical procedure for executing a quantum algorithm is the following.

1. Prepare a system of m qubits in some known state $|\psi_{in}\rangle$. It is often the case that $|\psi_{in}\rangle = |0\rangle^{\otimes m}$.

2. Apply a circuit U to the starting state, resulting in $|\psi_{out}\rangle = U|\psi_{in}\rangle$.

3. Fully or partially measure $|\psi_{out}\rangle$, typically in the computational basis.

The starting state, circuit, and measurement are chosen in such a way that the solution can be extracted from the outcome of the measurement of the final state. This procedure is often depicted in diagrams such as in figure 2.1. On the left are the qubits in the starting state. Next to each qubit is a wire and the wires travel through boxes which represent quantum gates. At the end of each wire is a lever, which represents a measurement being taken. As time passes we imagine the qubits traveling from left to right along the wires, transforming unitarily as they pass through each gate, until finally they reach the end and are measured.

We now present a description of some commonly used elementary quantum gates, how they act on the computational basis, and their circuit diagram depiction. We begin with single-qubit gates.

1. **Pauli gates.** The Pauli matrices are unitary and are therefore single qubit gates.

   (a) σ_X, also called *bit flip*, is the quantum equivalent of the classical NOT gate.

   $$\sigma_X|0\rangle = |1\rangle, \quad \sigma_X|1\rangle = |0\rangle. \tag{2.44}$$

   (b) σ_Z, also called *phase flip*, inverts the relative phase of the state vector.

   $$\sigma_Z|0\rangle = |0\rangle, \quad \sigma_Z|1\rangle = -|1\rangle. \tag{2.45}$$

   (c) σ_Y, also called *bit-phase flip*, is the composition of a bit and phase flip and multiplication by a global, and thus irrelevant, phase.

   $$\sigma_Y|0\rangle = i|1\rangle, \quad \sigma_Y|1\rangle - i|0\rangle. \tag{2.46}$$

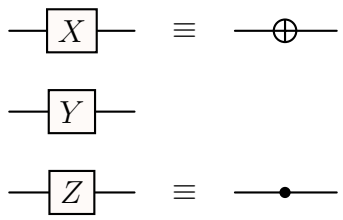

Figure 2.2: Pauli gates

2. **Hadamard gate.**

   $$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}. \tag{2.47}$$

   This gate sends the computational basis to the Hadamard basis.

   $$H|0\rangle = |+\rangle, \quad H|1\rangle = |-\rangle. \tag{2.48}$$

   Applying $H^{\otimes m}$ to $|0\rangle^{\otimes m}$ results in the uniform superposition of the basis states of $(\mathbb{C}^2)^{\otimes m}$.

   $$H^{\otimes m}|0\rangle^{\otimes m} = |+\rangle^{\otimes m} = \left(\frac{|0\rangle + |1\rangle}{\sqrt{2}}\right)^{\otimes m} = \frac{1}{2^{m/2}} \sum_{x=0}^{2^m-1} |x_1 \cdots x_m\rangle. \tag{2.49}$$

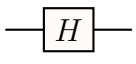

Figure 2.3: Hadamard gate

3. **Phase shift gates.** A family of gates represented by the matrix

$$P(\theta) = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\theta} \end{pmatrix}. \tag{2.50}$$

This transformation shifts the relative phase of the state vector by θ.

$$P(\theta)|0\rangle = |0\rangle, \quad P(\theta)|1\rangle = e^{i\theta}|1\rangle. \tag{2.51}$$

Some commonly used phase gates are

$$Z = P(\pi), \quad S = P(\pi/2), \quad \text{and} \quad T = P(\pi/4). \tag{2.52}$$

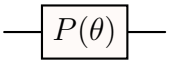

Figure 2.4: Phase shift gate

4. **Rotation gates.** Every state vector has a unique representation as a point on the surface of the unit sphere in $\mathbb{R}^3$. Consider $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$. We can rewrite

$$\alpha = |\alpha|e^{i\phi_1}, \text{ and } \beta = |\beta|e^{i\phi_2} \tag{2.53}$$

for some $\phi_1, \phi_2 \in [0, 2\pi)$. Since $|\alpha| \leq 1$, there is a $\theta \in [0, \pi]$ such that $|\alpha| = \cos(\theta/2)$ in which case

$$|\beta| = \sqrt{1 - |\alpha|^2} = \sqrt{1 - \cos^2(\theta/2)} = \sin(\theta/2). \tag{2.54}$$

Rewrite

$$\begin{aligned} |\psi\rangle &= e^{i\phi_1}\cos(\theta/2)|0\rangle + e^{i\phi_2}\sin(\theta/2)|1\rangle \\ &= e^{i\phi_1}(\cos(\theta/2)|0\rangle + e^{i(\phi_2-\phi_1)}\sin(\theta/2)|1\rangle). \end{aligned} \tag{2.55}$$

Let $\phi = \phi_2 - \phi_1$. Then $|\psi\rangle$ is proportional to the state

$$\cos(\theta/2)|0\rangle + e^{i\phi}\sin(\theta/2)|1\rangle. \tag{2.56}$$

We can therefore describe $|\psi\rangle$ with just $\phi \in [0, 2\pi]$ and $\theta \in [0, \pi]$. In spherical coordinates $(1, \theta, \phi)$ specifies a point on the surface of the unit sphere in $\mathbb{R}^3$.

The rotation gates R_X, R_Y and R_Z rotate this point about the x, y and z axes, respectively.

$$\begin{aligned} R_X(\theta) &= e^{-i\frac{\theta}{2}\sigma_X} &= \begin{pmatrix} \cos(\theta/2) & -i\sin(\theta/2) \\ -i\sin(\theta/2) & \cos(\theta/2) \end{pmatrix} \\ R_Y(\theta) &= e^{-i\frac{\theta}{2}\sigma_Y} &= \begin{pmatrix} \cos(\theta/2) & -\sin(\theta/2) \\ \sin(\theta/2) & \cos(\theta/2) \end{pmatrix} \\ R_Z(\theta) &= e^{-i\frac{\theta}{2}\sigma_Z} &= \begin{pmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{i\theta/2} \end{pmatrix} \end{aligned} \tag{2.57}$$

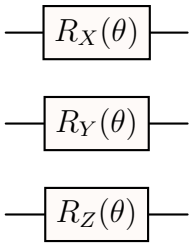

Figure 2.5: Rotation gates

5. **Parameter encoding gates.** These gates encode a parameter $\theta \in \mathbb{R}$ as a time evolution generated by the operator $\mathcal{G}$ over the period $\theta/2$.

$$U(\theta) = e^{-i\frac{\theta}{2}\mathcal{G}}. \tag{2.58}$$

The rotation gates are special cases of parameter encoding gates with $\mathcal{G} \in \{\sigma_X, \sigma_Y, \sigma_Z\}$.

The Pauli matrices along with H, S and T are all fixed gates while phase shift, rotation, and parameter encoding gates are examples of *parametric* gates. We can think of a parametric single-qubit gate U is a function

$$U : \mathbb{R} \to \mathcal{U}(2) \tag{2.59}$$

and $U(\theta)$, $\theta \in \mathbb{R}$, as a family of fixed unitaries. These types of gates are a key element of parametric circuits which we describe in section 2.3.

We now present some examples of commonly used two-qubit gates.

1. **Swap gate.** The swap gate swaps two qubits.

$$S|00\rangle = |00\rangle, \quad S|01\rangle = |10\rangle, \quad S|10\rangle = |01\rangle, \quad S|11\rangle = |11\rangle. \tag{2.60}$$

Figure 2.6: Swap gate

2. **Controlled gates.** Given a single qubit unitary U a controlled-U (C-U) gate takes as input a target $|t\rangle$ and control $|c\rangle$ qubit. It applies U to the target qubit if and only if the control qubit is $|1\rangle$.

$$C\text{-}U(|c\rangle \otimes |t\rangle) = \begin{cases} |c\rangle \otimes |t\rangle & \text{if } c = 0 \\ |c\rangle \otimes U|t\rangle & \text{if } c = 1 \end{cases}. \tag{2.61}$$

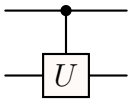

Figure 2.7: Controlled-U gate

One could construct a controlled gate with any unitary U. Two examples are

(a) **CNOT.** This is a controlled-U gate with $U = \sigma_X$.

$$CNOT(|c\rangle \otimes |t\rangle) = |c\rangle \otimes |t \oplus c\rangle \tag{2.62}$$

where $\oplus$ is the binary XOR.

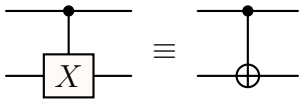

Figure 2.8: CNOT gate

(b) **C-Z.** This is a controlled-U gate with $U = \sigma_Z$.

$$C\text{-}Z(|c\rangle \otimes |t\rangle) = \begin{cases} -|c\rangle \otimes |t\rangle & \text{if } c = t = 1 \\ |c\rangle \otimes |t\rangle & \text{otherwise} \end{cases}. \tag{2.63}$$

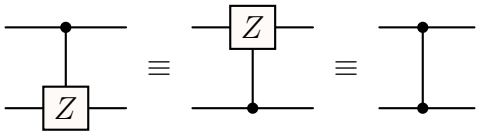

Figure 2.9: C-Z gate

We conclude this section with an example of a quantum circuit which uses two of the gates we have described.

Example 3. *Let $|\psi_{in}\rangle = |0\rangle^{\otimes 2}$ and let U be the circuit resulting from applying a Hadamard gate to the first qubit followed by a CNOT gate: $U = CNOT(H \otimes I_2)$. Then the final state is*

$$\begin{aligned} |\psi_{out}\rangle &= U|\psi_{in}\rangle \\ &= CNOT(H \otimes I_2)(|0\rangle \otimes |0\rangle) \\ &= CNOT\left(\frac{|0\rangle + |1\rangle}{\sqrt{2}} \otimes |0\rangle\right) \\ &= CNOT\left(\frac{|00\rangle + |10\rangle}{\sqrt{2}}\right) \\ &= \frac{|00\rangle + |11\rangle}{\sqrt{2}}. \end{aligned} \tag{2.64}$$

$|\psi_{out}\rangle$ is the Bell state $|\Psi^+\rangle$ that was introduced in example 2. This circuit is depicted in figure 2.10.

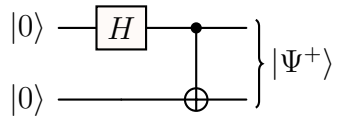

Figure 2.10: The diagram of a circuit which constructs the Bell state $|\Psi^+\rangle$.

2.2.3 Quantum advantage

Time complexity is a metric used to evaluate both classical and quantum algorithms. It does *not* refer to the length of time required for an algorithm to run, which will in any case depend on the hardware used, but rather the number of *elementary operations* required to execute the algorithm. In the quantum model one elementary operation is the application of one elementary quantum gate. In the classical model it refers to a basic operation that can be implemented in constant time. Although the literature does often include operations such as basic arithmetic and comparison of fixed width integers, for simplicity we will adopt the convention that one elementary operation corresponds to the application of one elementary logic gate.

Time complexity is expressed as a function of the size of the input to the algorithm, often using *asymptotic notation*. Asymptotic notation is a useful tool for characterizing the behavior of a function as its argument(s) grow.

Definition 9. Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$.

- $f(n) \in O(g(n))$ if there exist positive constants n_0 and c such that $|f(n)| \leq c|g(n)|$ for all $n \geq n_0$.
- $f(n) \in \Theta(g(n))$ if there exists positive constants n_0, c_1 and c_2 such that $c_1|g(n)| \leq |f(n)| \leq c_2|g(n)|$ for all $n \geq n_0$.
- $f(n) \in \Omega(g(n))$ if there exist positive constants n_0 and c such that $|f(n)| \geq c|g(n)|$ for all $n \geq n_0$.

The first point refers to *big-O* notation and it is useful in the context of algorithm analysis, especially when it comes to describing *efficiency*. It allows one to categorize an algorithm in terms of its run time in the worst case scenario. Let $f(n)$ be the time complexity of an algorithm with input size n. We say that the algorithm is efficient if there exists a polynomial p such that $f(n) \in O(p(n))$. Examples of efficient run times include constant: $O(1)$, linear: $O(n)$, logarithmic: $O(\log n)$, and $O(n^k)$ where k is a constant independent of n. A run time of $O(2^n)$ is exponential and not considered efficient.

Quantum computing is believed to be a more powerful computational model than classical computing. It includes elementary operations that are not possible under the classical model and this allows quantum computers to perform certain tasks more efficiently. A *reversible* logic gate implements a one-to-one boolean function between a number of inputs and the same number of outputs. It is so named because the input of a reversible gate is uniquely determined by the output and thus it is invertible. These gates are special because for every reversible gate there is a unitary operator which performs the same mapping. Furthermore, reversible gates are *universal* - every binary circuit can be constructed out of only reversible

gates. This implies that every binary circuit can be implemented on a quantum computer. On the other hand, there are quantum gates that have no classical counterpart. Simulating a quantum circuit using a classical computer requires, in the worst case, a time complexity that scales exponentially with the number of qubits.

Another distinguishing feature that highlights the power of quantum computing is *parallelism*. Consider a function $f : \{0,1\}^m \to \{0,1\}^n$. A quantum *oracle* for this function is a unitary $U_f \in \mathcal{L}((\mathbb{C}^2)^{\otimes(m+n)})$ such that for all $|x\rangle \in (\mathbb{C}^2)^{\otimes m}$ and $|y\rangle \in (\mathbb{C}^2)^{\otimes n}$,

$$U_f|x\rangle|y\rangle = |x\rangle|y \oplus f(x)\rangle. \tag{2.65}$$

Recall that applying $H^{\otimes m}$ to $|0\rangle^{\otimes m}$ results in the uniform superposition of the basis states of $(\mathbb{C}^2)^{\otimes m}$. By applying U_f to this state we get

$$U_f(H^{\otimes m}|0\rangle^{\otimes m})|0\rangle^{\otimes n} = \frac{1}{2^{m/2}} \sum_{x\in\{0,1\}^m} |x\rangle|f(x)\rangle. \tag{2.66}$$

Suppose we would like to determine some global property of an unknown function f and we have access to the oracle U_f. With just one application of U_f the resulting state contains information about $f(x)$ for all $x \in \{0,1\}^m$. With clever further processing one can construct a state from which the desired property can be extracted. This concept is known as quantum parallelism and it allows quantum algorithms to explore multiple solutions to a problem simultaneously, contributing to a faster run time.

The most famous example of this is Shor's algorithm which factors large numbers into prime factors in a time that grows only polynomially with the number of digits of the number to be factored, providing an exponential speed up over its best known classic counterpart [6]. Another prominent example is Grover's algorithm which provides a quadratic speedup for searching an unstructured database [7].

Time complexity, however, is not the whole picture when considering whether or not quantum computers provide a significant advantage over classical computers. We must also take into account the applicability of an algorithm and whether or not it can actually be carried out in practice. To achieve a quantum advantage a quantum computer must solve a practical, real-world problem more efficiently than any classical computer is capable of.

Factoring large numbers is a highly practical problem. The ability to do so efficiently would render certain widely used public-key cryptosystems insecure. However, using Shor's algorithm to factor cryptographically relevant numbers is not possible using current quantum computers and therefore does not provide a demonstrable advantage. In reality, few claims of a quantum achievement in practice have been made, and many have later been challenged or refuted [29].

The quantum computers of today suffer primarily from noise, decoherence, and limited qubits. Noise refers to random disruptions of the state of the qubits, whether from interference or imperfect hardware. Decoherence occurs when the information held by the state of the qubits is lost due to unavoidable interaction with their environment. Errors due to noise and decoherence accumulate with each gate that is applied, making deeper circuits especially prone to inaccuracies. And finally, today's quantum computers have on the order of tens to a thousand qubits with limited connectivity and no capability for error correction.

It is thought that we are years, maybe even decades, away from scalable, fault tolerant devices [27]. Consequently, the push to find a quantum advantage has centered around applications suited towards NISQ devices. VQAs have emerged as a leading contender in this vein.

2.3 Introduction to variational quantum algorithms

A VQA is a hybrid classical-quantum optimization based algorithm. It involves using a quantum subroutine to compute a parametric function, while a classical computer performs a parameter optimization. The VQA is a framework which can be applied to a variety of applications including, but not limited to, combinatorial optimization [30], solving systems of linear equations [31], quantum error correction [32], and quantum simulation [33]. As described in section 2.2.3 it has garnered interest in recent years due to its suitability to NISQ devices. In this section we give an introduction to this framework as well as two specific applications. In section 2.3.2 we present the variational quantum eigensolver (VQE) [34], a method for finding ground states of a physical system, and the first proposal of any VQA. In section 2.3.3 we describe deterministic quantum neural networks [19], a machine learning paradigm and the VQA example which the main result of this thesis is tailored towards. For a more in-depth look at the general VQA framework we refer the reader to [21].

2.3.1 The VQA framework

A fundamental component of a VQA is a *parametric circuit*. Parametric circuits, also called *variational circuits*, are quantum circuits in which one or more of the elementary gates composing the circuit are parametric. (See section 2.2.2 for a description of some parametric gates.) A specific choice and arrangement of such gates is called an *ansatz*. This ansatz, which we denote by U, takes as input a set of parameters Θ and defines a family of fixed circuits which we denote by $U(\Theta)$. Let

$$|\psi_{out}(\Theta)\rangle = U(\Theta)|\psi_{in}\rangle \tag{2.67}$$

be the output of the n-qubit circuit $U(\Theta)$ for a given input state $|\psi_{in}\rangle$, and let $\mathcal{O} \in \mathcal{L}((\mathbb{C})^{\otimes n})$ be an observable. The output of the VQA can be extracted from the measurement described by $\mathcal{O}$ of the target state $|\psi_{out}(\Theta^*)\rangle$, where Θ^* is a set of parameters which the algorithm has deemed to be optimal. In both of the examples we consider, the output of the VQA is given by the expected value of the measurement of the target state:

$$\langle\psi_{out}(\Theta^*)|\mathcal{O}|\psi_{out}(\Theta^*)\rangle. \tag{2.68}$$

To determine the expected value one must prepare $|\psi_{out}(\Theta^*)\rangle$ and measure it a number of times then take the average. Each of the measurements is called a *shot* and the number of shots increases with the desired accuracy.

Determining the optimal parameters Θ^* is one of the key tasks in executing a VQA. It is done by minimizing a *cost function* $\mathcal{C}$. The cost, which is a function of $\langle\psi_{out}(\Theta)|\mathcal{O}|\psi_{out}(\Theta)\rangle$

and possibly some other data, is specifically designed so that

$$\Theta^* = \arg\min_{\Theta} \mathcal{C}(\Theta). \tag{2.69}$$

The parameters which minimize the cost function prescribe a fixed quantum circuit which results in an output state that encodes the correct solution, or an approximation thereof, of the problem the VQA aims to solve.

The optimization is carried out via an iterative procedure, such as gradient descent, in which an initial guess is chosen and at each successive iteration the parameters are updated in the opposite direction of the gradient of $\mathcal{C}$. The evaluation of $\mathcal{C}$ and its gradients is left up to a quantum computer while the parameter update is done by a classical computer. In this sense the VQA is a hybrid algorithm.

Choosing an appropriate ansatz is another critical step when executing a VQA. A few metrics typically used to evaluate an ansatz are its *expressivity*, *trainability*, and *cost*. The first metric, expressivity, refers to the ability of an ansatz to prepare the correct target state or at least a good approximation of it which is necessary for the VQA to be successful. The probability that this is true is maximized if the family $U(\Theta)$ is capable of preparing states that are well representative of the Hilbert space they reside in [35].

Trainability refers to the ability of the optimization procedure to converge to the global minimum of the cost function. Certain circuit architectures suffer from a phenomenon known as *barren plateaus* in which, with high probability, the gradient of the cost function decays exponentially with the number of qubits [23]. This was shown to be true for deep unstructured variational circuits when randomly initialized in [22]. This pitfall causes the number of shots required to calculate the direction of the parameter update to increase exponentially, destroying any hope of a quantum advantage.

The last metric we will touch on is the cost which takes into account, among other things, the depth and connectivity of a circuit as well as the number of measurements required. Highly expressive circuits typically suffer from a higher cost and worse trainability therefore a delicate balance must be struck in order to choose an effective ansatz.

This is particularly true in the NISQ era where shallow circuits are preferred. Fewer gates and shorter execution time reduces the accumulation of errors and the chance of decoherence. In fact, the ability of a VQA to be implemented using a shallow circuit is one of the reasons why it is favored for NISQ devices. Another reason is its tolerance for noise. The averaging of the shots taken to compute the expected value of a measurement smooths out some of the noise introduced by the individual computations. In addition, VQAs benefit from the techniques of classical optimization which are well studied and well developed.

2.3.2 Variational quantum eigensolvers

In section 2.1.2 we introduced Schrödinger's equation which dictates the evolution of a physical system:

$$i\hbar \frac{d}{dt}|\psi_t\rangle = H|\psi_t\rangle. \tag{2.70}$$

H, called the Hamiltonian of the system, is a Hermitian operator and therefore an observable. By measuring this observable one measures the total energy of the system with each

eigenvalue of H corresponding to a possible energy level. For this reason, the spectrum of H is called the *energy spectrum*. The eigenvectors of H are called *energy eigenstates* or sometimes *stationary states*. To illustrate this, suppose that $|\psi_0\rangle$ is an energy eigenstate corresponding to the eigenvalue E_0. Then according to Schrodinger's equation we have

$$\left(i\hbar\frac{d}{dt}|\psi_t\rangle\right)_{t=0} = H|\psi_0\rangle = E_0|\psi_0\rangle, \tag{2.71}$$

and thus

$$|\psi_t\rangle = e^{-iEt/\hbar}|\psi_0\rangle. \tag{2.72}$$

$|\psi_t\rangle$ is proportional to $|\psi_0\rangle$ for all $t \in \mathbb{R}$, so as time varies the state of the system remains the same in every observable way.

The smallest eigenvalue of H is the lowest possible energy of the system and is called the *ground state energy*. The corresponding energy eigenstate is called the *ground state*. For many applications such as e.g. quantum chemistry [36], it is desirable to determine the ground state of a given physical system and its corresponding energy. However, even with full knowledge of the Hamiltonian, using direct methods to find these values can be computationally infeasible for very large systems. The VQE, first proposed in [37], aims to find an upper bound on the ground state energy and a corresponding approximation of the ground state. It is based on the *Ritz-Rayleigh principle* which states that if E_0 is the smallest eigenvalue of H, and $|E_0\rangle$ its corresponding eigenvector, then

$$E_0 \leq \frac{\langle\psi|H|\psi\rangle}{\langle\psi|\psi\rangle} \tag{2.73}$$

for all $|\psi\rangle \in \mathcal{H}$ with equality when $|\psi\rangle = |E_0\rangle$ [38]. Since $\langle E_0|E_0\rangle = 1$ we can thereby find E_0 and $|E_0\rangle$ by minimizing $\langle\psi|H|\psi\rangle$ over all possible state vectors in $\mathcal{H}$:

$$E_0 = \min_{\substack{|\psi\rangle\in\mathcal{H} \\ \||\psi\rangle\|=1}} \langle\psi|H|\psi\rangle. \tag{2.74}$$

To apply the VQE framework to this problem we first define a structure-preserving transformation

$$\mathcal{T} : \mathcal{L}(\mathcal{H}) \to \mathcal{L}((\mathbb{C}^2)^{\otimes n}) \tag{2.75}$$

and let $\hat{H} = \mathcal{T}(H)$. $\hat{H}$ is an observable that describes a measurement of trial states that can be prepared with a quantum computer. We use a variational circuit to prepare trial states as $|\psi_{out}(\Theta)\rangle = U(\Theta)|\psi_{in}\rangle$. The cost function is given by

$$\mathcal{C}(\Theta) = \langle\psi_{out}(\Theta)|\hat{H}|\psi_{out}(\Theta)\rangle. \tag{2.76}$$

According to (2.74), $E_0 \leq \mathcal{C}(\Theta)$ for all Θ, with equality if $|\psi_{out}(\Theta)\rangle = |E_0\rangle$. The output of the algorithm is therefore $\mathcal{C}(\Theta^*)$ and $|\psi_{out}(\Theta^*)\rangle$ which is an upper bound on the ground state energy and an approximation of the ground state.

2.3.3 Quantum neural networks

Before diving into quantum neural networks let us first recall some notions from machine learning. In the context of machine learning a model is an object capable of ingesting data and making predictions or decisions about that data based on patterns, relationships, and structure that it has "learned". Often a machine learning task comes with an input domain $\mathcal{X}$, an output domain $\mathcal{Y}$, and a corpus of *training data* in the form of input-output pairs:

$$\mathcal{D} = \{(x^{(i)}, y^{(i)}) \in \mathcal{X} \times \mathcal{Y} : i \in \{1, ..., n\}\}, \tag{2.77}$$

and the goal is to learn the relationship between $\mathcal{X}$ and $\mathcal{Y}$ based on $\mathcal{D}$. This task is called *supervised learning*. The two most common types of supervised learning are *regression*, in which the output domain is continuous, and *classification*, in which the output domain is a discrete set of labels. An example of a regression task is to predict the price of a house given its size and location. A classification task could be, given an image, to predict if it contains a human face or not. When a classification task has exactly two possible labels, then we say it is a *binary classification*.

In the case of supervised learning a model is typically a parametrized class of functions from the input to the output domain $f_\Theta : \mathcal{X} \to \mathcal{Y}$ and the learning procedure involves identifying the parameters Θ^* which result in the function that best fits the training data. This is done by minimizing a cost function which quantifies the discrepancy between the training outputs and predictions by the function on the inputs. One example of such a cost function is the *mean squared error*

$$\mathcal{C}(\Theta) = \sum_{i=1}^{n} (f_\Theta(x_i) - y_i)^2. \tag{2.78}$$

The goal is that for the Θ^* that minimizes $\mathcal{C}(\Theta)$, not only do we have

$$f_{\Theta^*}(x^{(i)}) \approx y^{(i)} \text{ for all } i \in \{1, ..., n\} \tag{2.79}$$

but also f_{Θ^*} is able to make accurate predictions on unseen $x \in \mathcal{X}$. The ability of a model function to do so is called its *generalization power*. In general, when a model function has a large amount of parameters, especially when the parameters far outnumber the training examples, it has more flexibility to fit said examples, including any random fluctuations which are not indicative of the overall relationship between the input and output domains. This can lead to *overfitting* which occurs when a model perfectly fits the training data but does not generalize well.

One of the most successful and widely used class of parametrized functions is the *deep neural network*. These functions excel at learning tasks involving highly dimensional data such as image and speech recognition, and have been observed to generalize well even when heavily over-parametrized.

Definition 10. Let $L \in \mathbb{N}$, $n_0, ..., n_{L+1} \in \mathbb{N}$ and $\phi : \mathbb{R} \to \mathbb{R}$. A depth L *fully connected feedforward neural network* with input dimension n_0, output dimension n_{L+1}, hidden layer widths $n_1, ..., n_L$, and activation ϕ is a function

$$x^{(0)} \in \mathbb{R}^{n_0} \to z^{(L+1)} \in \mathbb{R}^{n_{L+1}} \tag{2.80}$$

defined by the relations

$$\begin{cases} z_i^{(l)} = b_i^{(l)} + \sum_{j=1}^{n_{l-1}} W_{ij}^{(l)} x_j^{(l-1)}, & l = 1, ..., L+1 \\ x_j^{(l)} = \phi(z_j^{(l-1)}), & l = 1, ..., L \end{cases} \tag{2.81}$$

where $W^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}}$ and $b^{(l)} \in \mathbb{R}^{n_l}$ for all $l \in \{1, ..., L+1\}$.

L, $n = (n_0, ..., n_{L+1})$, and ϕ are called *hyperparameters*. They are fixed and determine what we call the network *architecture*. L is the number of hidden layers in the network. Networks with more than one hidden layer are considered *deep*. The matrices $W^{(l)}$ and vectors $b^{(l)}$ are the trainable parameters.

We are now ready to introduce a class of deterministic model functions that utilize variational quantum circuits. They are inspired by the classical neural networks and have thus been dubbed *quantum neural networks*.

Definition 11. Let $\mathcal{X}$ be an input domain, Ω a parameter domain, and $m \in \mathbb{N}$. Let $U(x, \Theta)$ be a variational circuit on m qubits which is parametrized by $x \in \mathcal{X}$ and $\Theta \in \Omega$. Let $\mathcal{O} \in \mathcal{L}((\mathbb{C}^2)^{\otimes m})$ be an observable, and

$$|\psi(x, \Theta)\rangle = U(x, \Theta)|0^m\rangle. \tag{2.82}$$

The function

$$f_\Theta(x) = \langle \psi(x, \Theta)|\mathcal{O}|\psi(x, \Theta)\rangle \tag{2.83}$$

defines a *deterministic quantum model*.

Borrowing the nomenclature from the classical version, we say that m is the *width* of the circuit. It is a hyperparameter while Θ are trainable parameters. The ansatz that defines the family of circuits $U(x, \Theta)$ is called the circuit *architecture* and the depth of the circuit is akin to the depth of the network.

Clearly the supervised learning of a deterministic quantum model can be phrased as a VQA. The problem the VQA aims to solve is to learn a function that fits a set of training data. The output of the algorithm is the model function $x \to f_{\Theta^*}(x)$ where Θ^* minimizes a cost function such as the mean squared error given by (2.78).

Chapter 3

The model function of quantum neural networks

In chapter 2 we introduced quantum neural networks. We now turn our attention to the model function they generate and how it can be viewed as a random variable which we will label $f(\Theta)$. The goal of this chapter is twofold. Firstly, to fix some assumptions on the architecture of the network. And secondly, to define a number of objects and quantities which describe this architecture. Understanding these quantities will be crucial to characterizing the Gaussianity of $f(\Theta)$, which is our ultimate goal and the main subject of chapter 4. Let us first set the stage by formally defining $f(\Theta)$.

Let $\mathcal{H} = (\mathbb{C}^2)^{\otimes m}$ be the Hilbert space representing a system of m qubits and let $|\psi\rangle \in \mathcal{H}$ be a vector representing the initial state of the system. Let $\mathcal{X}$ be an input domain and Ω a parameter domain, and let $U(x, \Theta) \in \mathcal{L}(\mathcal{H})$ be a unitary representing a variational circuit parametrized by $x \in \mathcal{X}$ and $\Theta \in \Omega$. Denote by $|\psi(x, \Theta)\rangle \in \mathcal{H}$ the result of applying $U(x, \Theta)$ to the initial state: $|\psi(x, \Theta)\rangle = U(x, \Theta)|\psi\rangle$. Finally, let $\mathcal{O} \in \mathcal{L}(\mathcal{H})$ be an observable describing a projective measurement of the system. As we saw in section 2.3, the model function of a quantum neural network is given by

$$f_\Theta(x) = \langle\psi(x, \Theta)|\mathcal{O}|\psi(x, \Theta)\rangle. \tag{3.1}$$

Let us fix $\hat{x} \in \mathcal{X}$ and let Θ be sampled from Ω according to the probability distribution $\mathbb{P}$, as is the case when the model is randomly initialized. Then

$$\langle\psi(\hat{x}, \Theta)|\mathcal{O}|\psi(\hat{x}, \Theta)\rangle \tag{3.2}$$

is a random variable on the probability space $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$ where $\mathcal{B}(\Omega)$ is the Borel sigma algebra generated by Ω. Denote by N its standard deviation. We define $f(\Theta)$ to be the random variable on the same probability space, normalized so that $\mathbb{V}[f(\Theta)] = 1$:

$$f(\Theta) := \frac{1}{N}\langle\psi(\hat{x}, \Theta)|\mathcal{O}|\psi(\hat{x}, \Theta)\rangle. \tag{3.3}$$

We will henceforth refer to N as the *normalization constant* of $f(\Theta)$. Given that $\hat{x}$ is fixed, going forward, we will refer to the unitary $U(\hat{x}, \Theta)$ as $U(\Theta)$ and the state $|\psi(\hat{x}, \Theta)\rangle$ as $|\psi(\Theta)\rangle$.

The rest of this chapter is outlined as follows. In section 3.1 we discuss in detail the observable $\mathcal{O}$ and introduce quantities which we call *observable weights*. In section 3.2 we discuss the circuit $U(\Theta)$ and introduce objects called *interactions* and *light cones*. In section 3.3 we define a subspace of $\mathcal{H}$ which will inform both how the light cones and observable weights affect the law of $f(\Theta)$, and whether we can expect to achieve a quantum advantage. We follow the notation of [25, chapter 2] in which light cones and observables of variational circuits are discussed. The notable difference is that we consider a broader class of observables. We will make clear any important distinctions.

3.1 The observable

As we saw in section 2.2.1, $\{\sigma_I, \sigma_X, \sigma_Y, \sigma_Z\}^{\otimes m}$ forms a basis of the observables in $\mathcal{H}$, therefore we can express $\mathcal{O}$ as

$$\mathcal{O} = \sum_{j=1}^{p} w_j \mathcal{O}_j \tag{3.4}$$

with $\mathcal{O}_j \in \{\sigma_I, \sigma_X, \sigma_Y, \sigma_Z\}^{\otimes m}$ called a *local observable* and $w = (w_1, ..., w_p) \in \mathbb{R}^p$. Furthermore, this representation is unique.

According to (3.4) we can express $f(\Theta)$ as the sum

$$f(\Theta) = \frac{1}{N}\langle\psi(\Theta)|\mathcal{O}|\psi(\Theta)\rangle = \sum_{l=1}^{p} f_j(\Theta) \tag{3.5}$$

where we define

$$f_j(\Theta) := \frac{w_j}{N}\langle\psi(\Theta)|\mathcal{O}_j|\psi(\Theta)\rangle. \tag{3.6}$$

When executing a VQA in practice, each of the $f_j(\Theta)$ are computed individually on a quantum device and the sum is computed classically. Since performing the measurement described by $\mathcal{O}_j$ on $|\psi(\Theta)\rangle$ collapses the state, it is not possible to then take another measurement described by $\mathcal{O}_{j'}$ for some $j' \neq j$. It is therefore necessary to prepare p independent copies of $|\psi(\Theta)\rangle$ in order to perform all p measurements. Consequently, a necessary condition for the VQA to be considered an efficient algorithm is that $p \in O(g(m))$ for some polynomial g.

Lemma 1. *Let* $W = |w|_\infty$, *then*

$$|f_j(\Theta)| \leq \frac{W}{N} \tag{3.7}$$

for all $j \in \{1, ..., p\}$ *implying* $|f(\Theta)| \leq \frac{pW}{N}$.

Proof. It is enough to show that $|\langle\psi|\mathcal{O}_j|\psi\rangle| \leq 1$ for all $|\psi\rangle \in \mathcal{H}$. The *operator norm* of $\mathcal{O}_j$ is given by

$$\|\mathcal{O}_j\|_{op} = \sup_{\||\psi\rangle\|=1} \|\mathcal{O}_j|\psi\rangle\|. \tag{3.8}$$

Since $\mathcal{O}_j^\dagger = \mathcal{O}_j$ and $\mathcal{O}_j^2 = \mathbb{1}_{\mathcal{H}}$, we have

$$\|\mathcal{O}_j|\psi\rangle\| = \sqrt{\langle\psi|\mathcal{O}_j^\dagger\mathcal{O}_j|\psi\rangle} = \sqrt{\langle\psi|\psi\rangle} = 1 \tag{3.9}$$

for all $|\psi\rangle$ such that $\||\psi\rangle\| = 1$, therefore $\|\mathcal{O}_j\|_{op} = 1$ for all $j \in \{1, ..., p\}$. The spectral radius of a linear operator is bounded by the operator norm of said operator [39], thus if $\mathcal{O}_j = \sum_i \lambda_i |\lambda_i\rangle\langle\lambda_i|$ is the spectral decomposition of $\mathcal{O}_j$, then $|\lambda_i| \leq 1 \forall i$. It follows that

$$\begin{aligned} |\langle\psi(\Theta)|\mathcal{O}_j|\psi(\Theta)\rangle| &= \left|\langle\psi(\Theta)\left(\sum_i \lambda_i|\lambda_i\rangle\langle\lambda_i|\right)|\psi(\Theta)\rangle\right| \\ &= \left|\sum_i \lambda_i\langle\psi(\Theta)|\lambda_i\rangle\langle\lambda_i|\psi(\Theta)\rangle\right| \\ &\leq \sum_i |\lambda_i| \cdot |\langle\lambda_i|\psi(\Theta)\rangle|^2 \\ &\leq \sum_i |\langle\lambda_i|\psi(\Theta)\rangle|^2 \\ &= 1 \end{aligned} \tag{3.10}$$

since $\{|\lambda_i\rangle\}_i$ is an orthonormal basis of $\mathcal{H}$. □

3.1.1 Observable weights

It is useful to write the local observable $\mathcal{O}_j$ explicitly as

$$\mathcal{O}_j = \sigma_{j_1} \otimes \sigma_{j_2} \otimes \cdots \otimes \sigma_{j_m} \qquad j_k \in \{I, X, Y, Z\}. \tag{3.11}$$

Definition 12. Given $\mathcal{O}_j$ as in (3.11), let $\mathcal{Q}_j$ be the subset of qubits which $\mathcal{O}_j$ acts non-trivially on:

$$\mathcal{Q}_j = \{k \in \{1, ..., m\} : \sigma_{j_k} \neq I\}. \tag{3.12}$$

We define the *Pauli weight* of $\mathcal{O}_j$ to be $|\mathcal{Q}_j|$ and the *maximal Pauli weight* of $\mathcal{O}$ to be

$$|\mathcal{Q}| = \max_{j \in \{1,...,p\}} |\mathcal{Q}_j|. \tag{3.13}$$

The Pauli weight is so named because each local observable is a tensor product of Pauli operators. In practice, when we perform the measurement described by $\mathcal{O}_j$ on the output state, we perform the measurement described by the Pauli operator σ_{j_k} on the qubit k for all $k \in \mathcal{Q}_j$. The identity terms do not contribute to the measurement.

Definition 13. Let $\mathcal{R}_k$ be the subset of local observables making up $\mathcal{O}$ which act non-trivially on the qubit k:

$$\mathcal{R}_k = \{j \in \{1, ..., p\} : \sigma_{j_k} \neq I\}. \tag{3.14}$$

We define the *qubit weight* of qubit k, relative to $\mathcal{O}$ to be $|\mathcal{R}_k|$ and the *maximal qubit weight* of $\mathcal{O}$ to be

$$|\mathcal{R}| = \max_{k \in \{1,...,m\}} |\mathcal{R}_k|. \tag{3.15}$$

Remark 3. *Clearly,*

$$k \in \mathcal{Q}_j \iff j \in \mathcal{R}_k \tag{3.16}$$

for all $k \in \{1, ..., m\}$ *and* $j \in \{1, ..., p\}$.

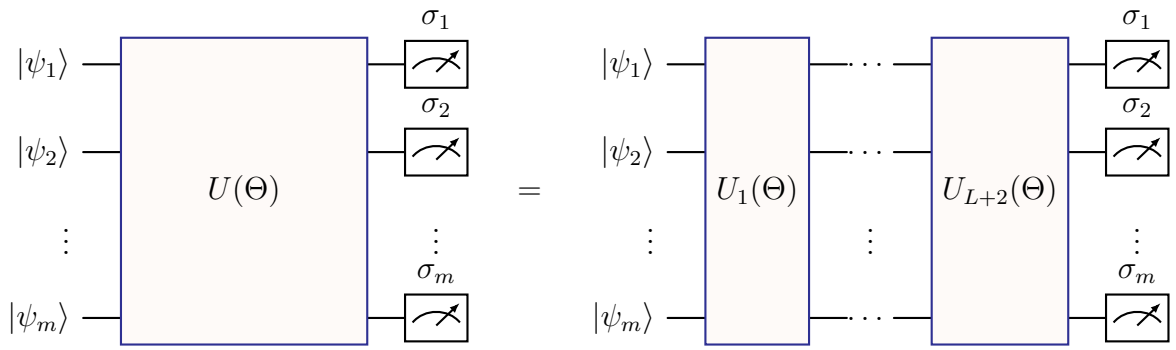

Figure 3.1: The circuit is a composition of layers.

Remark 4. *By Definition 13 we see that*

$$p = \left|\bigcup_{k=1}^{m} \mathcal{R}_k\right| \leq \sum_{k=1}^{m} |\mathcal{R}_k| \leq \sum_{k=1}^{m} |\mathcal{R}| = m|\mathcal{R}|. \tag{3.17}$$

Remark 5. *In [25, assumption 2.6] only observables of the following form are considered.*

$$\mathcal{O} = \sum_{k=1}^{m} \mathcal{O}_k \tag{3.18}$$

where for all $k \in \{1, ..., m\}$, $\mathcal{O}_k$ is an observable acting only on the qubit k:

$$\mathcal{O}_k = \mathbb{1}_1 \otimes \cdots \otimes \mathbb{1}_{k-1} \otimes \sigma_k \otimes \mathbb{1}_{k+1} \otimes \cdots \otimes \mathbb{1}_m \tag{3.19}$$

and $\sigma_k \in \{\sigma_X, \sigma_Y, \sigma_Z\}$. In other words, $p = m$, and $|\mathcal{Q}| = |\mathcal{R}| = W = 1$.

3.2 The circuit

So far we have simply described the circuit $U(\Theta)$ as a parametric unitary. In this section we will introduce some assumptions on the structure of the ansatz U and discuss their implications.

Fixing $L \geq 0$, the circuit $U(\Theta)$ is composed of $L + 2$ *layers*:

$$U(\Theta) = U_{L+2}(\Theta) U_{L+1}(\Theta) \cdots U_1(\Theta) \tag{3.20}$$

where each layer $U_l(\Theta)$ is a unitary gate that further decomposes into

$$U_l(\Theta) = V_l W_l(\Theta) \tag{3.21}$$

with the following structure imposed on V_l and W_l:

Assumption 1. *[25, definition 2.4]*

1. *V_l is a fixed unitary composed of one and two-qubit gates such that no more than one two-qubit gate acts on any single qubit.*

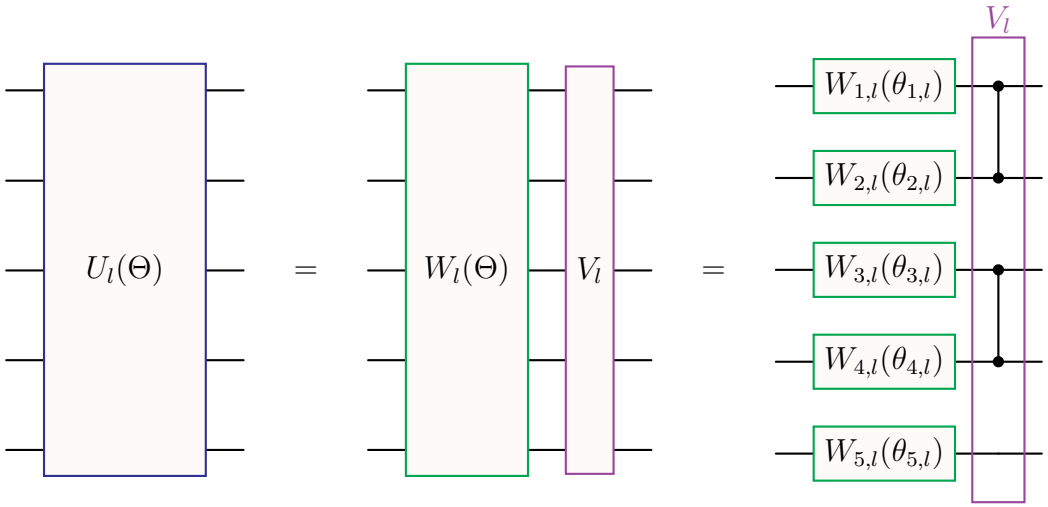

Figure 3.2: An example of one layer of a circuit with 5 qubits.

2. *$W_l(\Theta)$ is a tensor product of m single-qubit parametric gates, each depending on a single parameter.*

Remark 6. *For all $l \in \{1, ..., L\}$, V_l may depend on the input to the quantum neural network: $V_l(\hat{x})$, however, since $\hat{x}$ is fixed in this case, we consider V_l a fixed unitary.*

There are $L+2$ layers, each involving m parameters, therefore $|\Theta| = (L+2)m$. Noticing how, in a typical circuit diagram, the qubits are arranged in rows and the layers are arranged in columns, it is natural to enumerate the parameters using matrix notation:

$$\Theta = \begin{pmatrix} \theta_{1,1} & \cdots & \theta_{1,L+2} \\ \vdots & \ddots & \vdots \\ \theta_{m,1} & \cdots & \theta_{m,L+2} \end{pmatrix} \tag{3.22}$$

where $\theta_{k,l}$ parametrizes the single qubit unitary $W_{k,l}$ which acts on the qubit k in the layer l. We can then write a single layer as

$$U_l(\Theta) = V_l \left(W_{1,l}(\theta_{1,l}) \otimes \cdots \otimes W_{m,l}(\theta_{m,l}) \right). \tag{3.23}$$

On the other hand, it will be useful in certain cases to refer to a parameter or single qubit unitary by a single index $i \in \{1, ..., (L+2)m\}$ in which case, following [25, definition 2.5], we use the convention that

$$\left.\begin{aligned} \theta_{k,l} &= \theta_i \\ W_{k,l} &= W_i \end{aligned}\right\} \iff i = (l-1) \times m + k. \tag{3.24}$$

Assumption 2. *At initialization, each parameter is sampled independently, but not necessarily identically, from Ω. That is, $\theta_i \perp \theta_{i'}$ for all $i, i' \in \{1, ..., m(L+2)\}$ such that $i \neq i'$.*

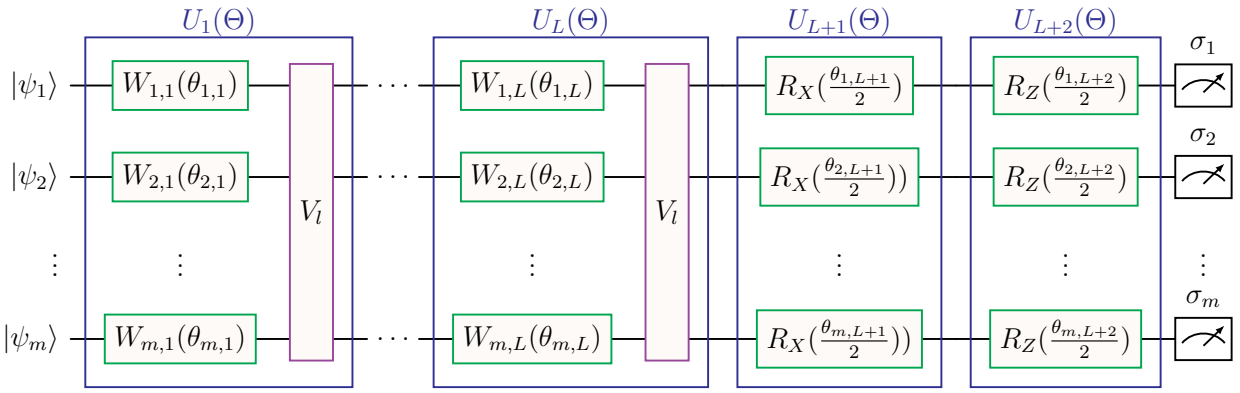

Figure 3.3: The Final Two Layers are made up of Pauli X and Pauli Z rotations.

Assumption 3. *[25, assumption 2.6] For all $i \in \{1, ..., (L+2)m\}$, $W_i(\theta_i)$ is given by the unitary evolution of a qubit generated by the Hamiltonian $\mathcal{G}_i$ over the time $\frac{\theta_i}{2}$, where $\mathcal{G}_i \in \mathcal{L}(\mathbb{C}^2)$ is Hermitian with spectrum in $\{-1, 1\}$:*

$$W_i(\theta_i) = e^{-i\frac{\theta_i}{2}\mathcal{G}_i}. \tag{3.25}$$

Remark 7. *[25, remark 2.7] Assumption 3 ensures that each W_i has period π up to an irrelevant sign, therefore we can restrict the domain of Θ to $\Omega = [0, \pi]^{(L+2)m}$.*

[25, assumption 2.24] enforces that the final layer of the circuit is chosen in such a way that $\mathbb{E}[f_k(\Theta)] = 0$ for all $k \in \{1, ..., m\}$, and an example of such a layer is given when the observable is a sum of single-qubit Pauli-Z observables, i.e. when

$$\mathcal{O} = \sum_{k=1}^{m} \mathbb{1}_1 \otimes \cdots \otimes \mathbb{1}_{k-1} \otimes \sigma_Z \otimes \mathbb{1}_{k+1} \otimes \cdots \mathbb{1}_m. \tag{3.26}$$

With assumption 4 we explicitly define the final two layers of $U(\Theta)$. In lemma 2, the proof of which is reserved for appendix A, we show that these layers achieve the same effect for *any* observable $\mathcal{O} \in \mathcal{L}(\mathcal{H})$.

Assumption 4. *V_{L+1} and V_{L+2} are the identity on $(\mathbb{C}^2)^{\otimes m}$ and*

$$W_{k,L+1}(\theta_{k,L+1}) = R_X(\theta_{k,L+1}) = e^{-i\frac{\theta_{k,L+1}}{2}\sigma_X} \tag{3.27}$$

$$W_{k,L+2}(\theta_{k,L+2}) = R_Z(\theta_{k,L+2}) = e^{-i\frac{\theta_{k,L+2}}{2}\sigma_Z} \tag{3.28}$$

for all $k \in \{1, ..., m\}$. Furthermore, the parameters used in the final two layers are independent and sampled uniformly from $[0, 2\pi]$ at initialization.

Lemma 2. *Assumption 4 ensures that at initialization*

1. *$\mathbb{E}[f_j(\Theta)] = 0$ for all $j \in \{1, ..., p\}$ such that $\mathcal{Q}_j \neq \emptyset$ and*
2. *$\mathrm{Cov}[f_j(\Theta), f_{j'}(\Theta)] = 0$ for all $j, j' \in \{1, ..., p\}$ such that $j \neq j'$.*

Proof. See Appendix A. □

Remark 8. *Suppose that there exists $j \in \{1, ..., p\}$ such that $\mathcal{Q}_j = \emptyset$. That would imply that $\mathcal{O}_j = \mathbb{1}_{\mathcal{H}}$ and therefore $f_j(\Theta) = \frac{w_j}{N}$ which does not depend on Θ. If we simply redefine $f(\Theta)$ to be*

$$f(\Theta) := \sum_{j'=1}^{p} f_{j'}(\Theta) - \frac{w_j}{N}, \tag{3.29}$$

then

$$\mathbb{E}[f(\Theta)] = \sum_{j':|\mathcal{Q}_{j'}|>0} \mathbb{E}[f_{j'}(\Theta)] + \frac{w_j}{N} - \frac{w_j}{N} = 0. \tag{3.30}$$

We can therefore assume, without loss of generality, that $\mathcal{Q}_j \neq \emptyset$ for all $j \in \{1, ..., p\}$ and therefore $\mathbb{E}[f(\Theta)] = 0$.

In order to have cleaner notation going forward we will sometimes refer to a parametric circuit $U(\Theta)$, layer $U_l(\Theta)$, or constituent thereof, $W_l(\Theta)$ or $W_{k,l}(\theta_{k,l})$, without explicit dependence on the parameters: $U, U_l, W_l, W_{k,l}$.

3.2.1 Light cones

In this section we describe how the dependence of the variables $f_j(\Theta)$ on the parameters is propagated backwards through the circuit as a result of qubit *interactions*.

Definition 14. Let $k, k' \in \{1, ..., m\}$ and $l \in \{1, ..., L+2\}$. We say that qubit k *interacts* with qubit k' in the layer l if V_l contains a two-qubit gate acting on qubits k and k'.

Following the notation of [25, section 2.3.1], we introduce some auxiliary sets which capture these interactions. Let $\mathcal{I}_l$ be a partition of the set $\{1, ..., m\}$ such that

- $\{k, k'\} \in \mathcal{I}_l$ if and only if the the qubits k and k' interact with each other in the layer l and
- $\{k\} \in \mathcal{I}_l$ if and only if qubit k does not interact with any other qubit in the layer l.

Let $\mathcal{I}_{l,k}$ be the unique element of $\mathcal{I}_l$ such that $k \in \mathcal{I}_{l,k}$; we note that if $\{k, k'\} \in \mathcal{I}_l$, then $\mathcal{I}_{l,k} = \mathcal{I}_{l,k'}$. Since $V_{L+1} = V_{L+2} = \mathbb{1}_{\mathcal{H}}$, there are no interactions in the last two layers and so $\mathcal{I}_{L+1,k} = \mathcal{I}_{L+2,k} = \{k\}$ for all $k \in \{1, ..., m\}$.

Remark 9. *[25, remark 2.16] According to the definition of $\mathcal{I}_l$ we can rewrite the unitary V_l as the composition*

$$V_l = \prod_{\{k\} \in \mathcal{I}_l} V_l^{(k)} \prod_{\{k,k'\} \in \mathcal{I}_l} V_l^{(k,k')} \tag{3.31}$$

where $V_l^{(k)}$ is a unitary acting only on qubit k and $V_l^{(k,k')}$ is a unitary acting on qubits k and k'. The order of the factors do not matter since each factor acts on a distinct Hilbert space and so they commute.

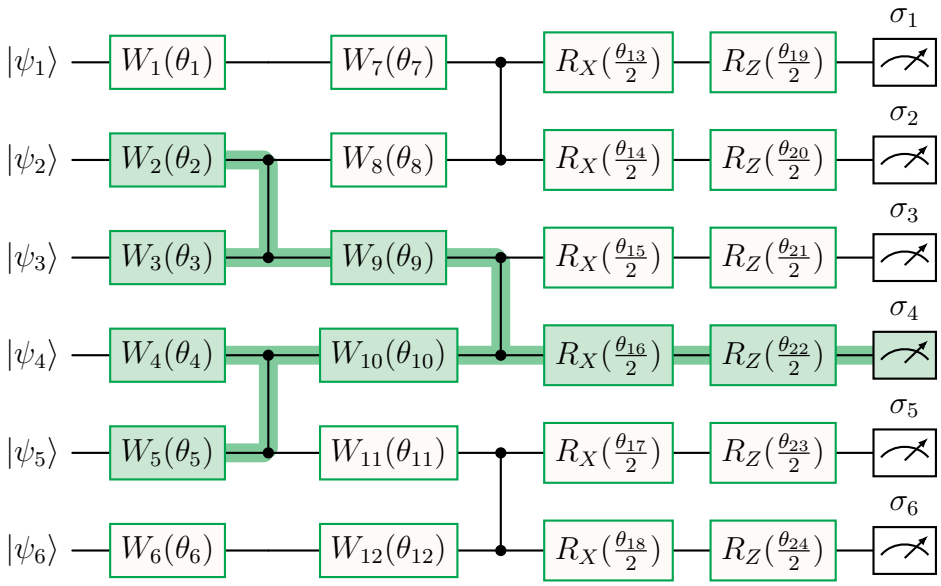

Figure 3.4: The past light cone of qubit 4 is $\mathcal{N}_6 = \{\theta_i : i \in \{2,3,4,5,9,10,16,22\}\}$. The corresponding auxiliary sets are $\mathcal{J}_4^1 = \{2,3,4,5\}, \mathcal{J}_4^2 = \{3,4\}, \mathcal{J}_4^3 = \{4\}$ and $\mathcal{J}_4^4 = \{4\}$.

For example, in Figure 3.2, $V_l = V_l^{(1,2)} \cdot V_l^{(3,4)} \cdot V_l^{(5)}$ where $V_l^{(1,2)}$ and $V_l^{(3,4)}$ are Controlled Z gates and $V_l^{(5)} = \mathbb{1}_5$.

The following sets, recursively defined, expand as we step backward through the circuit:

$$\mathcal{J}_k^l = \begin{cases} \mathcal{I}_{L,k} & \text{if } l = L+2 \\ \bigcup\limits_{k' \in \mathcal{J}_k^{l+1}} \mathcal{I}_{l,k'} & \text{if } l < L+2. \end{cases} \tag{3.32}$$

Given the lack of interactions in the final two layers it is implicit that $\mathcal{J}_k^{L+1} = \mathcal{J}_k^{L+2} = \{k\}$. At each decreasing layer $\mathcal{J}_k^l$ incorporates the indices of the qubits which interact with those in the preceding layer: $\mathcal{J}_k^1 \supseteq \mathcal{J}_k^2 \supseteq \cdots \supseteq \mathcal{J}_k^L \supseteq \mathcal{J}_k^{L+1} = \mathcal{J}_k^{L+2} = \{k\}$. Now to each $\mathcal{J}_k^l$ associate the set of parameters

$$\mathcal{N}_k^l = \bigcup_{k' \in \mathcal{J}_k^l} \{\theta_{k'l}\}. \tag{3.33}$$

Definition 15. [25, definition 2.12] For all $k \in \{1,...,m\}$, define the *past light cone* of qubit k as the subset of parameters given by

$$\mathcal{N}_k = \bigcup_{l=1}^{L+2} \mathcal{N}_k^l. \tag{3.34}$$

And for all $i \in \{1,...(L+2)m\}$, define the *future light cone* of the parameter θ_i as the subset of qubits given by

$$\mathcal{M}_i = \{k' \in \{1,...,m\} : \theta_i \in \mathcal{N}_{k'}\}. \tag{3.35}$$

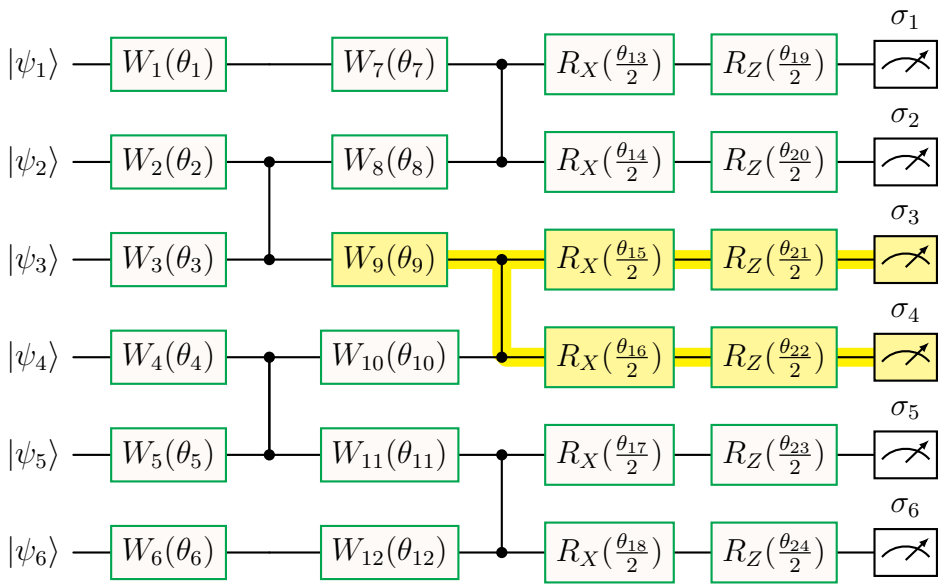

Figure 3.5: The future light cone of parameter θ_9 is $\mathcal{M}_9 = \{3, 4\}$.

Remark 10. *Clearly,*

$$\theta_i \in \mathcal{N}_k \iff k \in \mathcal{M}_i \tag{3.36}$$

for all $i \in \{1, ..., (L+2)m\}$ *and* $k \in \{1, ..., m\}$.

Definition 16. [25, definition 2.14] The *maximal cardinalities* of the past and future light cones are given by

$$|\mathcal{N}| = \max_{k \in \{1,...,m\}} |\mathcal{N}_k| \tag{3.37}$$

and

$$|\mathcal{M}| = \max_{i \in \{1,...,(L+2)m\}} |\mathcal{M}_i|. \tag{3.38}$$

Let us now clarify some subtleties between this work and [25]. In the previous paper the past light cone of qubit k is denoted $\mathcal{L}_k^p$ and defined as the subset of parameters on which the variable $f_k(\Theta)$ depends. They show that, since $\mathcal{O}_k$ acts non-trivially on only qubit k, $\mathcal{L}_k^p$ is a subset of $\mathcal{N}_k$ which they define as the *extended* past light cone of qubit k. Similarly, they denote by $\mathcal{L}_i^f$ the future light cone of the parameter θ_i and define it as the subset of local variables that depend on θ_i. They show that $\mathcal{L}_i^f$ is a subset of $\mathcal{M}_i$ which is named the *extended* future light cone of parameter θ_i.

Our case differs in that each observable may act non-trivially on an arbitrary number of qubits, therefore the dependency relations between variables and parameters cannot be characterized solely with what we refer to as light cones. In chapter 4 we will construct sets which capture these relationships using the $\mathcal{M}_i, \mathcal{N}_k, \mathcal{Q}_j$ and $\mathcal{R}_k$.

3.3 The Hilbert space of a local observable

Depending on the local observable $\mathcal{O}_j$ being measured, many of the gates making up the circuit do not contribute to the expected value of the measurement. Furthermore, only a subset of the qubits making up $\mathcal{H}$ are involved in the computation of $f_j(\Theta)$. We denote by $\mathcal{H}^j_{loc}$ the Hilbert space describing this subset of qubits. In this section we will illustrate which unitaries are unnecessary which will in turn inform a definition of $\mathcal{H}^j_{loc}$. This will be crucial to understanding both the interplay of the variables $f_j(\Theta)$ and the conditions under which we expect to gain a quantum advantage.

In [25, definition 2.17] the authors introduce for each $k \in \{1, ..., m\}$ a *pruning operation* which removes certain gates from a circuit and then show in [25, lemma 2.18] that pruning the circuit does not change the value of $f_k(\Theta)$. In [25, definition 2.21] they give a definition of $\mathcal{H}^k_{loc}$ and show in [25, lemma 2.22] that to compute $f_k(\Theta)$ requires only linear algebra operations in $\mathcal{H}^k_{loc}$. Finally, they estimate the dimension of $\mathcal{H}^k_{loc}$ in [25, lemma 2.23].

In definition 17 we introduce for each $j \in \{1, ..., p\}$ a similar operation which pares down the circuit $U(\Theta)$ according to the properties of $f(\Theta)_j$, taking into account that each $\mathcal{O}_j$ may act non-trivially on more than one qubit. In lemma 3 we similarly show that the pruning operation does not affect the value of $f_j(\Theta)$. In definition 18 we give a revised definition of $\mathcal{H}^j_{loc}$, show in lemma 4 that operations in $\mathcal{H}^j_{loc}$ are sufficient to compute $f_j(\Theta)$, and estimate the dimension of $\mathcal{H}^j_{loc}$ in lemma 5.

Definition 17. For each $j \in \{1, ..., p\}$ the *pruning operation* $[\,\cdot\,]_j$ acts on the circuit $U(\Theta)$ to produce the *pruned circuit* $[U(\Theta)]_j$. The pruned circuit is the composition of *pruned layers*:

$$[U(\Theta)]_j = \prod_{l=1}^{L+2} [U_l(\Theta)]_j = \prod_{l=1}^{L+2} [V_l]_j [W_l(\Theta)]_j. \tag{3.39}$$

The pruning operation on V_l is defined as

$$[V_l]_j = \prod_{\{k\} \in \mathcal{I}_l} [V_l^{(k)}]_j \prod_{\{k,k'\} \in \mathcal{I}_l} [V_l^{(k,k')}]_j \tag{3.40}$$

where

$$[V_l^{(k)}]_j = \begin{cases} \mathbb{1} & \text{if } k \notin \bigcup_{k' \in \mathcal{Q}_j} \mathcal{J}^l_{k'} \\ V_l^{(k)} & \text{otherwise} \end{cases} \tag{3.41}$$

$$[V_l^{(k,k')}]_j = \begin{cases} \mathbb{1} & \text{if } \{k,k'\} \in \mathcal{I}_l \text{ and } \{k,k'\} \cap \bigcup_{k'' \in \mathcal{Q}_j} \mathcal{J}^l_{k''} = \emptyset \\ V_l^{(k,k')} & \text{otherwise} \end{cases}. \tag{3.42}$$

And the pruning operation on $W_l(\Theta)$ is defined as

$$[W_l(\Theta)]_j = [W_{1,l}(\theta_{1,l})]_j \otimes \cdots \otimes [W_{1,m}(\theta_{1,m})]_j \tag{3.43}$$

where

$$[W_{k,l}(\theta_{k,l})]_j = \begin{cases} \mathbb{1} & \text{if } \theta_{k,l} \notin \bigcup_{k' \in \mathcal{Q}_j} \mathcal{N}^l_{k'} \\ W_{k,l}(\theta_{k,l}) & \text{otherwise} \end{cases}. \tag{3.44}$$

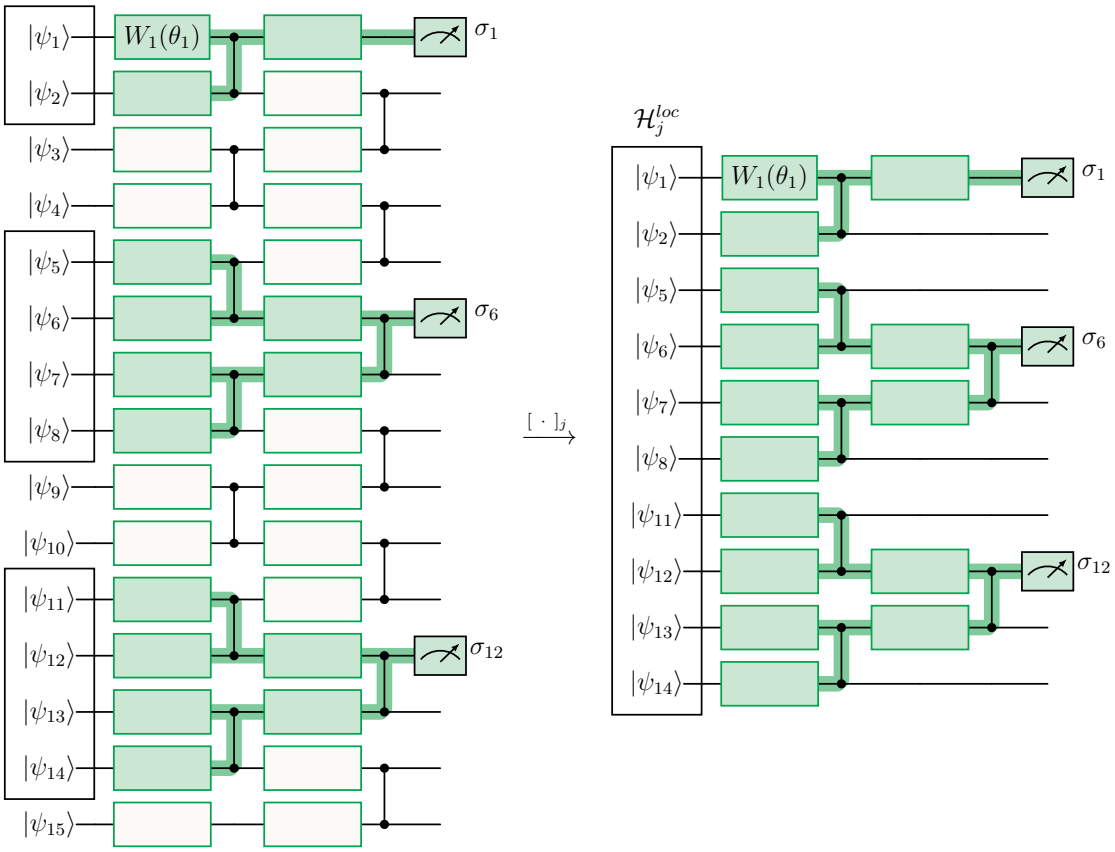

Figure 3.6: On the left is an example of a circuit $U(\Theta)$ with $m = 15$ and $L = 2$. On the right is the circuit $[U(\Theta)]_j$ pruned according to an observable $\mathcal{O}_j$ with $\mathcal{Q}_j = \{1, 6, 12\}$.

In other words, if $\mathcal{O}_j$ acts non-trivially on the qubit k and the parameter $\theta_{k',l}$ is in the past light cone of qubit k, then the pruning operation leaves unchanged $W_{k',l}$ and $V_l^{(k')}$ (or $V_l^{(k',k'')}$, whichever is appropriate). The rest of the unitaries are swapped with the identity.

Lemma 3. *For any $|\psi\rangle \in \mathcal{H}$ the following holds*

$$\langle\psi|[U(\Theta)]_j^\dagger \mathcal{O}_j [U(\Theta)]_j|\psi\rangle = \langle\psi|U^\dagger(\Theta)\mathcal{O}_j U(\Theta)|\psi\rangle. \tag{3.45}$$

Proof. We proceed by induction on the number of layers. First we show that (3.45) holds when $L = 0$. In this case the circuit consists of only the extra two layers: $U = W_2 W_1$, whence

$$\begin{aligned}\langle\psi|[U]_j^\dagger \mathcal{O}_j [U]_j|\psi\rangle &= \langle\psi|[W_1]_j^\dagger [W_2]_j^\dagger \mathcal{O}_j [W_2]_j [W_1]_j|\psi\rangle \\ &= \langle\psi|\left(\bigotimes_{k=1}^{m} [W_{k,1}]_j^\dagger [W_{k,2}]_j^\dagger \sigma_{j_k} [W_{k,2}]_j [W_{k,1}]_j\right)|\psi\rangle. \end{aligned} \tag{3.46}$$

If $k \notin \mathcal{Q}_j$, then $\sigma_{j_k} = I$ in which case

$$\begin{aligned}
[W_{k,1}]_j^\dagger [W_{k,2}]_j^\dagger \sigma_{j_k} [W_{k,2}]_j [W_{k,1}]_j &= ([W_{k,2}]_j [W_{k,1}]_j)^\dagger I [W_{k,2}]_j [W_{k,1}]_j \\
&= I \\
&= (W_{k,2} W_{k,1})^\dagger I W_{k,2} W_{k,1} \\
&= W_{k,1}^\dagger W_{k,2}^\dagger \sigma_{j_k} W_{k,2} W_{k,1}. \qquad (3.47)
\end{aligned}$$

On the other hand, suppose $k \in \mathcal{Q}_j$. Given that $\mathcal{J}_k^1 = \mathcal{J}_k^2 = \{k\}$, by definition, $\mathcal{N}_k^1 = \{\theta_{k,1}\}$ and $\mathcal{N}_k^2 = \{\theta_{k,2}\}$. Therefore $k \in \mathcal{Q}_j$ implies that

$$\begin{aligned}
\theta_{k,l} &\in \bigcup_{k' \in \mathcal{Q}_j} \{\theta_{k',l}\} \\
&= \bigcup_{k' \in \mathcal{Q}_j} \mathcal{N}_{k'}^l \qquad (3.48)
\end{aligned}$$

for $l = \{1, 2\}$ and thus

$$[W_{k,1}]_j^\dagger [W_{k,2}]_j^\dagger \sigma_{j_k} [W_{k,2}]_j [W_{k,1}]_j = W_{k,2}^\dagger W_{k,1}^\dagger \sigma_{j_k} W_{k,1} W_{k,2} \qquad (3.49)$$

trivially holds since $[W_{k,1}]_j = W_{k,1}$ and $[W_{k,2}]_j = W_{k,2}$. In either case, we have

$$\begin{aligned}
\langle \psi | [U]_j^\dagger \mathcal{O}_j [U]_j | \psi \rangle &= \langle \psi | \left(\bigotimes_{k=1}^m [W_{k,1}]_j^\dagger [W_{k,2}]_j^\dagger \sigma_{j_k} [W_{k,2}]_j [W_{k,1}]_j \right) | \psi \rangle \\
&= \langle \psi | \left(\bigotimes_{k=1}^m W_{k,1}^\dagger W_{k,2}^\dagger \sigma_{j_k} W_{k,2} W_{k,1} \right) | \psi \rangle \\
&= \langle \psi | U^\dagger \mathcal{O}_j U | \psi \rangle \qquad (3.50)
\end{aligned}$$

for all $|\psi\rangle \in \mathcal{H}$.

Assume now that the hypothesis holds when $L = L'$ for some $L' \geq 0$. Let U be a circuit with $L' + 2$ layers and U' the result of adding an additional layer U_0 to the beginning of U: $U' = UU_0$. Then for all $|\psi\rangle \in \mathcal{H}$, since $U_0|\psi\rangle \in \mathcal{H}$, by the inductive hypothesis,

$$\begin{aligned}
\langle \psi | U'^\dagger \mathcal{O}_j U' | \psi \rangle &= \langle \psi | U_0^\dagger U^\dagger \mathcal{O}_j U U_0 | \psi \rangle \\
&= \langle \psi | U_0^\dagger [U]_j^\dagger \mathcal{O}_j [U]_j U_0 | \psi \rangle \\
&= \langle \psi | W_0^\dagger V_0^\dagger [U]_j^\dagger \mathcal{O}_j [U]_j V_0 W_0 | \psi \rangle. \qquad (3.51)
\end{aligned}$$

We can split up V_0 into the product of terms which become the identity under the pruning operation and those which remain unchanged:

$$V_0 = V_0' V_0'' \quad \text{where} \quad [V_0']_j = V_0' \text{ and } [V_0'']_j = \mathbb{1}_{\mathcal{H}}. \qquad (3.52)$$

Clearly,

$$[V_0]_j = [V_0']_j [V_0'']_j = V_0' \cdot \mathbb{1}_{\mathcal{H}} = V_0'. \qquad (3.53)$$

If V_0 contains $V_0^{(k,k')}$, then

$$\{k,k'\} \in \mathcal{I}_l \text{ and } \{k,k'\} \cap \bigcup_{k'' \in \mathcal{Q}_j} \mathcal{J}^0_{k''} = \emptyset. \tag{3.54}$$

Since $\mathcal{J}^0_{k''} \supseteq \mathcal{J}^l_{k''}$ for all $l \in 1, ..., L+2$, we have also

$$\{k,k'\} \cap \bigcup_{k'' \in \mathcal{Q}_j} \mathcal{J}^l_{k''} = \emptyset, \quad l \in \{1, ..., L+2\}. \tag{3.55}$$

By the definition of the pruning operation, (3.55) implies $[U]_j$ acts as the identity on the qubits k and k' in which case $V_0^{(k,k')}$ commutes with $[U]_j$. And since $\mathcal{J}^{L+2}_{k''} = \{k''\}$, (3.55) implies that $\{k,k'\} \cap \mathcal{Q}_j = \emptyset$ in which case $V_0^{(k,k')}$ also commutes with $\mathcal{O}_j$. By the same logic, if $V_0^{(k)}$ divides V_0'', then $V_0^{(k)}$ commutes with $[U]_j$ and $\mathcal{O}_j$ and thus V_0'' commutes with $[U]_j$ and $\mathcal{O}_j$.

We can similarly express W_0 as

$$W_0 = W_0' W_0'' \quad \text{where} \quad [W_0']_j = W_0' \text{ and } [W_0'']_j = \mathbb{1}_{\mathcal{H}} \tag{3.56}$$

and thus $[W_0]_j = W_0'$.

If for some $k \in \{1, ..., m\}$, $\mathbb{1}_1 \otimes \cdots \otimes W_{k,0}(\theta_{k,0}) \otimes \cdots \otimes \mathbb{1}_m$ divides W_0'', then

$$\begin{aligned} \theta_{k,0} &\notin \bigcup_{k' \in \mathcal{Q}_j} \mathcal{N}^0_{k'} \\ &= \bigcup_{k' \in \mathcal{Q}_j} \left(\bigcup_{k'' \in \mathcal{J}^0_{k'}} \{\theta_{k''l}\} \right) \end{aligned} \tag{3.57}$$

implying $\nexists k' \in \mathcal{Q}_j$ such that $k \in \mathcal{J}^0_{k'}$. It follows that

$$k \notin \bigcup_{k' \in \mathcal{Q}_j} \mathcal{J}^l_{k'} \quad \forall l \in \{0, ..., L+2\} \tag{3.58}$$

therefore $\mathbb{1}_1 \otimes \cdots \otimes W_{k,0}(\theta_{k,0}) \otimes \cdots \otimes \mathbb{1}_m$ commutes with $V_0, [U]_j$ and $\mathcal{O}_j$, and thus so too does W_0''. Finally,

$$\begin{aligned} \langle \psi | U'^\dagger \mathcal{O}_j U' | \psi \rangle &= \langle \psi | W_0^\dagger V_0^\dagger [U]_j^\dagger \mathcal{O}_j [U]_j V_0 W_0 | \psi \rangle \\ &= \langle \psi | (W_0' W_0'')^\dagger (V_0' V_0'')^\dagger [U]_j^\dagger \mathcal{O}_j [U]_j V_0' V_0'' W_0' W_0'' | \psi \rangle \\ &= \langle \psi | W_0'^\dagger V_0'^\dagger [U]_j^\dagger \mathcal{O}_j [U]_j V_0' W_0' | \psi \rangle \\ &= \langle \psi | [W_0]_j^\dagger [V_0]_j^\dagger [U]_j^\dagger \mathcal{O}_j [U]_j [V_0]_j [W_0]_j | \psi \rangle \\ &= \langle \psi | [U_0]_j^\dagger [U]_j^\dagger \mathcal{O}_j [U]_j [U_0]_j | \psi \rangle \\ &= \langle \psi | [U']_j^\dagger \mathcal{O}_j [U']_j | \psi \rangle \end{aligned} \tag{3.59}$$

which shows that (3.45) holds when $L = L' + 1$ and thus for all $L \geq 0$. □

As a direct consequence of Lemma 3, we may define H^j_{loc} to be the Hilbert space describing the subset of qubits that are acted on non-trivially by the pruned circuit $[U]_j$.

Definition 18. Let $\mathcal{H}_k$ be the Hilbert space associated with qubit k. We define the *local Hilbert space* associated with the local observable $\mathcal{O}_j$ to be

$$\mathcal{H}^j_{loc} := \bigotimes_{k \in \bigcup_{k' \in \mathcal{Q}_j} \mathcal{J}^1_{k'}} \mathcal{H}_k. \tag{3.60}$$

Lemma 4. *The computation of $f_j(\Theta)$ only requires linear algebra in $\mathcal{H}^j_{loc}$.*

Proof. Suppose that $\mathcal{H}_k$ is not included in the tensor product that is $\mathcal{H}^j_{loc}$. Then $k \notin \bigcup_{k' \in \mathcal{Q}_j} \mathcal{J}^1_{k'}$. Since $\mathcal{J}^l_{k'} \subseteq \mathcal{J}^1_{k'}$ for all $l \geq 1$ we have that $k \notin \bigcup_{k' \in \mathcal{Q}_j} \mathcal{J}^1_{k'}$ for all $l \geq 1$. Then, by definition, $[U]_j$ acts as the identity on $\mathcal{H}_k$. By Lemma 3, operations on $\mathcal{H}_k$ are not needed in order to compute $f_j(\Theta)$. □

We now derive an upper bound on the dimension of the local Hilbert space given different assumptions on the circuit and observable.

Lemma 5. *For all $j \in \{1, ..., p\}$ it holds that*

$$\dim \mathcal{H}^j_{loc} \leq 2^{|\mathcal{Q}|2^L}. \tag{3.61}$$

Proof. Since $\dim \mathcal{H}_k = \dim \mathbb{C}^2 = 2$ for all $k \in \{1, ..., m\}$, the dimension of $\mathcal{H}^j_{loc}$ is given by 2^{m_j} where

$$m_j = \left| \bigcup_{k \in \mathcal{Q}_j} \mathcal{J}^1_k \right| \leq |\mathcal{Q}_j| \max_{k \in \mathcal{Q}_j} |\mathcal{J}^1_k| \leq |\mathcal{Q}| \max_{k \in \{1,...,m\}} |\mathcal{J}^1_k|. \tag{3.62}$$

We notice that for all $k \in \{1, ..., m\}$,

$$|\mathcal{J}^1_k| = \left| \bigcup_{k' \in \mathcal{J}^2_k} \mathcal{I}_{1,k'} \right| \leq \sum_{k' \in \mathcal{J}^2_k} |\mathcal{I}_{1,k'}| \leq \sum_{k' \in \mathcal{J}^2_k} 2 = 2|\mathcal{J}^2_k| \tag{3.63}$$

where we have used the fact that $|\mathcal{I}_{l,k}| \in \{1, 2\}$ for all $l \in \{1, ..., L+2\}$ and $k \in \{1, ..., m\}$ by Assumption 1. Inductively, we can infer the relation

$$|\mathcal{J}^1_k| \leq 2^l |\mathcal{J}^{l+1}_k| \tag{3.64}$$

for all $l \leq L$. Since $\mathcal{J}^{L+1}_k = \{k\}$, we have that $|\mathcal{J}^1_k| \leq 2^L |\mathcal{J}^{L+1}_k| = 2^L$. Since this is true for all $k \in \{1, ..., m\}$, we conclude that $m_j \leq |\mathcal{Q}| 2^L$ and thus $\dim \mathcal{H}^j_{loc} \leq 2^{|\mathcal{Q}|2^L}$. □

Lemma 5 makes no assumption on the layout of the qubits and no restrictions are imposed on which qubits may interact with each other. If, for example, the qubits are arranged in a d-dimensional lattice and $U(\Theta)$ is such that entangling gates may only act on adjacent qubits, then

$$|\mathcal{J}^1_k| \in O(L^d), \quad k \in \{1, ..., m\} \tag{3.65}$$

implying the improved bound

$$\dim \mathcal{H}_{loc}^{j} \in O(2^{|\mathcal{Q}|L^d}), \quad j \in \{1, ..., p\}. \tag{3.66}$$

Such a circuit is called *geometrically local*. The restriction of geometric locality is often imposed on a circuit due to limitations inherent to the physical implementation of the qubits.

Another particular case is when each of the local observables $\mathcal{O}_j$ act non-trivially on a set of not more than h adjacent qubits. In this case, we say that $\mathcal{O}$ is *spatially h-local* [40]. If $\mathcal{O}$ is spatially h-local, and $U(\Theta)$ is 1-dimensional and geometrically local, then we can find further improved bounds on the dimension of the local Hilbert space. In this configuration, for all $j \in \{1, ..., p\}$, $\mathcal{Q}_j = \{a, a+1, ..., a+h\}$ for some $a \in \{1, ..., m-h\}$ and $\mathcal{J}_k^1 \subseteq \{k-L, ..., k+L\}$ for all $k \in \{1, ..., m\}$. This implies

$$\begin{aligned} \bigcup_{k \in \mathcal{Q}_j} \mathcal{J}_k^1 &\subseteq \bigcup_{k \in \{a, a+1, ..., a+h\}} \{k - L, ..., k + L\} \\ &= \{a - L, ..., a + h + L\}, \end{aligned} \tag{3.67}$$

and thus

$$\dim \mathcal{H}_{loc}^{j} \leq 2^{h+2L}. \tag{3.68}$$

We will further examine these particular cases in chapter 4, however, it is important to note that we do not assume that $U(\Theta)$ is geometrically local, nor that $\mathcal{O}$ is spatially local unless otherwise specified.

3.3.1 Classical simulability

If we are able to efficiently simulate a quantum circuit with a classical device, then we gain no advantage in using a quantum computer. The most direct method of simulation is the state-vector approach where the initial state of an m qubit system is encoded in a 2^m dimensional vector with each entry representing the amplitude of a corresponding basis state. The application of unitary gates to the initial state is simulated via matrix multiplication of the state vector. Under this scheme, in the worst case, the run time of the circuit simulation is no more than $Cg2^m$ where g is the number of elementary gates composing the circuit, and C is a constant not depending on the circuit [41].

By Lemmas 4 and 5 it is possible to simulate the computation of $f_j(\Theta)$ using a classical computer with

$$O(L \dim \mathcal{H}_{loc}^{j}) = O(L2^{|\mathcal{Q}|2^L}) \tag{3.69}$$

operations, and thus to compute $f(\Theta)$, $O(pL2^{|\mathcal{Q}|2^L})$ operations are needed. In the case that $U(\Theta)$ is geometrically local, $O(pL2^{|\mathcal{Q}|L^d})$ operations are needed, and in the case that $U(\Theta)$ is 1-dimensional geometrically local, and $\mathcal{O}$ is spatially h-local, $O(pL2^{h+2L})$ operations are needed. In section 4.3 we will discuss specific examples of $|\mathcal{Q}|$ and L and see whether or not they afford the possibility of a quantum advantage.

Chapter 4

Convergence to a Gaussian distribution

In chapter 3 we introduced the random variable $f(\Theta)$ generated by a randomly initialized quantum neural network with a fixed input. In this chapter we aim to assess how well the law of this variable approximates the Gaussian distribution.

Suppose we have a sequence of randomly initialized quantum neural networks with diverging width and denote by $f^{(m)}(\Theta)$ the variable generated by the circuit with width m. In [25, Theorem 3.14] the author shows that if $|\mathcal{Q}|= |\mathcal{R}|= W = 1$ and

$$\lim_{m\to 0} \frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3} = 0, \tag{4.1}$$

then $f^{(m)}(\Theta)$ converges in distribution to a standard Gaussian random variable as $m \to \infty$. [1] In this chapter we aim to answer the question: How "fast" does $f^{(m)}(\Theta)$ converge? Answering this amounts to, for a fixed m, determining the "distance" between the law of $f^{(m)}(\Theta)$ and that of its limiting value. The measure we use to quantify this is the *Kolmogorov distance*. This metric is widely used to quantify the rate of convergence in the context of Gaussian approximation. A prominent example of such an application is the Berry-Esseen theorem (see Theorem 2) which provides a bound on the rate of convergence in the Central Limit Theorem (CLT) (see Theorem 1). Just as the authors of [25] take inspiration from the CLT to show convergence of $f^{(m)}(\Theta)$, we take inspiration from the Berry-Esseen theorem to find the rate at which this convergence takes place.

In section 4.1 we introduce the Berry-Esseen theorem along with some necessary preliminaries. In section 4.2 we state and prove the main result of this thesis: an upper bound on the rate of convergence of $f^{(m)}(\Theta)$ to a standard Gaussian random variable. In section 4.3 we discuss how we expect this upper bound to behave for certain dependencies of L and $|\mathcal{Q}|$ on m. Finally, in section 4.4 we compare our bound with a result previously established in [26] based on the same framework. We stress that our result includes observables in which $|\mathcal{Q}|, |\mathcal{R}|, W > 1$ which is a scenario not considered by [25] or [26]. In what follows, we return to the notation $f(\Theta)$, omitting the dependency on m.

[1] In fact, a stronger result is proved: the random function $\{f^{(m)}(x, \Theta)\}_{x\in\mathcal{X}}$ converges in distribution to a mean zero Gaussian process.

4.1 The Berry-Esseen theorem

Definition 19. The *Gaussian* distribution with mean μ and variance σ^2, denoted $\mathcal{N}_{\mu,\sigma^2}$, is the distribution with density

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{4.2}$$

$\mathcal{N}_{0,1}$ is called the *standard Gaussian* distribution.

The Gaussian distribution plays a central role in statistics due in part to its connection with various convergence theorems - most notably the CLT. In the context of random variables there are multiple notions of convergence. In this work we use *convergence in distribution*.

Definition 20. Let (X_n) be a sequence of random variables with CDFs $(F_n(x))$ and X a random variable with CDF $F(x)$. If

$$\lim_{n\to\infty} F_n(x) = F(x) \tag{4.3}$$

for every point x at which $F_n(x)$ is continuous, then (X_n) *converges in distribution* to X and we say $X_n \xrightarrow{d} X$.

This type of convergence is the subject of the CLT. Its classic version is as follows.

Theorem 1 (Central Limit Theorem)**.** *Let (X_i) be a sequence of independently and identically distributed random variables with $\mathbb{E}[X_1] = \mu$ and $\mathbb{V}[X_1] = \sigma^2 < \infty$. Let (S_n) be the sequence defined by*

$$S_n = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu). \tag{4.4}$$

Then

$$S_n \xrightarrow{d} Z, \quad Z \sim \mathcal{N}_{0,1} \tag{4.5}$$

as $n \to \infty$.

Proof. See [42]. □

There are several metrics one can use to measure the difference between probability measures. In this work we use the Kolmogorov distance which is well suited for quantifying the rate of convergence in distribution.

Definition 21. Let P_X and P_Y be probability measures on $\mathbb{R}$, with $F_X(x)$ and $F_Y(x)$ their corresponding CDFs. The *Kolmogorov distance* between P_X and P_Y is

$$d_K(P_X, P_Y) := \sup_{x\in\mathbb{R}} |F_X(x) - F_Y(x)|. \tag{4.6}$$

Intuitively, the Kolmogorov distance between two probability measures gives the maximum discrepancy between the probabilities assigned to an event by the two measures.

Remark 11. *Convergence in Kolmogorov distance implies convergence in distribution as it implies uniform convergence of the CDF, a stronger condition than pointwise convergence which is required for convergence in distribution.*

Remark 12. $d_K(P_X, P_Y) \leq 1$ *for all probability measures* P_X *and* P_Y *on* $\mathbb{R}$.

Theorem 2 (Berry-Esseen)**.** *Let* (X_i) *be a sequence of independently and identically distributed random variables with* $\mathbb{E}[X_1] = 0, \mathbb{V}[X_1] = \sigma^2$ *and* $\mathbb{E}[|X_1|^3] = \rho < \infty$. *Let* S_n *be the random variable defined by*

$$S_n = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} X_i \tag{4.7}$$

and P_{S_n} *the law of* S_n. *There exists a constant* C *such that*

$$d_K(P_{S_n}, \mathcal{N}_{0,1}) \leq \frac{C\rho}{\sigma^2\sqrt{n}}. \tag{4.8}$$

Proof. See [43, 44]. □

More generally, a Berry-Esseen type bound is any result that provides a similar quantitative bound on the rate of convergence of a sequence of random variables to its limit, usually under weaker but somewhat similar conditions to that of Theorem 2. A widely used approach to establishing these type of bounds is via the following result.

Theorem 3. *Suppose that* F *is a distribution function, that* $G : \mathbb{R} \to \mathbb{R}$ *satisfies*

$$G(-\infty) = \lim_{x\to-\infty} G(x) = 0 \quad \textit{and} \quad G(\infty) = \lim_{x\to\infty} G(x) = 1, \tag{4.9}$$

that G *is differentiable and of bounded variation and that its derivative satisfies*

$$M = \sup_{x\in\mathbb{R}} |G'(x)| < \infty \tag{4.10}$$

and that

$$\int_{\mathbb{R}} |F - G| dx < \infty. \tag{4.11}$$

Write

$$\Delta = \frac{1}{2M} \sup_{x\in\mathbb{R}} |F(x) - G(x)| \tag{4.12}$$

and

$$\varphi_f(t) = \int_{\mathbb{R}} e^{itx} dF(x), \quad \varphi_g(t) = \int_{\mathbb{R}} e^{itx} dG(x). \tag{4.13}$$

Then for all $T > 0$,

$$\Delta \leq \frac{1}{\pi M} \int_0^T \frac{|\varphi_f(t) - \varphi_g(t)|}{t} dt + \frac{12}{\pi T}. \tag{4.14}$$

Proof. See [43, 44]. □

The preceding theorem introduces the function

$$\varphi_f(t) = \int_{\mathbb{R}} e^{itx} dF(x). \tag{4.15}$$

If F is the CDF of a random variable, then $\varphi_f(t)$ takes on a special interpretation - it is the *characteristic function* of that variable.

Definition 22. The *characteristic function* of a random variable X with values in $\mathbb{R}$ is the function $\varphi_X(t) : \mathbb{R} \to \mathbb{C}$ defined by

$$\varphi_X(t) := \mathbb{E}[e^{itX}]. \tag{4.16}$$

The characteristic function of $Z \sim \mathcal{N}_{0,1}$ is

$$\varphi_Z(t) = e^{-\frac{1}{2}t^2}. \tag{4.17}$$

Theorem 3 demonstrates that for all random variables $X \sim P_X$ a bound on $|\varphi_X(t) - \varphi_Z(t)|$ directly translates to a corresponding bound on $d_K(P_X, \mathcal{N}_{0,1})$. This is unsurprising due to the following theorem which provides an equivalence between convergence in distribution and convergence of the characteristic function.

Theorem 4 (Lévy's Theorem)**.** *Let (X_n) and (φ_{X_n}) be a sequence of random variables in $\mathbb{R}$ and the corresponding sequence of characteristic functions, respectively. If there exists a random variable X such that*

$$\varphi(t) = \lim_{n\to\infty} \varphi_{X_n}(t), \quad t \in \mathbb{R} \tag{4.18}$$

is the characteristic function of X, then $X_n \xrightarrow{d} X$.

Proof. See [42]. □

When X is a sum of independently and identically distributed random variables, the bound on the difference between the characteristic functions is favorable and easy to derive. While $f(\Theta)$ *is* a sum of random variables, the variables *are not* independent, nor are they necessarily identically distributed. However, if the circuit is wide and shallow enough, then the variables, as we will see, exhibit weak dependence allowing us to, in a similar manner, leverage Theorem 3 for the purpose of deriving Berry-Esseen type bounds for $f(\Theta)$.

4.2 Rate of convergence of the output function at initialization

We are now ready to present the main result of this thesis: an upper bound on the Kolmogorov distance between the law of $f(\Theta)$ and the standard Gaussian distribution. This

bound is given in terms of the quantities defined in chapter 3. For convenience we summarize them below.

m : number of qubits of the circuit
W : maximum coefficient of any summand of $\mathcal{O}$ in the Pauli basis
$|\mathcal{Q}|$: maximal Pauli weight of $\mathcal{O}$
$|\mathcal{R}|$: maximal qubit weight of $\mathcal{O}$
$|\mathcal{N}|$: maximal cardinality of the past light cone of a qubit
$|\mathcal{M}|$: maximal cardinality of the future light cone of a parameter
N : normalization constant
L : number of layers of the circuit

Thanks to [25, Theorem 3.14] we know that for certain choices of these quantities, if the number of qubits utilized by the circuit is very large, then the law of $f(\Theta)$ is approximately $\mathcal{N}_{0,1}$. Our result (Theorem 5) will tell us, for any choice of these quantities, how accurate this approximation is. More precisely, it gives an upper bound on the maximum error in probability ascribed by $\mathcal{N}_{0,1}$ to the outcome represented by $f(\Theta)$. Conversely, if one were to define a variable width architecture in which these quantities were a function of m, then this bound determines, for a desired precision $\epsilon > 0$, how many qubits are required to ensure that the error is not more than ϵ.

Theorem 5. *Let $P_{f(\Theta)}$ be the law of $f(\Theta)$. If assumptions 1, 2, 3 and 4 hold, then*

$$d_K(P_{f(\Theta)}, \mathcal{N}_{0,1}) \leq \frac{64}{\pi}\left(\sqrt{\frac{6(eW)^3}{\sqrt{2\pi}}\frac{m|\mathcal{M}|^2|\mathcal{N}|^2|\mathcal{Q}|^2|\mathcal{R}|^3}{N^3}} + \frac{96(eW)^3}{\sqrt{2\pi}}\frac{m|\mathcal{M}|^2|\mathcal{N}|^2|\mathcal{Q}|^2|\mathcal{R}|^3}{N^3}\right). \tag{4.19}$$

Corollary 1. *If assumptions 1, 2, 3 and 4 hold and*

$$\lim_{m\to\infty} \frac{m|\mathcal{M}|^2|\mathcal{N}|^2|\mathcal{Q}|^2|\mathcal{R}|^3}{N^3} = 0 \tag{4.20}$$

then

$$f^{(m)}(\Theta) \xrightarrow{d} Z, \quad Z \sim \mathcal{N}_{0,1} \tag{4.21}$$

as $m \to \infty$.

Proof. An immediate consequence of (4.20) and Theorem 5 is that

$$\lim_{m\to\infty} d_K(P_{f(\Theta)}, \mathcal{N}_{0,1}) = 0 \tag{4.22}$$

which implies $f^{(m)}(\Theta) \xrightarrow{d} Z$ as $m \to \infty$ by remark 11. □

4.2.1 Proof of Theorem 5

Let $F(x)$ and $\Phi(x)$ be the CDFs of $f(\Theta)$ and $Z \sim \mathcal{N}_{0,1}$, respectively:

$$F(x) = \mathbb{P}(f(\Theta) \leq x), \quad \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy. \tag{4.23}$$

Then

$$d_K(P_{f(\Theta)}, \mathcal{N}_{0,1}) = \sup_{x \in \mathbb{R}} |F(x) - \Phi(x)|. \tag{4.24}$$

Our strategy involves applying Theorem 3 to F and Φ. We proceed by following the three steps outlined below, each of which is presented in its own subsection.

1. We use the properties of the Gaussian distribution to show that $F(x)$ and $\Phi(x)$ meet the criteria of Theorem 3, implying that

$$d_K(P_{f(\Theta)}, \mathcal{N}_{0,1}) \leq \frac{2}{\pi} \int_0^T \frac{|\varphi_f(t) - \varphi_Z(t)|}{t} dt + \frac{24M}{\pi T} \tag{4.25}$$

   for all $T > 0$ where $\varphi_f(t)$ and $\varphi_Z(t)$ are the characteristic functions of $f(\Theta)$ and Z.

2. We use the properties of $f(\Theta)$ to find a function $\epsilon(t)$ such that $|\varphi_f(t) - \varphi_Z(t)| \leq \epsilon(t)$ in a neighborhood of 0. This involves quantifying the extent to which the $f_j(\Theta)$ are dependent on one another and it accounts for the bulk of the work. Much of this section follows the work of [25, chapter 3], in which the convergence of $f(\Theta)$ is proved via Levy's theorem (Theorem 4) by showing that $\varphi_f(t)$ converges pointwise to $\varphi_Z(t)$. Once again we will highlight any important distinctions.

3. We estimate the integral

$$\int_0^T \frac{\epsilon(t)}{t} dt \tag{4.26}$$

   and optimize (4.25) over T.

The distribution functions

By definition, F is a distribution function and so is Φ which means that Φ satisfies condition (4.9). It is non-decreasing on $\mathbb{R}$ thus its total variation is given by

$$\Phi(\infty) - \Phi(-\infty) = 1 - 0 = 1 < \infty$$

and is therefore of bounded variation. We have

$$\begin{aligned} M &= \sup_{x \in \mathbb{R}} |\Phi'(x)| \\ &= \sup_{x \in \mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \\ &= \frac{1}{\sqrt{2\pi}} < \infty. \end{aligned} \tag{4.27}$$

Furthermore, we claim F and Φ satisfy (4.11).

Lemma 6.

$$\int_{-\infty}^{\infty} |F(x) - \Phi(x)| dx \leq \frac{4pW}{N} + \frac{\sqrt{2}}{\pi} \tag{4.28}$$

Proof. Recall that by Lemma 1, $|f(\Theta)| \leq \frac{pW}{N}$ which implies that $F(x) = 0$ for all $x < -\frac{pW}{N}$ and $F(x) = 1$ for all $x > \frac{pW}{N}$. It follows that

$$\begin{aligned} & \int_{-\infty}^{\infty} |F(x) - \Phi(x)| dx \\ = & \int_{-\infty}^{\frac{pW}{N}} |\Phi(x)| dx + \int_{-\frac{pW}{N}}^{\frac{pW}{N}} |F(x) - \Phi(x)| dx + \int_{\frac{pW}{N}}^{\infty} |1 - \Phi(x)| dx. \end{aligned} \tag{4.29}$$

Since F and Φ are distribution functions, they are upper bounded by one, thus

$$\begin{aligned} \int_{-\frac{pW}{N}}^{\frac{pW}{N}} |F(x) - \Phi(x)| dx &\leq \int_{-\frac{pW}{N}}^{\frac{pW}{N}} |F(x)| + |\Phi(x)| dx \\ &\leq \int_{-\frac{pW}{N}}^{\frac{pW}{N}} 2dx \\ &= \frac{4pW}{N}. \end{aligned} \tag{4.30}$$

It is useful now to introduce the *error function* which is a function $erf : \mathbb{R} \to \mathbb{R}$ defined as

$$erf(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt, \tag{4.31}$$

and the *complimentary error function*, a function $erfc : \mathbb{R} \to \mathbb{R}$ defined as

$$erfc(z) = 1 - erf(z). \tag{4.32}$$

We recall, without proof, that Φ can be expressed as a function of erf:

$$\Phi(x) = \frac{1}{2} + \frac{1}{2} erf\left(\frac{x}{\sqrt{2}}\right), \tag{4.33}$$

that erf is odd: $erf(-x) = -erf(x)$, that $erfc(x) \geq 0$ for all $x \in \mathbb{R}$, and that

$$\int_0^{\infty} erfc(x) = \frac{1}{\pi}. \tag{4.34}$$

We now use these properties to bound the tails of our integral.

$$\begin{aligned}
&\int_{-\infty}^{-\frac{pW}{N}} |\Phi(x)|dx + \int_{\frac{pW}{N}}^{\infty} |1-\Phi(x)|dx \\
= \;& \frac{1}{2}\int_{-\infty}^{-\frac{pW}{N}} \left|1+erf\left(\frac{x}{\sqrt{2}}\right)\right| dx + \frac{1}{2}\int_{\frac{pW}{N}}^{\infty} \left|1-erf\left(\frac{x}{\sqrt{2}}\right)\right| dx \\
= \;& \int_{\frac{pW}{N}}^{\infty} \left|1-erf\left(\frac{x}{\sqrt{2}}\right)\right| dx \\
= \;& \int_{\frac{pW}{N}}^{\infty} \left|erfc\left(\frac{x}{\sqrt{2}}\right)\right| dx \\
\leq \;& \int_{0}^{\infty} \left|erfc\left(\frac{x}{\sqrt{2}}\right)\right| dx \\
= \;& \sqrt{2}\int_{0}^{\infty} |erfc(x)|dx \\
= \;& \frac{\sqrt{2}}{\pi}. \qquad (4.35)
\end{aligned}$$

All together,

$$\int_{-\infty}^{\infty} |F(x)-\Phi(x)|dx \leq \frac{4pW}{N} + \frac{\sqrt{2}}{\pi}. \qquad (4.36)$$

□

With that we have shown that F and Φ satisfy the hypotheses of Theorem 3, and therefore

$$\sup_{x\in\mathbb{R}} |F(x)-\Phi(x)| \leq \frac{2}{\pi}\int_0^T \frac{|\varphi_f(t)-\varphi_Z(t)|}{t}dt + \frac{24}{\pi\sqrt{2\pi}T} \qquad (4.37)$$

for all $T > 0$.

The characteristic function

Without knowledge of $U(\Theta)$ or $\mathcal{O}$ it is impossible to find an expression for $\varphi_f(t)$, and even with knowledge of $U(\Theta)$ and $\mathcal{O}$ it may be very hard. Thankfully, we need not find $\varphi_f(t)$ itself - an approximation of $|\varphi_f(t)-\varphi_Z(t)|$ will do. For this we use the method of *cumulants*, a well known tool for comparing random variables with Gaussian ones [45].

Definition 23. The *cumulant generating function* of a real valued random variable X is

$$\mathcal{K}_X(t) := \log \varphi_X(t). \qquad (4.38)$$

The Maclaurin series of the cumulant generating function is

$$\sum_{r=0}^{\infty} \frac{\mathcal{K}_r(X)}{r!}(it)^r \qquad (4.39)$$

where for all $r \in \mathbb{N}$,

$$\mathcal{K}_r(X) = (-i)^r \frac{d^r}{dt^r} \log \mathbb{E}[e^{itX}]|_{t=0} \tag{4.40}$$

is the r^{th} *cumulant* of X.

The first and second cumulants of any random variable are its mean and variance, respectively: $\mathcal{K}_1(X) = \mathbb{E}[X]$ and $\mathcal{K}_2(X) = \mathbb{V}[X]$. This tells us that the first two cumulants of $f(\Theta)$ and Z accord since

$$\mathbb{E}[f(\Theta)] = \mathbb{E}[Z] = 0 \tag{4.41}$$

by Lemma 2, and

$$\mathbb{V}[f(\Theta)] = \mathbb{V}[Z] = 1. \tag{4.42}$$

The third and higher order cumulants of Z are zero as evidenced by the fact that

$$\log \varphi_Z(t) = -\frac{1}{2}t^2. \tag{4.43}$$

We can write

$$\begin{aligned} & \log \varphi_f(t) = -\frac{1}{2}t^2 + T(t) \\ \Rightarrow \quad & \varphi_f(t) = e^{-\frac{1}{2}t^2 + T(t)} \end{aligned} \tag{4.44}$$

where $T(t)$ is the tail of $\mathcal{K}_{f(\Theta)}(t)$:

$$T(t) := \sum_{r=3}^{\infty} \frac{\mathcal{K}_r(f(\Theta))}{r!}(it)^r. \tag{4.45}$$

Notice that

$$\begin{aligned} |\varphi_f(t) - \varphi_Z(t)| &= |e^{-\frac{1}{2}t^2 + T(t)} - e^{-\frac{1}{2}t^2}| \\ &= e^{-\frac{1}{2}t^2}|e^{T(t)} - 1| \\ &= e^{-\frac{1}{2}t^2}(e^{|T(t)|} - 1). \end{aligned} \tag{4.46}$$

This tells us that bounding $|\varphi_f(t) - \varphi_Z(t)|$ amounts to bounding $|T(t)|$ which can be done by bounding the third and higher order cumulants of $f(\Theta)$. Intuitively, this makes sense, since the third and higher order cumulants of Z are zero. To do so we use the following theorem:

Theorem 6 ([46])**.** *For any integer $r \geq 1$, there exists a constant C_r with the following property. Let $\{X_\alpha\}_{\alpha \in V}$ be a family of random variables with dependency graph G. We denote with $|V|$ the number of vertices of G and D the maximal degree of G. Assume that the variables X_α are uniformly bounded by a constant A. Then, if*

$$X = \sum_{\alpha \in V} X_\alpha, \tag{4.47}$$

one has

$$|\mathcal{K}_r(X)| \leq C_r |V| (D+1)^{r-1} A^r \tag{4.48}$$

with $C_r = 2^{r-1} r^{r-2}$.

Corollary 2. *Assuming the hypotheses of Theorem 6 hold,*

$$|\mathcal{K}_r(X)| \leq \frac{|V|}{D}(4eAD)^r r!\,. \tag{4.49}$$

Proof. For all $|V|, D, A > 0$ and $r > 1$,

$$\begin{aligned} 2^{r-1}r^{r-2}|V|(D+1)^{r-1}A^r &= \frac{|V|(2rA(D+1))^r}{2r^2(D+1)} \\ &\leq \frac{|V|}{D}(2rA(D+1))^r \\ &\leq \frac{|V|}{D}(4rAD)^r. \end{aligned} \tag{4.50}$$

Using the fact that $r^r \leq e^r r!$ we conclude

$$|\mathcal{K}_r(X)| \leq \frac{|V|}{D}(4eAD)^r r!\,. \tag{4.51}$$

□

$f(\Theta)$ is the sum of p random variables that are uniformly bounded by $\frac{W}{N}$. Therefore we can equate V in Theorem 6 with the set $\{1, ..., p\}$, and A with $\frac{W}{N}$. What we are missing is a *dependency graph* for the collection $\{f_j(\Theta)\}_{j=1}^p$.

Definition 24. Let $\{X_\alpha\}_{\alpha \in V}$ be a collection of random variables indexed by some set V. A *dependency graph* for $\{X_\alpha\}_{\alpha \in V}$ is a graph $\mathcal{G} = (V, E)$ such that the following property holds: whenever V_1 and V_2 are disjoint subsets of V such that there are no edges in $\mathcal{G}$ with one end in V_1 and one in V_2, the collections $\{X_\alpha\}_{\alpha \in V_1}$ and $\{X_\alpha\}_{\alpha \in V_2}$ are independent.

To apply Theorem 6 we must construct a dependency graph for $\{f_j(\Theta)\}_{j=1}^p$ and find its maximal degree D. Such a graph is not, in general, unique. For example, the complete graph is a valid dependency graph. However, since a smaller D leads to a tighter bound on the cumulants of $f(\Theta)$, we look for the sparsest graph possible. Following a similar procedure as in [25] in which a dependency graph is constructed for the set $\{f_k(\Theta)\}_{k=1}^p$ where $\mathcal{O}_k$ acts non-trivially on only the qubit k, we construct a dependency graph for the general case of $\{f_j(\Theta)\}_{j=1}^p$, albeit with a higher maximal degree.

We begin by defining for each $j \in \{1, ..., p\}$ the set P_j which is analogous to the $\mathcal{P}_k$ introduced in [25, lemma 2.25].

$$\begin{aligned} \mathcal{P}_j = \{j' \in \{1, ..., p\} \quad : \quad & f_j(\Theta) \text{ is not independent of} \\ & f_{j'}(\Theta) \text{ at initialization}\}. \end{aligned} \tag{4.52}$$

Observe that since by assumption 2 the individual parameters in Θ are independent of one another, then $f_j(\Theta)$ and $f_{j'}(\Theta)$ are independent if and only if there is no parameter θ of which they are both a function. With this in mind, it is useful to introduce the auxiliary sets $\mathcal{S}_j$ and $\mathcal{T}_i$ which are analogous to the past and future light cones ($\mathcal{L}_k^p$ and $\mathcal{L}_i^f$) of [25, definition 2.11].

$$\mathcal{S}_j = \{\theta_i \in \Theta : f_j(\Theta) \text{ depends on } \theta_i\}, \tag{4.53}$$

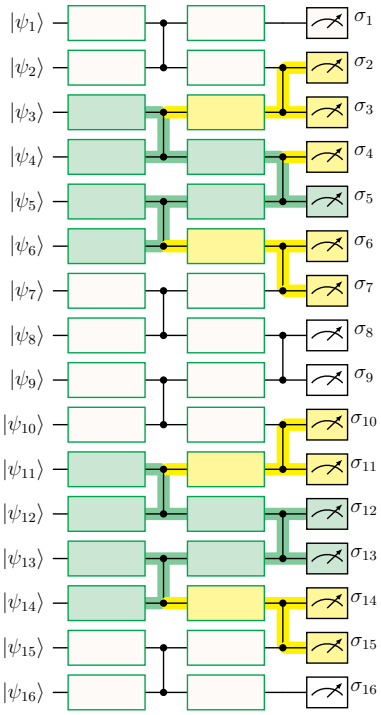

Figure 4.1: An example of a circuit with $m = 16$ and $L = 2$. Let $f_j(\Theta)$ be such that $\mathcal{Q}_j = \{5, 12, 13\}$, and let $f_{j'}(\Theta)$ be such that $\mathcal{Q}_{j'} \cap ([2, 7] \cup [10, 15]) \neq \emptyset$. Then $\mathcal{S}_j \cap \mathcal{S}_{j'} \neq \emptyset$ and so $j' \in \mathcal{P}_j$.

and

$$\mathcal{T}_i = \{j \in \{1, ..., p\} : f_j(\Theta) \text{ depends on } \theta_i\}, \tag{4.54}$$

and notice that

$$P_j \subseteq \{j' \in \{1, ..., p\} \ : \ \mathcal{S}_j \cap \mathcal{S}_{j'} \neq \emptyset\} \tag{4.55}$$

since if $j' \in \mathcal{P}_j$, there is a parameter on which both f_j and $f_{j'}$ depend.

Remark 13. *Clearly,*

$$i \in \mathcal{S}_j \iff j \in \mathcal{T}_i \tag{4.56}$$

for all $i \in \{1, ..., (L + 2)m\}$ *and* $j \in \{1, ..., p\}$.

Lemma 7. *[25, lemma 3.9] Let* $\mathcal{G} = (V, E)$ *be a graph with vertices* $V = \{1, ..., p\}$ *and edges*

$$E = \{(j, j') \in V^2 : j' \in \mathcal{P}_j\}. \tag{4.57}$$

$\mathcal{G}$ *is a dependency graph for the collection* $\{f_j(\Theta)\}_{j=1}^p$.

Proof. Let $V_1, V_2 \subset \{1, ..., p\}$ such that $V_1 \cap V_2 = \emptyset$ and that $\nexists (j, j') \in E$ with $j \in V_1$ and $j' \in V_2$. We show that the collections $\{f_j(\Theta)\}_{j \in V_1}$ and $\{f_{j'}(\Theta)\}_{j' \in V_2}$ are independent.

Let $j \in V_1$ and $j' \in V_2$, then $(j, j') \notin E$, by definition, implies $j' \notin \mathcal{P}_j$. It follows that

$$j' \in \{\{1, ..., p\} : \mathcal{S}_j \cap \mathcal{S}_{j'} = \emptyset\}. \tag{4.58}$$

The set of parameters that both $\{f_j(\Theta)\}_{j\in V_1}$ and $\{f_{j'}(\Theta)\}_{j'\in V_2}$ depend on is

$$\left(\bigcup_{j\in V_1}\mathcal{S}_j\right)\cap\left(\bigcup_{j'\in V_2}\mathcal{S}_{j'}\right)=\bigcup_{\substack{j\in V_1\\ j'\in V_2}}\mathcal{S}_j\cap\mathcal{S}_{j'}=\bigcup_{\substack{j\in V_1\\ j'\in V_2}}\emptyset=\emptyset.$$

Since each collection depends on a distinct set of independent parameters, they are independent. □

Now that we have a valid dependency graph for $\{f_j(\Theta)\}_{j=1}^p$, we establish its maximal degree. Let us first recall the definitions of the sets we introduced in chapter 3. Let

$$\mathcal{O}_j=\sigma_{j_1}\otimes\cdots\otimes\sigma_{j_m},\quad j_k\in\{I,X,Y,Z\} \tag{4.59}$$

be the observable associated with $f_j(\Theta)$. The sets $\mathcal{Q}_j$ and $\mathcal{R}_k$ are defined as

$$\mathcal{Q}_j=\{k\in\{1,...,m\}:\sigma_{j_k}\neq I\}\text{ and }\mathcal{R}_k=\{j\in\{1,...,p\}:\sigma_{j_k}\neq I\}. \tag{4.60}$$

The past light cone of qubit k is the subset of Θ given by

$$\mathcal{N}_k=\bigcup_{l=1}^{L+2}\mathcal{N}_k^l \tag{4.61}$$

where

$$\mathcal{N}_k^l=\bigcup_{k'\in\mathcal{J}_k^l}\{\theta_{k'l}\}. \tag{4.62}$$

And the future light cone of the parameter θ_i is the subset of qubits given by

$$\mathcal{M}_i=\{k'\in\{1,...,m\}:\theta_i\in\mathcal{N}_{k'}\}. \tag{4.63}$$

The following lemma is a generalization of [25, corollary 2.19] in which it is shown that the extended light cones generalize light cones:

$$\mathcal{L}_k^p\subseteq\mathcal{N}_k,\quad \mathcal{L}_i^f\subseteq\mathcal{M}_i. \tag{4.64}$$

Lemma 8. *The relations*

$$\mathcal{S}_j\subseteq\bigcup_{k\in\mathcal{Q}_j}\mathcal{N}_k\quad\textit{and}\quad\mathcal{T}_i\subseteq\bigcup_{k\in\mathcal{M}_i}\mathcal{R}_k \tag{4.65}$$

hold.

Proof. If $\theta_i\in\mathcal{S}_j$, then $f_j(\Theta)$ depends on the parameter θ_i, and by Lemma 3, θ_i must appear in the pruned circuit $[U(\Theta)]_j$. Let k,l be such that $i=(l-1)\times m+k$. By definition of the pruning operation,

$$\theta_i=\theta_{k,l}\in\bigcup_{k'\in\mathcal{Q}_j}\mathcal{N}_{k'}^l\subseteq\bigcup_{k'\in\mathcal{Q}_j}\mathcal{N}_{k'} \tag{4.66}$$

since $\mathcal{N}_{k'}^l\subseteq\mathcal{N}_{k'}$ for all $l\in\{1,...,L+2\}$. We conclude that $\mathcal{S}_j\subseteq\bigcup_{k\in\mathcal{Q}_j}\mathcal{N}_k$.

On the other hand, let $j \in \mathcal{T}_i$, then $\theta_i \in \mathcal{S}_j \subseteq \bigcup_{k\in\mathcal{Q}_j} \mathcal{N}_k$, as just demonstrated. This is equivalent to say that

$$\exists k \in \mathcal{Q}_j \text{ such that } \theta_i \in \mathcal{N}_k$$

and by Remark 10,

$$\exists k \in \mathcal{Q}_j \text{ such that } k \in \mathcal{M}_i.$$

Rewriting we get

$$\exists k \in \mathcal{M}_i \text{ such that } k \in \mathcal{Q}_j$$

and by Remark 3

$$\exists k \in \mathcal{M}_i \text{ such that } j \in \mathcal{R}_k$$

which implies $j \in \bigcup_{k\in\mathcal{M}_i} \mathcal{R}_k$ and therefore $\mathcal{T}_i \subseteq \bigcup_{k\in\mathcal{M}_i} \mathcal{R}_k$. □

Corollaries 3 and 4 are generalizations of [25, lemma 2.25] in which it is shown that if $|\mathcal{Q}|= |\mathcal{R}|= 1$, then $|\mathcal{P}_k|\leq |\mathcal{M}||\mathcal{N}|$.

Corollary 3. *For all $j \in \{1, ..., p\}$ and $i \in \{1, ..., (L + 2)m\}$, it holds that*

$$|\mathcal{S}_j|\leq |\mathcal{Q}||\mathcal{N}| \quad \text{and} \quad |\mathcal{T}_i|\leq |\mathcal{M}||\mathcal{R}|. \tag{4.67}$$

Proof. We have

$$|\mathcal{S}_j|= \left|\bigcup_{k\in\mathcal{Q}_j} \mathcal{N}_k\right| \leq \sum_{k\in\mathcal{Q}_j} |\mathcal{N}_k|\leq \sum_{k\in\mathcal{Q}_j} |\mathcal{N}|= |\mathcal{Q}_j||\mathcal{N}|\leq |\mathcal{Q}||\mathcal{N}|, \tag{4.68}$$

and $|\mathcal{T}_i|\leq |\mathcal{M}||\mathcal{R}|$ follows by the same logic. □

Corollary 4. *For all $j \in \{1, ..., p\}$, $\mathcal{P}_j$ satisfies*

$$|\mathcal{P}_j|\leq |\mathcal{M}||\mathcal{N}||\mathcal{Q}||\mathcal{R}|. \tag{4.69}$$

Proof.

$$\begin{aligned} P_j &\subseteq \{j' \in \{1, ..., p\} \,:\, \mathcal{S}_j \cap \mathcal{S}_{j'} \neq \emptyset\} \\ &= \{j' \in \{1, ..., p\} \,:\, \exists i \in \mathcal{S}_j \cap \mathcal{S}_{j'}\} \\ &= \bigcup_{i\in\mathcal{S}_j} \{j' \in \{1, ..., p\} \,:\, i \in \mathcal{S}_{j'}\} \\ &= \bigcup_{i\in\mathcal{S}_j} \{j' \in \{1, ..., p\} \,:\, j' \in \mathcal{T}_i\} \\ &= \bigcup_{i\in\mathcal{S}_j} \mathcal{T}_i. \end{aligned} \tag{4.70}$$

Therefore

$$\begin{aligned} |\mathcal{P}_j| &\leq |\mathcal{S}_j| \max_i |\mathcal{T}_i| \\ &\leq |\mathcal{Q}||\mathcal{N}||\mathcal{M}||\mathcal{R}|. \end{aligned} \tag{4.71}$$

□

Corollary 5. *If D is the maximum degree of $\mathcal{G}$, then $D \leq |\mathcal{M}||\mathcal{N}||\mathcal{Q}||\mathcal{R}|$.*

Proof. Let D be the maximum degree of $\mathcal{G}$, then

$$\begin{aligned} D &= \max_{j \in V} |\{(j, j') \in E\}| \\ &= \max_{j \in V} |\{(j, j') \in V^2 : j' \in \mathcal{P}_j\}| \\ &= \max_{j \in V} |\mathcal{P}_j| \\ &\leq |\mathcal{M}||\mathcal{N}||\mathcal{Q}||\mathcal{R}| \end{aligned} \tag{4.72}$$

by Corollary 4. □

We now have all the ingredients to apply Theorem 6.

$$\begin{aligned} |\mathcal{K}_r(f(\Theta))| &\leq \frac{N}{D}(4eAD)^r r! \,. \\ &\leq \frac{m|\mathcal{R}|}{|\mathcal{M}||\mathcal{N}||\mathcal{Q}||\mathcal{R}|} \left(\frac{4eW|\mathcal{M}||\mathcal{N}||\mathcal{Q}||\mathcal{R}|}{N} \right)^r r! \\ &= \frac{m}{|\mathcal{M}||\mathcal{N}||\mathcal{Q}|} \left((4eW) \frac{|\mathcal{M}||\mathcal{N}||\mathcal{Q}||\mathcal{R}|}{N} \right)^r r! \,. \end{aligned} \tag{4.73}$$

This implies the bound on the tail function

$$\begin{aligned} |T(t)| &= \left| \sum_{r=3}^{\infty} \frac{\mathcal{K}_r(f(\Theta))}{r!} t^r \right| \\ &\leq \sum_{r=3}^{\infty} \frac{|\mathcal{K}_r(f(\Theta))|}{r!} t^r \\ &\leq \frac{m}{|\mathcal{M}||\mathcal{N}||\mathcal{Q}|} \sum_{r=3}^{\infty} \left((4eWt) \frac{|\mathcal{M}||\mathcal{N}||\mathcal{Q}||\mathcal{R}|}{N} \right)^r \\ &\leq (4eWt)^3 \frac{m|\mathcal{M}|^2|\mathcal{N}|^2|\mathcal{Q}|^2|\mathcal{R}|^3}{N^3} \sum_{r=0}^{\infty} \left((4eWt) \frac{|\mathcal{M}||\mathcal{N}||\mathcal{Q}||\mathcal{R}|}{N} \right)^r . \end{aligned} \tag{4.74}$$

This sum converges for all t such that

$$|4eWt| \frac{|\mathcal{M}||\mathcal{N}||\mathcal{Q}||\mathcal{R}|}{N} < 1. \tag{4.75}$$

If we further enforce that

$$|4eWt| \frac{|\mathcal{M}||\mathcal{N}||\mathcal{Q}||\mathcal{R}|}{N} < \frac{1}{2}, \tag{4.76}$$

i.e. that

$$|t| < \frac{N}{8eW|\mathcal{M}||\mathcal{N}||\mathcal{Q}||\mathcal{R}|}, \tag{4.77}$$

then

$$|T(t)| \leq 2(4Wet)^3 \frac{m|\mathcal{M}|^2|\mathcal{N}|^2|\mathcal{Q}|^2|\mathcal{R}|^3}{N^3} \tag{4.78}$$

implying

$$|\varphi_f(t) - \varphi_Z(t)| \leq e^{-\frac{1}{2}t^2}(e^{at^3} - 1) \tag{4.79}$$

where we have defined

$$a := 2(4We)^3 \frac{m|\mathcal{M}|^2|\mathcal{N}|^2|\mathcal{Q}|^2|\mathcal{R}|^3}{N^3}. \tag{4.80}$$

The integral and optimization

Our next task is to solve the integral

$$\begin{aligned} \int_0^T \frac{|\varphi_f(t) - \varphi_Z(t)|}{t} dt &\leq \int_0^T \frac{e^{-\frac{1}{2}t^2}(e^{at^3} - 1)}{t} dt \\ &\leq \int_0^T \frac{e^{-\frac{1}{2}t^2}(e^{aTt^2} - 1)}{t} dt. \end{aligned} \tag{4.81}$$

Using the fact that $e^{|z|} - 1 \leq |z|e^{|z|}$ for all $z \in \mathbb{R}$, we have

$$\begin{aligned} \int_0^T \frac{e^{-\frac{1}{2}t^2}(e^{aTt^2} - 1)}{t} dt &\leq aT \int_0^T te^{-\frac{1}{2}t^2(1-2aT)} dt \\ &= \frac{aT}{1-2aT}(1 - e^{-\frac{1}{2}T^2(1-2aT)}). \end{aligned} \tag{4.82}$$

Substituting (4.82) into (4.37) gives

$$\begin{aligned} \sup_{x\in\mathbb{R}} |F(x) - \Phi(x)| &\leq \frac{2aT}{\pi(1-2aT)}(1 - e^{-\frac{1}{2}T^2(1-2aT)}) + \frac{24}{\pi\sqrt{2\pi}T} \\ &\leq \frac{2}{\pi}\left(\frac{aT}{(1-2aT)} + \frac{12}{\sqrt{2\pi}T}\right) \end{aligned} \tag{4.83}$$

for all $T \in \left(0, \frac{N}{8eW|\mathcal{M}||\mathcal{N}||\mathcal{Q}||\mathcal{R}|}\right]$ such that $1 - 2aT \geq 0$.

We now find an appropriate $\tilde{T}$ such that (4.83) is minimized. The first order condition tells us that

$$\begin{aligned} 0 &= \frac{a}{(1-2a\tilde{T})^2} - \frac{12}{\sqrt{2\pi}\tilde{T}^2} \\ &= a\sqrt{2\pi}\tilde{T}^2 - 12(1-2a\tilde{T})^2 \\ &= (a\sqrt{2\pi} - 48a^2)\tilde{T}^2 + 48a\tilde{T} - 12, \end{aligned} \tag{4.84}$$

thus

$$\tilde{T} = \frac{2\sqrt{3}}{4\sqrt{3}a \pm \sqrt{a\sqrt{2\pi}}}, \tag{4.85}$$

which implies

$$1 - 2a\tilde{T} = \frac{\pm\sqrt{a\sqrt{2\pi}}}{4\sqrt{3}a + \sqrt{a\sqrt{2\pi}}}. \tag{4.86}$$

We therefore select $\tilde{T} = \frac{2\sqrt{3}}{4\sqrt{3}a+\sqrt{a\sqrt{2\pi}}}$ to ensure that $1 - 2a\tilde{T} \geq 0$. This also ensures that the second order condition, namely that

$$0 \leq \frac{4a^2}{(1-2a\tilde{T})^3} + \frac{24}{\sqrt{2\pi}\tilde{T}^4}, \tag{4.87}$$

is satisfied. Finally, we show that $\tilde{T} \leq \frac{N}{8eW|\mathcal{M}||\mathcal{N}||\mathcal{Q}||\mathcal{R}|}$.

$$\begin{aligned}
\tilde{T} &= \frac{2\sqrt{3}}{4\sqrt{3}a + \sqrt{a\sqrt{2\pi}}} \\
&\leq \frac{2\sqrt{3}}{4\sqrt{3}a} \\
&= \frac{1}{2a} \\
&= \frac{N^3}{4(4eW)^3 m|\mathcal{M}|^2|\mathcal{N}|^2|\mathcal{Q}|^2|\mathcal{R}|^3} \\
&= \frac{N}{8eW|\mathcal{M}||\mathcal{N}||\mathcal{Q}||\mathcal{R}|}\left(\frac{N^2}{2(4eW)^2 m|\mathcal{M}||\mathcal{N}||\mathcal{Q}||\mathcal{R}|^2}\right).
\end{aligned} \tag{4.88}$$

We use the fact that $\mathcal{K}_2(f(\Theta)) = 1$, so that by (4.73)

$$1 = |\mathcal{K}_2(f(\Theta))| \leq \frac{m}{|\mathcal{M}||\mathcal{N}||\mathcal{Q}|}\left((4eW)\frac{|\mathcal{M}||\mathcal{N}||\mathcal{Q}||\mathcal{R}|}{N}\right)^2 2! \tag{4.89}$$

implying

$$\frac{N^2}{2(4eW)^2 m|\mathcal{M}||\mathcal{N}||\mathcal{Q}||\mathcal{R}|^2} \leq 1 \tag{4.90}$$

which proves the claim.

All together we have

$$\begin{aligned}
d_K(P_{f(\Theta)}, \mathcal{N}_{0,1}) &\leq \frac{2}{\pi}\left(\frac{a\tilde{T}}{(1-2a\tilde{T})} + \frac{12}{\sqrt{2\pi}\tilde{T}}\right) \\
&= \frac{2}{\pi}\left(\frac{2\sqrt{3}a}{\sqrt{a\sqrt{2\pi}}} + \frac{12(4\sqrt{3}a + \sqrt{a\sqrt{2\pi}})}{2\sqrt{6\pi}}\right) \\
&= \frac{8}{\pi}\left(\sqrt{\frac{3a}{\sqrt{2\pi}}} + \frac{6a}{\sqrt{2\pi}}\right) \\
&= \frac{64}{\pi}\left(\sqrt{\frac{6(eW)^3}{\sqrt{2\pi}}\frac{m|\mathcal{M}|^2|\mathcal{N}|^2|\mathcal{Q}|^2|\mathcal{R}|^3}{N^3}} + \frac{96(eW)^3}{\sqrt{2\pi}}\frac{m|\mathcal{M}|^2|\mathcal{N}|^2|\mathcal{Q}|^2|\mathcal{R}|^3}{N^3}\right).
\end{aligned} \tag{4.91}$$

4.3 Some estimates and examples

The variable quantity of interest with regard to Theorem 5 is

$$\frac{m|\mathcal{M}|^2|\mathcal{N}|^2|\mathcal{Q}|^2|\mathcal{R}|^3}{N^3}. \tag{4.92}$$

Since we know a priori that $d_K(P_{f(\Theta)}, \mathcal{N}_{0,1}) < 1$, the conclusion of Theorem 5 is trivial unless $f(\Theta)$ is such that (4.92) is very small. In fact, we are interested in the case when it vanishes in the large m limit. In this section we investigate under what conditions this may occur. In order to do so it is necessary to gain an understanding of the behavior of $|\mathcal{M}|$, $|\mathcal{N}|$, $|\mathcal{Q}|$, $|\mathcal{R}|$ and N in relation to m. Clearly, the growth of $|\mathcal{M}|$ and $|\mathcal{N}|$ is controlled by L. Similarly, we can control the growth of $|\mathcal{R}|$ with $|\mathcal{Q}|$. Finally, we will see that N is a function of m, L and $|\mathcal{Q}|$. The dependencies, or lack thereof, of L and $|\mathcal{Q}|$ on m will therefore determine how (4.92) behaves as $m \to \infty$.

Light cones

In [25, sec 2.3.4], the following architecture independent bounds are given:

$$|\mathcal{M}| \leq 2^L \quad \text{and} \quad |\mathcal{N}| \leq 2^{L+1}. \tag{4.93}$$

And if $U(\Theta)$ is d-dimensional and geometrically local, then

$$|\mathcal{M}| \in O(L^d) \quad \text{and} \quad |\mathcal{N}| \in O(L^{d+1}). \tag{4.94}$$

Observable weights

The theoretical upper bound of p is $|\{\sigma_I, \sigma_X, \sigma_Y, \sigma_Z\}^{\otimes m}| = 4^m$ in which case $|\mathcal{Q}| = m$ and

$$|\mathcal{R}| = \sum_{i=1}^{m} \binom{m-1}{i-1} 3^i = 3 \cdot 4^{m-1}. \tag{4.95}$$

However, for an efficient VQA, we need $p \in O(m^c)$ for some constant c. For fixed $|\mathcal{Q}|$ an upper bound on p and $|\mathcal{R}|$ is

$$p \leq \sum_{i=1}^{|\mathcal{Q}|} \binom{m}{i} 3^i = \sum_{i=1}^{|\mathcal{Q}|} \frac{m!}{(m-i)!\, i!} 3^i \leq m^{|\mathcal{Q}|} \sum_{i=1}^{|\mathcal{Q}|} \frac{3^i}{i!} \leq e^3 m^{|\mathcal{Q}|} \tag{4.96}$$

and

$$|\mathcal{R}| \leq \sum_{i=1}^{|\mathcal{Q}|} \binom{m-1}{i-1} 3^i \leq e^3 m^{|\mathcal{Q}|-1}. \tag{4.97}$$

If $|\mathcal{Q}| \in \Theta(1)$, then the efficiency requirement is satisfied.

We can find improved bounds if $\mathcal{O}$ is spatially h-local. In this case, obviously $|\mathcal{Q}| = h$. An upper bound on $|\mathcal{R}|$ will depend on how the qubits are arranged. For example, if the qubits are arranged in a 1-dimensional lattice - a line in other words - then we can estimate

$$|\mathcal{R}| \leq \sum_{i=1}^{h} i \cdot 3^i \leq \frac{3}{4}(2 \cdot 3^h \cdot h - 3^h + 1) \leq \frac{3}{2} 3^h \cdot h \tag{4.98}$$

and

$$p \leq \frac{3}{2} 3^h \cdot hm. \tag{4.99}$$

The normalization constant

We start by deriving an upper bound on N.

$$1 = \mathbb{V}[f(\Theta)] = \sum_{j=1}^{p} \mathbb{V}[f_j(\Theta)] + \sum_{j=1}^{p} \sum_{j' \neq j} \mathrm{Cov}[f_j(\Theta) f_{j'}(\Theta)] = \sum_{j=1}^{p} \mathbb{E}[f_j^2(\Theta)] \tag{4.100}$$

by Lemma 2 and

$$\sum_{j=1}^{p} \mathbb{E}[f_j^2(\Theta)] \leq \sum_{j=1}^{p} \left(\frac{W}{N}\right)^2 = \frac{pW^2}{N^2} \tag{4.101}$$

by Lemma 1, thus $N \leq \sqrt{p} W \leq \sqrt{m|\mathcal{R}|} W$.

On the other hand, having no knowledge of $U(\Theta)$, it is much harder to determine a lower bound for N besides the trivial one: $N \geq 0$. In [40] a class of circuits is introduced in which the entangling gates V are randomly sampled from $\mathcal{U}(4)$ according to a *unitary 2-design*. A unitary 2-design on $\mathcal{U}(d)$ is an approximation of the *Haar measure* which is the unique translation invariant measure on $\mathcal{U}(d)$. The Haar measure, denoted μ_H, is often referred to as the uniform measure on the unitary group because it assigns an equal probability to each element of the group.

Definition 25. Let ν be a probability distribution over a set of unitary matrices $S \subset \mathcal{U}(d)$. ν is a *unitary t-design* if

$$\mathbb{E}_{V \sim \nu}[V^{\otimes k} O V^{\dagger \otimes k}] = \mathbb{E}_{U \sim \mu_H}[U^{\otimes k} O U^{\dagger \otimes k}] \tag{4.102}$$

for all $O \in \mathcal{L}((\mathbb{C}^d)^{\otimes k})$.

Generating Haar random unitaries is computationally expensive, whereas generating random unitaries according to a k-design can be done efficiently. This makes them a desirable alternative for applications in which only the lower order moments of the Haar measure need be replicated [47].

The author of [40] has shown that if the entangling gates V are sampled according to a unitary 2-design ν, then there exists a constant c such that

$$\mathbb{E}_{V \sim \nu}[N^2] \geq \frac{p}{2^{cL|\mathcal{Q}|}}. \tag{4.103}$$

This suggests that the variance of $f(\Theta)$ decays exponentially in $L|\mathcal{Q}|$ and so for the purpose of this discussion we assume that there exists a constant c such that

$$N \geq \frac{\sqrt{p}}{2^{cL|\mathcal{Q}|}} \geq \frac{\sqrt{m}}{2^{cL|\mathcal{Q}|}} \tag{4.104}$$

for all configurations of the entangling gates of $U(\Theta)$. If $U(\Theta)$ is 1 dimensional geometrically local and $\mathcal{O}$ is spatially h-local, then [40] gives the improved bound of

$$N \geq \frac{\sqrt{m}}{2^{c(L+h)}}. \tag{4.105}$$

4.3.1 Logarithmic depth circuits and fixed weight observables

We now analyze the behavior of $f(\Theta)$ in the large m limit based on the bounds we have derived in this section. In particular, we examine the case that $L \in \Theta(\log_2 m)$ and $|\mathcal{Q}| \in \Theta(1)$. We also consider the possibility of a quantum advantage. Recall that in section 3.3.1 we showed that the computation of $f(\Theta)$ is classically simulable with $O(pL2^{|\mathcal{Q}|2^L})$ operations in general, $O(pL2^{|\mathcal{Q}|L^d})$ operations if $U(\Theta)$ is d-dimensional geometrically local, and $O(pL2^{h+2L})$ operations if $U(\Theta)$ is 1-dimensional geometrically local and $\mathcal{O}$ is spatially h-local.

1. In the most general case

$$\frac{m|\mathcal{M}|^2|\mathcal{N}|^2|\mathcal{Q}|^2|\mathcal{R}|^3}{N^3} \in O\left(\frac{2^{L(4+3c|\mathcal{Q}|)}|\mathcal{Q}|^2}{m^{1/2-3(|\mathcal{Q}|-1)}}\right). \tag{4.106}$$

   For this to converge it must be true that $3(|\mathcal{Q}|-1) < \frac{1}{2}$ i.e. that $|\mathcal{Q}| = 1$ which is the case when $\mathcal{O}$ is the sum of single qubit observables.

   (a) If we choose $|\mathcal{Q}| = 1$ and $L \in \Theta(1)$, then (4.106) behaves as $O(m^{-1/2})$ which clearly tends to 0. However, with this choice of $|\mathcal{Q}|$ and L we do not achieve a quantum advantage since $f(\Theta)$ is efficiently simulable with linear run time.

   (b) If we choose $|\mathcal{Q}| = 1$ and $L = \epsilon \log_2(m)$ for some $\epsilon > 0$, then (4.106) behaves as $O(m^{-1/2+\epsilon(4+3c)})$ which tends to 0 as long as $\epsilon < \frac{1}{2(4+3c)}$. Furthermore, with this choice of $|\mathcal{Q}|$ and L, we may achieve a quantum advantage since to compute $f(\Theta)$ classically, $O(m \log_2(m^\epsilon) 2^{m^\epsilon})$ operations are needed.

2. If $U(\Theta)$ is d-dimensional geometrically local, then

$$\frac{m|\mathcal{M}|^2|\mathcal{N}|^2|\mathcal{Q}|^2|\mathcal{R}|^3}{N^3} \in O\left(\frac{2^{3cL|\mathcal{Q}|}L^{4d+2}|\mathcal{Q}|^2}{m^{1/2-3(|\mathcal{Q}|-1)}}\right). \tag{4.107}$$

   Again, letting $|\mathcal{Q}| = 1$ and $L = \epsilon \log_2 m$, (4.107) behaves as

$$O\left(\frac{\log_2(m^\epsilon)^{4d+2}}{m^{1/2-\epsilon 3c}}\right) \tag{4.108}$$

   which converges as long as $\epsilon < \frac{1}{6c}$. Classical simulation is possible with

$$O(\log_2(m^\epsilon) m^{1+\epsilon^d \log_2^{d-1} m}) \tag{4.109}$$

   operations which is super-polynomial if $d \geq 2$.

3. If $\mathcal{O}$ is spatially h-local and $U(\Theta)$ is 1 dimensional but not geometrically local, then

$$\frac{m|\mathcal{M}|^2|\mathcal{N}|^2|\mathcal{Q}|^2|\mathcal{R}|^3}{N^3} \in O\left(\frac{2^{L(4+3ch)}3^{3h}h^5}{\sqrt{m}}\right). \tag{4.110}$$

   In this configuration $|\mathcal{Q}| = h$ need not be 1. If $h \in \Theta(1)$ and $L = \epsilon \log_2(m)$ for some $\epsilon > 0$, then (4.110) behaves as $O(m^{-1/2+\epsilon(4+3ch)})$ which converges as long as $\epsilon < \frac{1}{2(4+3ch)}$. Once again, we expect a quantum advantage since $O(m \log_2(m^\epsilon) 2^{hm^\epsilon})$ operations are needed for classical simulation.

4. If $\mathcal{O}$ is spatially h-local and $U(\Theta)$ is 1 dimensional and geometrically local, then

$$\frac{m|\mathcal{M}|^2|\mathcal{N}|^2|\mathcal{Q}|^2|\mathcal{R}|^3}{N^3} \in O\left(\frac{2^{3c(L+h)}3^{3h}h^5L^6}{\sqrt{m}}\right). \tag{4.111}$$

Due to the improved lower bound on N given by (4.105) we may have that both L and h are logarithmic. Let $h = \epsilon_1 \log_3 m$ and $L = \epsilon_2 \log_2 m$, then 4.111 behaves as

$$O\left(\frac{\log_3^5(m^{\epsilon_1})\log_2^6(m^{\epsilon_2})}{m^{1/2-3(c(\epsilon_1+\epsilon_2)+\epsilon_1)}}\right) \tag{4.112}$$

which converges as long as $\epsilon_1(1+c)+\epsilon_2 \leq \frac{1}{6}$. However, in this configuration we do not obtain a quantum advantage since classical simulation is possible with

$$O(m^{2(\epsilon_1+\epsilon_2)+1}\log_3(m^{\epsilon_1})\log_2(m^{\epsilon_2})) \tag{4.113}$$

operations.

Based on these examples it would appear that to achieve a desirable result, $\mathcal{O}$ must be spatially local with constant weight, L must be at most logarithmic in m, and $U(\Theta)$ must either be not geometrically local, or have dimension greater than 1. We stress, however, that this is not necessarily the case. The bounds presented in this section are very crude as they must account for the very worst case. It is for this reason that we have not simplified the conclusion of Theorem 5 to rely on just m, L and $|\mathcal{Q}|$.

4.4 Comparison with previous work

An alternative metric to the Kolmogorov distance is the *Wasserstein distance*.

Definition 26. Let μ and ν be probability measures on $\mathbb{R}$ and let $p \in [1, \infty)$. The *Wasserstein distance of order p* (or *p-Wasserstein distance*) between μ and ν is defined by

$$d_{W_p}(\mu,\nu) = \left(\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathbb{R}}|x-y|^p d\pi(x,y)\right)^{1/p} \tag{4.114}$$

where $\Pi(\mu,\nu)$ is the set of probability measures on $\mathbb{R}^2$ which admit μ and ν as marginals.

Intuitively, the 1-Wasserstein distance represents the minimum "cost" associated with transforming one probability measure into the other. If we visualize two probability measures as two piles of dirt distributed over an area, then the cost takes on the interpretation of the amount of work required to transform one pile into the other. It depends on the amount of dirt which must be moved and how far. For this reason it is often called the earth-mover's distance. For a comprehensive treatment of this topic see [48]. The following proposition illustrates the relationship between the Kolmogorov and 1-Wasserstein distances.

Proposition 1 ([49]). *Let μ and ν be probability measures on $\mathbb{R}$. If the density of ν with respect to the Lebesgue measure is bounded by C, then*

$$d_K(\mu,\nu) \leq \sqrt{2C \cdot d_{W_1}(\mu,\nu)}. \tag{4.115}$$

In [26], under the same framework as ours, the authors use Stein's method to estimate the 1-Wasserstein distance between a centered Gaussian process and the distribution of the outputs of an untrained quantum neural network over a set of inputs. We can use this result to infer an estimate in the univariate case. [26, Theorem 5.1] provides the following bound when $|\mathcal{Q}|= |\mathcal{R}|= W = 1$.

$$d_{W_1}(P_{f(\Theta)}, \mathcal{N}_{0,1}) \leq 8\frac{m|\mathcal{M}|^{7/2}|\mathcal{N}|^{7/2}}{N^3}(1+\log N). \tag{4.116}$$

The Lebesgue density of $\mathcal{N}_{0,1}$ is bounded by $C = 1/\sqrt{2\pi}$, therefore (4.116) combined with Proposition 1 gives $d_K(P_{f(\Theta)}, \mathcal{N}_{0,1}) < \mathcal{B}_1$, where

$$\mathcal{B}_1 = \sqrt{\frac{16}{\sqrt{2\pi}}\frac{m|\mathcal{M}|^{7/2}|\mathcal{N}|^{7/2}}{N^3}(1+\log N)} \tag{4.117}$$

whereas Theorem 5 tells us that $d_K(P_{f(\Theta)}, \mathcal{N}_{0,1}) < \mathcal{B}_2$ where

$$\mathcal{B}_2 = \frac{64}{\pi}\left(\sqrt{\frac{6e^3}{\sqrt{2\pi}}\frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3}} + \frac{96e^3}{\sqrt{2\pi}}\frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3}\right). \tag{4.118}$$

So which bound is better? On one hand, $\mathcal{B}_2$ undoubtedly suffers from huge constant terms. But on the other hand,

$$\frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3} \leq \frac{m|\mathcal{M}|^{7/2}|\mathcal{N}|^{7/2}}{N^3}(1+\log N) \tag{4.119}$$

for all m, implying that $\mathcal{B}_2$ is more asymptotically favorable. To illustrate this, let us assume that

$$\lim_{m\to\infty} \frac{m|\mathcal{M}|^{7/2}|\mathcal{N}|^{7/2}}{N^3}(1+\log_2 N) = 0. \tag{4.120}$$

Then

$$\lim_{m\to\infty} \frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3} = 0 \tag{4.121}$$

and

$$\lim_{m\to\infty} |\mathcal{M}|^{3/2}|\mathcal{N}|^{3/2}(1+\log N) = \infty. \tag{4.122}$$

Then there exists $M_1 > 0$ such that for all $m \geq M_1$,

$$\frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3} \leq \frac{\sqrt{2\pi}}{96e^3} \tag{4.123}$$

and there exists $M_2 > 0$ such that for all $m \geq M_2$,

$$|\mathcal{M}|^{3/2}|\mathcal{N}|^{3/2}(1+\log N) \geq \frac{26112e^3}{\pi^2}. \tag{4.124}$$

It follows that for all $m \geq \max\{M_1, M_2\}$ we have

$$\begin{aligned}
\mathcal{B}_2 &= \frac{64}{\pi}\left(\sqrt{\frac{6e^3}{\sqrt{2\pi}}\frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3}} + \frac{96e^3}{\sqrt{2\pi}}\frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3}\right) \\
&\leq \frac{64}{\pi}\sqrt{\frac{102e^3}{\sqrt{2\pi}}\frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3}} \\
&= \frac{\sqrt{26112e^3}}{\pi}\sqrt{\frac{16}{\sqrt{2\pi}}\frac{m|\mathcal{M}|^2|\mathcal{N}|^2}{N^3}} \\
&= \frac{\sqrt{26112e^3}}{\pi\sqrt{|\mathcal{M}|^{3/2}|\mathcal{N}|^{3/2}(1+\log N)}}\sqrt{\frac{16}{\sqrt{2\pi}}\frac{m|\mathcal{M}|^{7/3}|\mathcal{N}|^{7/3}}{N^3}(1+\log N)} \\
&= \frac{\sqrt{26112e^3}}{\pi\sqrt{|\mathcal{M}|^{3/2}|\mathcal{N}|^{3/2}(1+\log N)}}\mathcal{B}_1 \\
&\leq \mathcal{B}_1. \qquad (4.125)
\end{aligned}$$

Likewise, one could just as easily show that there exists $M_3 > 0$ such that for all $m \leq M_3$, the opposite relation is true: $\mathcal{B}_2 \geq \mathcal{B}_1$. Theorem 5 therefore provides a tighter bound than [26, Theorem 5.1] when m is large, and the opposite is true when m is small. Precisely how large or small depends on how $|\mathcal{M}|, |\mathcal{N}|$ and N grow in relation to m. In any case, one may use the two results in conjunction:

$$d_K(f(\Theta), Z) \leq \min\{\mathcal{B}_1, \mathcal{B}_2\}. \qquad (4.126)$$

Chapter 5

Conclusion

We have proved an upper bound on the Kolmogorov distance between the output of a randomly initialized variational quantum circuit and the Gaussian random variable with the same mean and variance. Our bound, which is easily computable with the knowledge of a given circuit architecture, provides insight into the number of qubits required in order to achieve any desired precision in the approximation of the output of the circuit with a Gaussian random variable. Due to the large constant terms, the number of qubits required to make the bound non-trivial is unfortunately beyond the reach of any currently available hardware.

For example, we saw in section 4.3.1 that the best asymptotic bound we can hope for, which happens when $|\mathcal{Q}|= 1$ and L is constant, is when

$$\frac{m|\mathcal{M}|^2|\mathcal{N}|^2|\mathcal{Q}|^2|\mathcal{R}|^3}{N^3} \in O(m^{-1/2}). \tag{5.1}$$

Supposing that $W = 1$ and plugging this into Theorem 5, we get

$$\begin{aligned} d_K(P_{f(\Theta)}, \mathcal{N}_{0,1}) &\leq \frac{64}{\pi}\left(\sqrt{\frac{6e^3C}{\sqrt{2\pi m}}} + \frac{96e^3C}{\sqrt{2\pi m}}\right) \\ &\approx 141\sqrt{\frac{C}{\sqrt{m}}} + 15671\frac{C}{\sqrt{m}} \end{aligned} \tag{5.2}$$

for some constant C independent of m. For this bound to be less than 1 we require $m > 2.11 \times 10^9$ qubits.

A natural follow-up to this work would be to determine whether a lower and more realistic number of qubits are still enough for the Gaussian approximation to hold. Moreover, it would be interesting to investigate whether the relaxed constraints on the observable improve either the expressibility or trainability of the model function defined by the quantum neural network. Finally, it would be fundamental to extend our Berry-Esseen type bounds to trained quantum neural networks as has been done in [26] with the Wasserstein distance of order 1.

Appendix A

Consequence of the final layers

We provide a proof of Lemma 2 on the implications of the final two layers of the circuit. For convenience Lemma 2 is restated along with Assumption 4 in which the final two layers are characterized.

Assumption 4. *V_{L+1} and V_{L+2} are the identity on $(\mathbb{C}^2)^{\otimes m}$ and*

$$W_{k,L+1}(\theta_{k,L+1}) = R_X(\theta_{k,L+1}) = e^{-i\frac{\theta_{k,L+1}}{2}\sigma_X} \tag{3.27}$$

$$W_{k,L+2}(\theta_{k,L+2}) = R_Z(\theta_{k,L+2}) = e^{-i\frac{\theta_{k,L+2}}{2}\sigma_Z} \tag{3.28}$$

for all $k \in \{1, ..., m\}$. Furthermore, the parameters used in the final two layers are independent and sampled uniformly from $[0, 2\pi]$ at initialization.

Lemma 2. *Assumption 4 ensures that at initialization*

1. $\mathbb{E}[f_j(\Theta)] = 0$ *for all* $j \in \{1, ..., p\}$ *such that* $\mathcal{Q}_j \neq \emptyset$ *and*

2. $\mathrm{Cov}[f_j(\Theta), f_{j'}(\Theta)] = 0$ *for all* $j, j' \in \{1, ..., p\}$ *such that* $j \neq j'$.

Proof. Consider the value of $f_j(\Theta)$ with the layers of the circuit expanded:

$$\begin{aligned} f_j(\Theta) &= \frac{w_j}{N}\langle\psi|U^\dagger \mathcal{O}_j U|\psi\rangle \\ &= \frac{w_j}{N}\langle\psi|U_1^\dagger \cdots U_{L+1}^\dagger U_{L+2}^\dagger \mathcal{O}_j U_{L+2} U_{L+1} \cdots U_1|\psi\rangle \\ &= \frac{w_j}{N}\langle\psi|U_1^\dagger \cdots U_L^\dagger \left(\bigotimes_{k=1}^{m} e^{i\frac{\theta_{k,L+1}}{2}\sigma_X} e^{i\frac{\theta_{k,L+2}}{2}\sigma_Z} \sigma_{j_k} e^{-i\frac{\theta_{k,L+2}}{2}} e^{-i\frac{\theta_{k,L+1}}{2}\sigma_X} \right) U_L \cdots U_1|\psi\rangle. \end{aligned}$$

Using the identity

$$e^{-i\theta\sigma_\alpha} = \cos(\theta)I - i\sin(\theta)\sigma_\alpha, \quad \alpha \in \{I, X, Y, Z\} \tag{A.1}$$

we see that

$$\begin{aligned} & e^{i\frac{\theta}{2}\sigma_X}\sigma_\beta e^{-i\frac{\theta}{2}\sigma_X} \\ = \; & (\cos(\theta/2)I + i\sin(\theta/2)\sigma_X)\sigma_\beta(\cos(\theta/2)I - i\sin(\theta/2)\sigma_X) \\ = \; & \cos^2(\theta/2)\sigma_\beta + \sin^2(\theta/2)\sigma_X\sigma_\beta\sigma_X + i\sin(\theta)\cos(\theta/2)(\sigma_X\sigma_\beta - \sigma_\beta\sigma_X) \\ = \; & \begin{cases} \sigma_I & \beta = I \\ \sigma_X & \beta = X \\ \cos(\theta)\sigma_Y - \sin(\theta)\sigma_Z & \beta = Y \\ \cos(\theta)\sigma_Z + \sin(\theta)\sigma_Y & \beta = Z \end{cases}. \end{aligned} \tag{A.2}$$

By the same logic

$$e^{i\frac{\theta}{2}\sigma_Z}\sigma_\beta e^{-i\frac{\theta}{2}\sigma_Z} = \begin{cases} \sigma_I & \beta = I \\ \cos(\theta)\sigma_X - \sin(\theta)\sigma_Y & \beta = X \\ \cos(\theta)\sigma_Y + \sin(\theta)\sigma_X & \beta = Y \\ \sigma_Z & \beta = Z \end{cases}, \tag{A.3}$$

and thus

$$\begin{aligned} & e^{i\frac{\theta_1}{2}\sigma_X}e^{i\frac{\theta_2}{2}\sigma_Z}\sigma_\beta e^{-i\frac{\theta_2}{2}\sigma_Z}e^{-i\frac{\theta_1}{2}\sigma_X} \\ = \; & \begin{cases} \sigma_I & \beta = I \\ \cos(\theta_2)\sigma_X - \cos(\theta_1)\sin(\theta_2)\sigma_Y + \sin(\theta_1)\sin(\theta_2)\sigma_Z & \beta = X \\ \sin(\theta_2)\sigma_X + \cos(\theta_1)\cos(\theta_2)\sigma_Y - \sin(\theta_1)\cos(\theta_2)\sigma_Z & \beta = Y \\ \sin(\theta_1)\sigma_Y + \cos(\theta_1)\sigma_Z & \beta = Z \end{cases}. \end{aligned} \tag{A.4}$$

Define the following functions:

$$h_I(\theta_1, \theta_2; \beta) = \begin{cases} 1 & \beta = I \\ 0 & \text{otherwise} \end{cases}, \tag{A.5}$$

$$h_X(\theta_1, \theta_2; \beta) = \begin{cases} 0 & \beta = I \\ \cos(\theta_2) & \beta = X \\ \sin(\theta_2) & \beta = Y \\ 0 & \beta = Z \end{cases}, \tag{A.6}$$

$$h_Y(\theta_1, \theta_2; \beta) = \begin{cases} 0 & \beta = I \\ -\cos(\theta_1)\sin(\theta_2) & \beta = X \\ \cos(\theta_1)\cos(\theta_2) & \beta = Y \\ \sin(\theta_1) & \beta = Z \end{cases}, \tag{A.7}$$

$$h_Z(\theta_1, \theta_2; \beta) = \begin{cases} 0 & \beta = I \\ \sin(\theta_1)\sin(\theta_2) & \beta = X \\ -\sin(\theta_1)\cos(\theta_2) & \beta = Y \\ \cos(\theta_1) & \beta = Z \end{cases}, \tag{A.8}$$

and notice that if θ_1 and θ_2 are independently sampled from the uniform distribution on $[0, 2\pi]$, then

- $\mathbb{E}[h_\alpha(\theta_1, \theta_2; \beta)] = 0$ for all $\alpha \in \{I, X, Y, Z\}$ and $\beta \in \{X, Y, Z\}$ and
- $\mathbb{E}[h_{\alpha_1}(\theta_1, \theta_2; \beta_1) h_{\alpha_2}(\theta_1, \theta_2; \beta_2)] = 0$ for all $\alpha_1, \alpha_2 \in \{I, X, Y, Z\}$ and $\beta_1, \beta_2 \in \{X, Y, Z\}$ such that $\alpha_1 = \alpha_2 \cap \beta_1 = \beta_2$ is not true.

We can now compactly write

$$e^{i\frac{\theta_{k,L+1}}{2}\sigma_X} e^{i\frac{\theta_{k,L+2}}{2}\sigma_Z} \sigma_{j_k} e^{-i\frac{\theta_{k,L+2}}{2}\sigma_Z} e^{-i\frac{\theta_{k,L+1}}{2}\sigma_X} = \sum_{\alpha \in \{I,X,Y,Z\}} h_\alpha(\theta_{k,L+1}, \theta_{k,L+2}; \sigma_{j_k})\sigma_\alpha \tag{A.9}$$

and letting it be understood that the summation runs over $\alpha \in \{I, X, Y, Z\}$:

$$\begin{aligned}
&\bigotimes_{k=1}^{m} e^{i\frac{\theta_{k,L+1}}{2}\sigma_X} e^{i\frac{\theta_{k,L+2}}{2}\sigma_Z} \sigma_{j_k} e^{-i\frac{\theta_{k,L+2}}{2}} e^{-i\frac{\theta_{k,L+1}}{2}\sigma_X} \\
= &\bigotimes_{k=1}^{m} \sum_{\alpha} h_\alpha(\theta_{k,L+1}, \theta_{k,L+2}; \sigma_{j_k})\sigma_\alpha \\
= &\sum_{\alpha_1,\ldots,\alpha_m} h_{\alpha_1}(\theta_{1,L+1}, \theta_{1,L+2}; \sigma_{j_1})\sigma_{\alpha_1} \otimes \cdots \otimes h_{\alpha_m}(\theta_{m,L+1}, \theta_{m,L+2}; \sigma_{j_m})\sigma_{\alpha_m} \\
= &\sum_{\alpha_1,\ldots,\alpha_m} \left(\prod_{k=1}^{m} h_{\alpha_k}(\theta_{k,L+1}, \theta_{k,L+2}; \sigma_{j_k}) \right) \sigma_{\alpha_1} \otimes \cdots \otimes \sigma_{\alpha_m}.
\end{aligned} \tag{A.10}$$

To ease notation, we use the abbreviation $h^j_{\alpha_k} = h_{\alpha_k}(\theta_{k,L+1}, \theta_{k,L+2}; \sigma_{j_k})$, and define

$$f_{\alpha_1,\ldots,\alpha_m} := \langle\psi| U_1^\dagger \cdots U_L^\dagger (\sigma_{\alpha_1} \otimes \cdots \otimes \sigma_{\alpha_m}) U_L \cdots U_1 |\psi\rangle. \tag{A.11}$$

We can then write

$$f_j(\Theta) = \frac{w_j}{N} \sum_{\alpha_1,\ldots,\alpha_m} \left(\prod_{k=1}^{m} h^j_{\alpha_k} \right) f_{\alpha_1,\ldots,\alpha_m}. \tag{A.12}$$

First we show that if $\mathcal{Q}_j \neq \emptyset$, then $\mathbb{E}[f_j(\Theta)] = 0$. Note that the terms $f_{\alpha_1,\ldots,\alpha_m}$ are independent of the $h^j_{\alpha_k}$ since they only depend on the parameters in the first L layers, thus

$$\mathbb{E}[f_j(\Theta)] = \frac{w_j}{N} \sum_{\alpha_1,\ldots,\alpha_m} \mathbb{E}\left[\prod_{k=1}^{m} h^j_{\alpha_k} \right] \mathbb{E}[f_{\alpha_1,\ldots,\alpha_m}].$$

Furthermore, $h^j_{\alpha_k}$ is independent of $h^j_{\alpha_{k'}}$ for all $k \neq k'$ so

$$\mathbb{E}\left[\prod_{k=1}^{m} h^j_{\alpha_k} \right] = \prod_{k=1}^{m} \mathbb{E}[h^j_{\alpha_k}]. \tag{A.13}$$

Partitioning $\{1,...,m\}$ into $\mathcal{Q}_j$ and $\mathcal{Q}_j^c := \{1,...,m\}\backslash\mathcal{Q}_j$ gives

$$\begin{aligned}\prod_{k=1}^{m}\mathbb{E}[h_{\alpha_k}^j] &= \prod_{k\in\mathcal{Q}_j}\mathbb{E}[h_{\alpha_k}(\theta_{k,L+1},\theta_{k,L+2};\sigma_{j_k})]\prod_{k\in\mathcal{Q}_j^c}\mathbb{E}[h_{\alpha_k}(\theta_{k,L+1},\theta_{k,L+2};I)]\\ &= \prod_{k\in\mathcal{Q}_j}0\prod_{k\in\mathcal{Q}_j^c}\mathbb{E}[h_{\alpha_k}(\theta_{k,L+1},\theta_{k,L+2};I)]\\ &= 0 \end{aligned} \tag{A.14}$$

since it was assumed that $\mathcal{Q}_j \neq 0$. This shows that $\mathbb{E}\left[\prod_{k=1}^m h_{\alpha_k}^j\right] = 0$ for all $\alpha_1,...,\alpha_m \in \{I,X,Y,Z\}$ therefore $\mathbb{E}[f_j(\Theta)] = 0$.

Next we show that, if $j \neq j'$, then $\mathbb{E}[f_j(\Theta)f_{j'}(\Theta)] = 0$. If there were $j,j' \in \{1,...,p\}$ such that $\mathcal{O}_j = \mathcal{O}_{j'}$, then we could combine the terms f_j and $f_{j'}$ as $f_j + f_{j'} = (w_j + w_{j'})\langle\psi|U^\dagger(\Theta)\mathcal{O}_jU(\Theta)|\psi\rangle$, therefore, without loss of generality, we may assume that $\mathcal{O}_j \neq \mathcal{O}_{j'}$ for all $j \neq j'$. In other words, for all $j \neq j'$, there exists $k \in \{1,...,m\}$ such that $\sigma_{j_k} \neq \sigma_{j'_k}$.

Following the notation introduced thus far,

$$f_j(\Theta)f_{j'}(\Theta) = \frac{w_jw_{j'}}{N^2}\sum_{\substack{\alpha_1,...,\alpha_m\\ \alpha'_1,...,\alpha'_m}}\left(\prod_{k=1}^m h_{\alpha_k}^j h_{\alpha'_k}^{j'}\right)f_{\alpha_1,...,\alpha_m}f_{\alpha'_1,...,\alpha'_m}. \tag{A.15}$$

Taking the expected value,

$$\begin{aligned}\mathbb{E}[f_j(\Theta)f_{j'}(\Theta)] &= \frac{w_jw_{j'}}{N^2}\sum_{\substack{\alpha_1,...,\alpha_m\\ \alpha'_1,...,\alpha'_m}}\prod_{k=1}^m\mathbb{E}[h_{\alpha_k}^j h_{\alpha'_k}^{j'}]\mathbb{E}[f_{\alpha_1,...,\alpha_m}f_{\alpha'_1,...,\alpha'_m}]\\ &= \frac{w_jw_{j'}}{N^2}\sum_{\alpha_1,...,\alpha_m}\prod_{k=1}^m\mathbb{E}[h_{\alpha_k}^j h_{\alpha_k}^{j'}]\mathbb{E}[f_{\alpha_1,...,\alpha_m}^2] \end{aligned} \tag{A.16}$$

since $\mathbb{E}[h_{\alpha_k}^j h_{\alpha'_k}^{j'}] = 0$ unless $\alpha_k = \alpha'_k$. Furthermore, since $j \neq j'$, there exists $k \in \{1,...,m\}$ such that $\sigma_{j_k} \neq \sigma_{j'_k}$ in which case $\mathbb{E}[h_{\alpha_k}^j h_{\alpha_k}^{j'}] = 0$ for all $\alpha_k \in \{I,X,Y,Z\}$. Each of the terms in (A.16) are then necessarily zero and thus $\mathbb{E}[f_j(\Theta)f_{j'}(\Theta)] = 0$. □

Bibliography

[1] Giuseppe Bisicchia et al. *From Quantum Mechanics to Quantum Software Engineering: A Historical Review*. 2024. arXiv: 2404.19428 [quant-ph].

[2] Paramita Basak Upama et al. "Evolution of Quantum Computing: A Systematic Survey on the Use of Quantum Computing Tools". In: *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*. 2022, pp. 520–529. DOI: 10.1109/COMPSAC54236.2022.00096.

[3] Rafael B. Audibert et al. "On the Evolution of A.I. and Machine Learning: Towards a Meta-level Measuring and Understanding Impact, Influence, and Leadership at Premier A.I. Conferences". In: (2024). arXiv: 2205.13131 [cs.AI].

[4] Richard P. Feynman. "Simulating Physics with Computers". In: *International Journal of Theoretical Physics* 21 (1982).

[5] David Deutsch. "Quantum Theory as a Universal Physical Theory". In: *International Journal of Theoretical Physics* 24 (1985).

[6] Peter W. Shor. "Algorithms for Quantum Computation: Discrete Logarithms and Factoring". In: *Proceedings of the 35th Annual Symposium on Foundations of Computer Science (FOCS 1994)*. IEEE. 1994, pp. 124–134. DOI: 10.1109/SFCS.1994.365700.

[7] Lov K. Grover. *A fast quantum mechanical algorithm for database search*. 1996. arXiv: quant-ph/9605043 [quant-ph].

[8] Aram W. Harrow, Avinatan Hassidim, and Seth Lloyd. "Quantum Algorithm for Linear Systems of Equations". In: *Physical Review Letters* 103.15 (2009). DOI: 10.1103/physrevlett.103.150502.

[9] Lieven M. K. Vandersypen et al. "Experimental realization of Shor's quantum factoring algorithm using nuclear magnetic resonance". In: *Nature* 414.6866 (2001), pp. 883–887.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks". In: *Communications of the ACM* 60 (2012), pp. 84 –90.

[11] Olga Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision* 115 (2015), pp. 211–252.

[12] David Silver et al. "Mastering the game of Go with deep neural networks and tree search". In: *Nature* 529 (2016), pp. 484–489.

[13] John Tromp. "The Number of Legal Go Positions". In: *Computers and Games*. Springer International Publishing, 2016, pp. 183–190.

[14] Tom Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.

[15] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2.5 (1989), pp. 359–366. DOI: https://doi.org/10.1016/0893-6080(89)90020-8.

[16] Radford M. Neal. "Priors for Infinite Networks". In: *Bayesian Learning for Neural Networks*. New York, NY: Springer New York, 1996, pp. 29–53. DOI: 10.1007/978-1-4612-0745-0_2.

[17] Jaehoon Lee et al. "Deep Neural Networks as Gaussian Processes". In: (2018). arXiv: 1711.00165 [stat.ML].

[18] Jaehoon Lee et al. "Wide neural networks of any depth evolve as linear models under gradient descent". In: *Journal of Statistical Mechanics: Theory and Experiment* 2020.12 (2020), p. 124002. DOI: 10.1088/1742-5468/abc62b.

[19] Maria Schuld and Francesco Petruccione. *Machine Learning with Quantum Computers*. 2nd ed. Springer Nature Switzerland, 2021.

[20] Edward Farhi and Hartmut Neven. *Classification with Quantum Neural Networks on Near Term Processors*. 2018. arXiv: 1802.06002 [quant-ph].

[21] M. Cerezo et al. "Variational quantum algorithms". In: *Nature Reviews Physics* 3.9 (2021), 625–644. DOI: 10.1038/s42254-021-00348-9.

[22] Jarrod R. McClean et al. "Barren pleateaus in quantum neural network training landscapes". In: *Nature Communications* 9.1 (2018).

[23] Jack Cunningham and Jun Zhuang. "Investigating and Mitigating Barren Plateaus in Variational Quantum Circuits: A Survey". In: (2025). arXiv: 2407.17706 [quant-ph].

[24] M. Cerezo et al. *Does provable absence of barren plateaus imply classical simulability? Or, why we need to rethink variational quantum computing*. 2024. arXiv: 2312.09121 [quant-ph].

[25] Filippo Girardi and Giacomo De Palma. *Trained quantum neural networks are Gaussian processes*. 2024. arXiv: 2402.08726 [quant-ph].

[26] Anderson Melchor Hernandez et al. *Quantitative convergence of trained quantum neural networks to a Gaussian process*. 2024. arXiv: 2412.03182 [quant-ph].

[27] John Preskill. "Quantum Computing in the NISQ era and beyond". In: *Quantum* 2 (2018), p. 79. DOI: 10.22331/q-2018-08-06-79.

[28] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. 10th Anniversary. Cambridge, UK: Cambridge University Press, 2010.

[29] Ryan LaRose. *A brief history of quantum vs classical computational advantage*. 2024. arXiv: 2412.14703 [quant-ph].

[30] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. *A Quantum Approximate Optimization Algorithm*. 2014. arXiv: 1411.4028 [quant-ph].

[31] Carlos Bravo-Prieto et al. "Variational Quantum Linear Solver". In: *Quantum* 7 (2023), p. 1188. DOI: 10.22331/q-2023-11-22-1188.

[32] Peter D. Johnson et al. "QVECTOR: an algorithm for device-tailored quantum error correction". In: (2017). arXiv: 1711.02249 [quant-ph].

[33] Xiao Yuan et al. "Theory of variational quantum simulation". In: *Quantum* 3 (2019), p. 191.

[34] Jules Tilly et al. "The Variational Quantum Eigensolver: A review of methods and best practices". In: *Physics Reports* 986 (2022), 1–128. DOI: 10.1016/j.physrep.2022.08.003.

[35] Sukin Sim, Peter D. Johnson, and Alán Aspuru-Guzik. "Expressibility and Entangling Capability of Parameterized Quantum Circuits for Hybrid Quantum-Classical Algorithms". In: *Advanced Quantum Technologies* 2.12 (2019). DOI: 10.1002/qute.201900070.

[36] Ira N. Levine. *Quantum Chemistry*. 7th ed. Pearson, 2014.

[37] Alberto Peruzzo et al. "A variational eignevalue solver on a photonic quantum processor". In: *Nature Communications* (2015).

[38] R. Horn and C. Johnson. *Matrix Analysis*. 2nd ed. Cambridge University Press, 2013.

[39] Erwin Kreyszig. *Introductory Functional Analysis with Applications*. New York, NY: Wiley, 1978.

[40] John Napp. *Quantifying the barren plateau phenomenon for a model of unstructured variational ansätze*. 2022. arXiv: 2203.06174 [quant-ph].

[41] Xiaosi Xu et al. *A Herculean task: Classical simulation of quantum computers*. 2023. arXiv: 2302.08880 [quant-ph].

[42] Rick Durrett. *Probability: theory and examples*. Cambridge University Press, 2019.

[43] V. V. Petrov. "Chapter V. Estimates of the Distance Between the Distribution of a Sum of Independent Random Variables and the Normal Distribution". In: *Sums of Independent Random Variables*. Berlin, Heidelberg, New York: Springer-Verlag, 1975, pp. 104–133.

[44] Jordan Bell. *The Berry-Esseen Theorem*. Tech. rep. University of Toronto, 2015.

[45] Hanna Döring, Sabine Jansen, and Kristina Schubert. *The Method of Cumulants for the Normal Approximation*. 2021. arXiv: 2102.01459 [math.PR].

[46] Valentin Féray, Pierre-Loïc Méliot, and Ashkan Nikeghbali. "Fluctuations in the case of lattice distributions". In: *Mod-phi Convergence*. Springer International Publishing, 2016, 17–32. DOI: 10.1007/978-3-319-46822-8_3.

[47] Antonio Anna Mele. "Introduction to Haar Measure Tools in Quantum Information: A Beginner's Tutorial". In: *Quantum* 8 (2024), p. 1340. DOI: 10.22331/q-2024-05-08-1340.

[48] Cédric Villani. *Optimal Transport: Old and New*. Springer Berlin, Heidelberg, 2008.

[49] Nathan Ross. *Fundamentals of Stein's method*. 2011. arXiv: 1109.1880 [math.PR].