

# Quantum Machine Learning

## A Hands-on Tutorial for Machine Learning Practitioners and Researchers

Yuxuan Du<sup>1,\*†</sup>  
Zhan Yu<sup>2,\*</sup>

Xinbiao Wang<sup>1,\*</sup>  
Yang Qian<sup>1,\*</sup>

Naixu Guo<sup>2,\*</sup>  
Kaining Zhang<sup>1,\*</sup>

Min-Hsiu Hsieh<sup>3,†</sup>

Patrick Rebentrost<sup>2,⊥</sup>

Dacheng Tao<sup>1,‡</sup>

<sup>1</sup> College of Computing and Data Science, Nanyang Technological University,  
639798, Singapore

<sup>2</sup> Centre for Quantum Technologies, National University of Singapore, 117543,  
Singapore

<sup>3</sup> Hon Hai Research Institute, Taipei, 114, Taiwan

† [duyuxuan123@gmail.com](mailto:duyuxuan123@gmail.com)

‡ [min-hsiu.hsieh@foxconn.com](mailto:min-hsiu.hsieh@foxconn.com)

⊥ [cqtfpr@nus.edu.sg](mailto:cqtfpr@nus.edu.sg)

‡ [dacheng.tao@ntu.edu.sg](mailto:dacheng.tao@ntu.edu.sg)

\* Equal contributions



# Abstract

This tutorial intends to introduce readers with a background in AI to quantum machine learning (QML)—a rapidly evolving field that seeks to leverage the power of quantum computers to reshape the landscape of machine learning. For self-consistency, this tutorial covers foundational principles, representative QML algorithms, their potential applications, and critical aspects such as trainability, generalization, and computational complexity. In addition, practical code demonstrations are provided in <https://qml-tutorial.github.io/> to illustrate real-world implementations and facilitate hands-on learning. Together, these elements offer readers a comprehensive overview of the latest advancements in QML. By bridging the gap between classical machine learning and quantum computing, this tutorial serves as a valuable resource for those looking to engage with QML and explore the forefront of AI in the quantum era.



# Contents

<b>Preface</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 A First Glimpse of Quantum Machine Learning . . . . .	3
1.1.1 Quantum computers . . . . .	3
1.1.2 Different measures of quantum advantages . . . . .	7
1.1.3 Explored tasks in quantum machine learning . . . . .	9
1.2 Progress of Quantum Machine Learning . . . . .	11
1.2.1 Progress of quantum computers . . . . .	11
1.2.2 Progress of quantum machine learning under FTQC .	13
1.2.3 Progress of quantum machine learning under NISQ .	15
1.2.4 A brief review of quantum machine learning . . . . .	18
1.3 Organization of This Tutorial . . . . .	19
<b>2 Basics of Quantum Computing</b>	<b>21</b>
2.1 From Classical Bits to Quantum Bits . . . . .	21
2.1.1 Classical bits . . . . .	22
2.1.2 Quantum bits (Qubits) . . . . .	22
2.1.3 Density matrix . . . . .	26
2.2 From Digital Logical Circuit to Quantum Circuit Model .	28
2.2.1 Classical digital logical circuit . . . . .	28
2.2.2 Quantum circuit . . . . .	30
2.3 Quantum Read-in and Read-out protocols . . . . .	41
2.3.1 Quantum read-in protocols . . . . .	42
2.3.2 Quantum read-out protocols . . . . .	46
2.4 Quantum Linear Algebra . . . . .	52
2.4.1 Block encoding . . . . .	53
2.4.2 Basic arithmetic for block encodings . . . . .	54
2.4.3 Quantum singular value transformation . . . . .	56

2.5	Code Demonstration . . . . .	58
2.5.1	Read-in implementations . . . . .	58
2.5.2	Block encoding . . . . .	60
2.6	Bibliographic Remarks . . . . .	61
2.6.1	Advanced quantum read-in protocols . . . . .	61
2.6.2	Advanced quantum read-out protocols . . . . .	63
2.6.3	Advanced quantum linear algebra . . . . .	64
<b>3</b>	<b>Quantum Kernel Methods</b>	<b>67</b>
3.1	Classical Kernel Machines . . . . .	68
3.1.1	Motivation of kernel methods . . . . .	68
3.1.2	Dual representation . . . . .	71
3.1.3	Kernel construction . . . . .	73
3.2	Quantum Kernel Machines . . . . .	76
3.2.1	Motivations for quantum kernel machines . . . . .	76
3.2.2	Quantum feature maps and quantum kernel machines	77
3.2.3	Relation between quantum and classical kernel machines	80
3.2.4	Concrete examples of quantum kernels . . . . .	81
3.3	Theoretical Foundations of Quantum Kernel Machines . . . . .	86
3.3.1	Expressivity of quantum kernel machines . . . . .	87
3.3.2	Generalization of quantum kernel machines . . . . .	91
3.4	Code Demonstration . . . . .	98
3.4.1	Classification on MNIST dataset . . . . .	99
3.5	Bibliographic Remarks . . . . .	106
3.5.1	Quantum kernel design . . . . .	107
3.5.2	Theoretical studies of quantum kernels . . . . .	108
3.5.3	Applications of quantum kernels . . . . .	110
<b>4</b>	<b>Quantum Neural Networks</b>	<b>113</b>
4.1	Classical Neural Networks . . . . .	114
4.1.1	Perceptron . . . . .	114
4.1.2	Multilayer perceptron . . . . .	118
4.2	Fault-tolerant Quantum Perceptron . . . . .	122
4.2.1	Grover search . . . . .	122
4.2.2	Online quantum perceptron with quadratic speedups .	124
4.3	Near-term Quantum Neural Networks . . . . .	128
4.3.1	General framework . . . . .	129
4.3.2	Discriminative learning with QNNs . . . . .	134
4.3.3	Generative learning with QNNs . . . . .	135
4.4	Theoretical Foundations of Quantum Neural Networks . . . . .	140

4.4.1	Expressivity and generalization of quantum neural networks . . . . .	142
4.4.2	Trainability of quantum neural networks . . . . .	150
4.5	Code Demonstration . . . . .	155
4.5.1	Quantum classifier . . . . .	155
4.5.2	Quantum patch GAN . . . . .	163
4.6	Bibliographic Remarks . . . . .	171
4.6.1	Discriminative learning with QNN . . . . .	171
4.6.2	Generative learning with QNNs . . . . .	175
<b>5</b>	<b>Quantum Transformer</b>	<b>179</b>
5.1	Classical Transformer . . . . .	180
5.1.1	Tokenization and embedding . . . . .	180
5.1.2	Self-attention . . . . .	182
5.1.3	Residual connection . . . . .	183
5.1.4	Feed-forward network . . . . .	184
5.1.5	Optimization and inference . . . . .	185
5.2	Fault-tolerant Quantum Transformer . . . . .	186
5.2.1	Quantum self-attention . . . . .	188
5.2.2	Quantum residual connection and layer normalization	194
5.2.3	Quantum feedforward neural network . . . . .	196
5.3	Runtime Analysis with Quadratic Speedups . . . . .	198
5.3.1	Overview . . . . .	198
5.3.2	Numerical evidence . . . . .	199
5.4	Code Demonstration . . . . .	201
5.5	Bibliographic Remarks . . . . .	203
<b>6</b>	<b>Conclusion</b>	<b>205</b>
<b>A</b>	<b>Notations Summary</b>	<b>207</b>
<b>B</b>	<b>Concentration Inequality</b>	<b>209</b>
<b>C</b>	<b>Haar Measure and Unitary t-design</b>	<b>213</b>



# Preface

Quantum computers, as next-generation computational devices, harness the quantum principles of superposition and entanglement to process information in ways fundamentally different from classical computers. These unique properties enable quantum computers to address many practical problems that are intractable for classical computers. Although quantum computing is still in its early stages, we have entered an era since 2019 where quantum supremacy has been experimentally demonstrated by several research groups and industrial organizations, underscoring the immense potential of quantum technologies to transform various aspects of everyday life.

Machine learning (ML) is widely regarded as one of the most promising and impactful applications of quantum computing. The ability of quantum computing to accelerate advancements in foundational models, such as generative pre-trained transformers (GPTs), and even pave the way toward artificial general intelligence (AGI), is particularly compelling. Recent progress in both theories and experiments has exhibited the power of quantum machine learning (QML), where the integration of quantum computing with machine learning may lead to novel approaches that outperform classical algorithms by offering faster runtimes, better performance, and reduced data requirements in areas such as computer vision, natural language processing, drug discovery, finance, and fundamental science.

As an interdisciplinary field, the development of QML requires close collaboration between leading scientists and engineers in both quantum computing and artificial intelligence (AI). At the same time, as QML advances alongside the continuous progress of quantum hardware, there is a growing need for expertise from the AI community to drive this emerging field forward. However, the distinct conceptual frameworks and terminologies of quantum and classical computing present significant barriers for researchers and practitioners with a classical machine learning background in understanding the mechanisms behind QML algorithms and the benefits they may offer. Reducing this barrier to entry remains a major challenge within

the community.

To overcome this challenge, we have written this tutorial to deliver a comprehensive introduction to the latest developments in QML, specifically designed for readers with expertise in machine learning. Whether you are an AI researcher, a machine learning practitioner, or a computer science student, this resource will equip you with a solid foundation in the principles and techniques of QML. By bridging the gap between classical ML and quantum computing, this tutorial could serve as a useful resource for those looking to engage with quantum machine learning and explore the forefront of AI in the quantum era.

The Authors  
Feb, 2025

# Chapter 1

## Introduction

The advancement of computational power has always been a driving force behind every major industrial revolution. The invention of the modern computer, followed by the central processing unit (CPU), led to the “digital revolution”, transforming industries through process automation and the rise of information technology. More recently, the development of graphical processing units (GPUs) has powered the era of artificial intelligence (AI) and big data, enabling breakthroughs in areas such as intelligent transportation systems, autonomous driving, scientific simulations, and complex data analysis. However, as we approach the physical and practical limits of Moore’s law—the principle that the number of transistors on a chip doubles approximately every two years—traditional computing devices like CPUs and GPUs are nearing the end of their scaling potential. The ever-growing demand for computational power, driven by the exponential increase in data and the complexity of modern applications, necessitates new computing paradigms. Among the leading candidates to meet these challenges are **quantum computers** ([Feynman, 2017](#)), which leverage the unique principles of quantum mechanics such as superposition and entanglement to process information in ways that classical systems cannot, with the potential to revolutionize diverse aspects of daily life.

One of the most concrete and direct ways to understand the potential of quantum computers is through the framework of complexity theory ([Watrous, 2008](#)). Theoretical computer scientists have demonstrated that quantum computers can efficiently solve problems within the BQP (Bounded-Error Quantum Polynomial Time) complexity class, meaning these problems can be solved in polynomial time by a quantum computer. In contrast, classical computers are limited to efficiently solving problems within the P

(Polynomial Time) complexity class. While it is widely believed, though not proven, that  $P \subseteq BQP$ , this suggests that quantum computers can provide exponential speedups for certain problems in  $BQP$  that are intractable for classical machines.

A prominent example of such a problem is large-number factorization, which forms the basis of RSA cryptography. Shor’s algorithm ([Shor, 1999](#)), a quantum algorithm, can factor large numbers in polynomial time, while the most efficient known classical factoring algorithm requires super-polynomial time. For instance, breaking an RSA-2048 bit encryption key would take a classical computer approximately 300 trillion years, whereas an *ideal* quantum computer could complete the task in around 10 seconds. However, constructing ‘ideal’ quantum computers remains a significant challenge. As will be discussed in later chapters, based on current fabrication techniques, this task could potentially be completed in approximately 8 hours using a noisy quantum computer with a sufficient number of **qubits**—the fundamental units of quantum computation ([Gidney and Ekerå, 2021](#)).

The convergence of the computational power offered by quantum machines and the limitations faced by AI models has led to the rapid emergence of the field: **quantum machine learning** (QML) ([Biamonte et al., 2017](#)). In particular, the challenges in modern AI stem from the neural scaling law ([Kaplan et al., 2020](#)), which posits that “bigger is often better.” Since 2020, this principle has driven the development of increasingly colossal models, featuring more complex architectures and an ever-growing number of parameters. However, this progress comes at an immense cost. For instance, training a model like ChatGPT on a single GPU would take approximately 355 years, while the cloud computing costs for training such large models can reach tens of thousands of dollars.

These staggering costs present a critical barrier to the future growth of AI. Quantum computing, celebrated for its extraordinary computational capabilities, holds the potential to overcome these limitations. It offers the possibility of advancing models like generative pre-trained transformers (GPTs) and accelerating progress toward artificial general intelligence (AGI). Quantum computing, and more specifically QML, represents a paradigm shift, moving from the classical “it from bit” framework to the quantum “it from qubit” era, with the potential to reshape the landscape of AI and computational science.

## 1.1 A First Glimpse of Quantum Machine Learning

So, what exactly is quantum machine learning (QML)? In its *simplest* terms, the focus of this tutorial on QML can be summarized as follows (see Chapter 1.1.3 for the systematic overview).

### Quantum machine learning (informal)

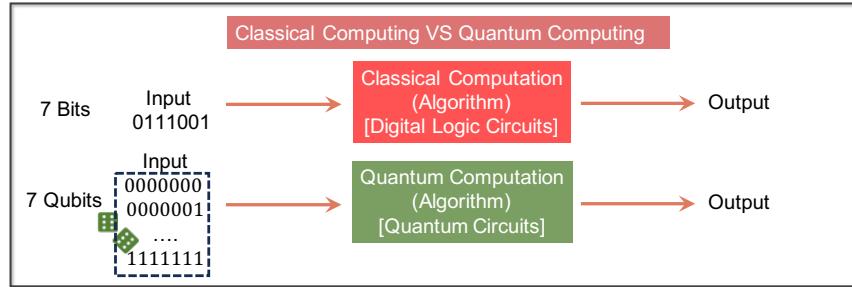
QML explores learning algorithms that can be executed on quantum computers to accomplish specified tasks with potential advantages over classical implementations.

The three key elements in the above interpretation are: *quantum processors*, *specified tasks*, and *advantages*. In what follows, let us elucidate the specific meaning of each of these terms, providing the necessary foundation for a deeper understanding of the mechanisms and potential of QML.

### 1.1.1 Quantum computers

The origins of quantum computing can be traced back to 1980 when Paul Benioff introduced the *quantum Turing machine* (Benioff, 1982), a quantum analog of the classical Turing machine designed to describe computational models through the lens of quantum theory. Since then, several models of quantum computation have emerged, including *circuit-based quantum computation* (Nielsen and Chuang, 2011), *one-way quantum computation* (Raussendorf and Briegel, 2001), *adiabatic quantum computation* (Albash and Lidar, 2018), and *topological quantum computation* (Kitaev, 2003). All of them have been shown to be computationally equivalent to the quantum Turing machine, meaning that a perfect implementation of any one of these models can simulate the others with no more than polynomial overhead. Given the prominence of **circuit-based quantum computers** in both the research and industrial communities and their rapid advancement, this tutorial will focus primarily on this model of quantum computing.

Quantum computing gained further momentum in the early 1980s when physicists faced an exponential increase in computational overhead while simulating quantum dynamics, particularly as the number of particles in a system grew. This “curse of dimensionality” prompted Yuri Manin and Richard Feynman to independently propose leveraging quantum phenomena to build quantum computers, arguing that such devices would be far more



**Figure 1.1: The paradigm between classical and quantum computing.** The mechanisms between classical and quantum computing are very similar, where both of them involve input, computation, and output. In classical computing, the input refers to a bit-string, the computation part refers to the digital logic circuits, and the output also refers to a bit-string. In quantum computing, the input is a single- or multi-qubit state. The computation involves quantum circuits. And the output of quantum computers requires quantum measurement, which aims to extract information from the quantum world to the classical world.

efficient for simulating quantum systems than classical computers.

However, as a universal computing device, the potential of quantum computers extends well beyond quantum simulations. In the 1990s, [Shor \(1999\)](#) developed a groundbreaking quantum algorithm for large-number factorization, posing a serious threat to widely used encryption protocols such as RSA and Diffie–Hellman. In 1996, Grover’s algorithm demonstrated a quadratic speedup for unstructured search problems ([Grover, 1996](#)), a task with broad applications. Since then, the influence of quantum computing has expanded into a wide range of fields, with new quantum algorithms being developed to achieve runtime speedups in areas such as finance ([Herman et al., 2023](#)), drug design ([Santagati et al., 2024](#)), optimization ([Abbas et al., 2024](#)), and, most relevant to this tutorial, machine learning.

An intuitive way to understand why quantum computers can outperform classical computers is by comparing their fundamental components. As illustrated in Figure 1.1, both types of computers consist of three fundamental components: input, the computational process, and the output. The implementation of these three components in classical and quantum computing is summarized in Table 1.1.

The advantages of quantum computers stem primarily from the key distinctions between classical bits and quantum bits (**qubits**), as well as be-

Table 1.1: **Comparison between classical and quantum computing.**

	Classical	Quantum
Input	Binary bits	Quantum bits
Computation	Digital logical circuits	Quantum circuits
Output	Retrieve solution	Quantum measurements

tween digital logic circuits and **quantum circuits**, as outlined below:

- *Bits versus Qubits.* A classical bit is a binary unit that takes on a value of either 0 or 1. In contrast, a quantum bit, or qubit, can exist in a superposition of both 0 and 1 simultaneously, represented by a *two-dimensional vector* where the entries correspond to the *probabilities* of the qubit being in each state.

Furthermore, while classical bits follow the Cartesian product rule, qubits adhere to the tensor product rule. This distinction implies that an  $N$ -qubit system is described by a  $2^N$ -dimensional vector, allowing quantum systems to encode information exponentially with  $N$ —far surpassing the capacity of classical bits. Table 1.2 summarizes the mathematical expressions of classical and quantum bits.

Table 1.2: **Mathematical representations of  $N$ -(quantum) bits in classical and quantum computers.** Here the symbols ‘ $\dagger$ ’ and  $\mathbb{C}$  denote the transpose conjugation and complex space, respectively.

	Classical	Quantum
Single bit ( $N = 1$ )	$x \in \{0, 1\}$	$[a_1, a_2]^\dagger \in \mathbb{C}^2$ s.t. $a_1^2 + a_2^2 = 1$
Multiple bits ( $N > 1$ )	$x \in \{0, 1\}^N$	s.t. $[a_1, a_2, \dots, a_{2N}]^\dagger \in \mathbb{C}^{2^N}$ s.t. $a_1^2 + a_2^2 + \dots + a_{2N}^2 = 1$

- *Digital logic circuits versus quantum circuits.* Classical computers rely on digital logic circuits composed of logic gates that perform operations on bits in a deterministic manner, as illustrated in Figure 1.1. In contrast, quantum circuits consist of **quantum gates**, which act on single or multiple qubits to modify their states—the probability amplitudes  $a_1, \dots, a_{2N}$ , as shown in Table 1.2. Owing to the universality of quantum gates, for any given input qubit state, there always exists a specific quantum circuit capable of transforming the input

state into one corresponding to the target solution—a particular probability distribution. For certain probability distributions, a quantum computer can use a polynomial number of quantum gates relative to the qubit count  $N$  to generate the distribution, whereas classical computers require an exponential number of gates with  $N$  to achieve the same result. This difference underpins the quantum advantage.

- The readout process in quantum computing differs fundamentally from that in classical computing, as it involves **quantum measurements**, which extract information from a quantum system and translate it into a form that can be interpreted by classical systems. For problems in quantum physics and chemistry, quantum measurements can reveal far more useful information than classical simulations of the same systems, enabling significant runtime speedups in obtaining the desired physical properties.

The formal definitions of quantum computing are presented in Chapter 2. As we will see, the power of quantum computers is primarily determined by two factors: the number of qubits and the quantum gates, as well as their respective qualities. The term “qualities” refers to the fact that fabricating quantum computers is highly challenging, as both qubits and quantum gates are prone to errors. These qualities are measured using various physical metrics. One commonly used metric is *quantum volume*  $V_Q$  (Cross et al., 2019), which quantifies a quantum computer’s capabilities by accounting for both its error rates and overall performance. Mathematically, the quantum volume represents the maximum size of square quantum circuits that the computer can successfully implement to achieve the *heavy output generation problem*, i.e.,

$$\log_2(V_Q) = \arg \max_m \min(m, d(m)), \quad (1.1)$$

where  $m \leq N$  is a number of qubits selected from the given  $N$ -qubit quantum computer, and  $d(m)$  is the number of qubits in the largest square circuits for which we can reliably sample heavy outputs with probability greater than  $2/3$ . The heavy output generation problem discussed here stems from proposals aimed at demonstrating quantum advantage. That is, if a quantum computer is of sufficiently high quality, we should expect to observe heavy outputs frequently across a range of random quantum circuit families. For illustration, Table 1.3 summarizes the progress of quantum computers as of 2024.

Table 1.3: Progress of quantum computers up to December 2024.

Date	$\log_2(V_Q)$	N	Manufacturer	System Name
Dec, 2024	-	105	Google	Willow
Aug, 2024	21	56	Quantinuum	H2-1
Jul, 2024	9	156	IBM	Heron
Jun, 2023	19	20	Quantinuum	H1-1
Sep, 2022	13	20	Quantinuum	H1-1
Apr, 2022	12	12	Quantinuum	H1-2
Jul, 2021	10	10	Honeywell	H1
Nov, 2020	7	10	Honeywell	H1
Aug, 2020	6	27	IBM	Falcon r4

**Remark**

Note that quantum volume is not the unique metric for evaluating the performance of quantum computers. There are several other metrics that assess the power of quantum processors from different perspectives. For instance, *Circuit Layer Operations Per Second* (CLOPS) ([Wack et al., 2021](#)) measures the computing speed of quantum computers, reflecting the feasibility of running practical calculations that involve a large number of quantum circuits. Additionally, *effective quantum volume* ([Kchedzhi et al., 2024](#)) provides a more nuanced comparison between noisy quantum processors and classical computers, considering factors such as error rates and noise levels. These metrics, among others, offer a more comprehensive understanding of the strengths and limitations of quantum computers across various applications.

### 1.1.2 Different measures of quantum advantages

What do we mean when we refer to quantum advantage? Broadly, quantum advantage is demonstrated when quantum computers can solve a problem *more efficiently* than classical computers. However, the notion of “efficiency” in this context is not uniquely defined.

The most common measure of efficiency is *runtime complexity*. By harnessing quantum effects, certain computations can be accelerated significantly—sometimes even exponentially—enabling tasks that are otherwise infeasible for classical computers. A prominent example is Shor’s algorithm, which achieves an exponential speedup in large-number factorization relative to the best classical algorithms. In terms of runtime complexity, the

quantum advantage is achieved when the upper bound of a quantum algorithm's runtime for a given task is *lower than* the theoretical lower bound of all possible classical algorithms for the same task.

In quantum learning theory (Arunachalam and De Wolf, 2017), efficiency can also be measured by *sample complexity*, particularly within the Probably Approximately Correct (PAC) learning framework, which is central to this tutorial. In this context, sample complexity is defined as the number of interactions (e.g., quires of target quantum systems or measurements) required for a learner to achieve a desired prediction accuracy below a specified threshold. Here, the quantum advantage is realized when the upper bound on the sample complexity of a quantum learning algorithm for a given task is *lower than* the lower bound of all classical learning algorithms. While low sample complexity is a necessary condition for efficient learning, it does not guarantee practical efficiency alone; for example, identifying useful training examples within a small sample size may still require substantial computational time.

#### Remark

(Difference of sample complexity in classical and quantum ML). In classical ML, sample complexity typically refers to the number of training examples required for a model to generalize effectively, such as the number of labeled images needed to train an image classifier. In quantum ML, however, sample complexity can take on varied meanings depending on the context, as shown below.

- Quantum state tomography (see Chapter 2.3.2). Here the sample complexity refers to the number of measurements required to accurately reconstruct the quantum state of a system.
- Evaluation of the generalization ability of quantum neural networks (see Chapter 4.4). Here the sample complexity refers to the number of input-output pairs needed to train the network to approximate a target function, similar to classical ML.
- Quantum system learning. Here the sample complexity often refers to the number of queries to interact with the target quantum system, such as the number of times a system must be probed to learn its Hamiltonian dynamics.

In addition to sample complexity, another commonly used measure in

quantum learning theory is *quantum query complexity*, particularly within the frameworks of quantum statistical learning and quantum exact learning. As these frameworks are not the primary focus of this tutorial, interested readers are referred to ([Anshu and Arunachalam, 2024](#)) for a more detailed discussion.

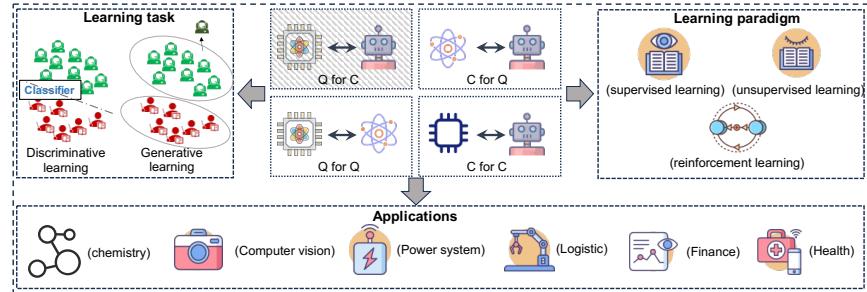
Quantum advantage can be pursued through *two main approaches*. The first involves identifying problems with quantum circuits that demonstrate provable advantages over classical counterparts in the aforementioned measures ([Harrow and Montanaro, 2017](#)). Such findings deepen our understanding of quantum computing’s potential and expand its range of applications. However, these quantum circuits often require substantial quantum resources, which are currently beyond the reach of near-term quantum computers. Additionally, for many tasks, analytically determining the upper bound of classical algorithm complexities is challenging.

These challenges have motivated a second approach: demonstrating that current quantum devices can perform accurate computations on a scale that exceeds brute-force classical simulations—a milestone known as “quantum utility.” Quantum utility refers to quantum computations that yield reliable, accurate solutions to problems beyond the reach of brute-force classical methods and otherwise accessible only through classical approximation techniques ([Kim et al., 2023](#)). This approach represents a step toward practical computational advantage with noise-limited quantum circuits. Reaching the era of quantum utility signifies that quantum computers have attained a level of scale and reliability enabling researchers to use them as effective tools for scientific exploration, potentially leading to groundbreaking new insights.

### 1.1.3 Explored tasks in quantum machine learning

What are the main areas of focus in QML? QML research is extensive and can be broadly divided into four primary sectors, each defined by the nature of the computing resources (whether the computing device is quantum (**Q**) or classical (**C**)) and the type of data involved (whether generated by a quantum (**Q**) or classical (**C**) system). The explanations of these four sectors are as follows.

**CC Sector.** The **CC** sector refers to classical data processed on classical systems, representing traditional machine learning. Here, classical ML algorithms run on classical processors (e.g., CPUs and GPUs) and



**Figure 1.2: Different research directions in QML.** QML can be categorized into four types based on the interplay of quantum (**Q**) and classical (**C**) systems: **Q** for **Q** (quantum algorithms for quantum data), **Q** for **C** (quantum algorithms for classical data), **C** for **Q** (classical algorithms for quantum data), and **C** for **C** (classical algorithms for classical data). This tutorial primarily focuses on the **Q** for **C** category. Beyond the role of the learner and system, QML can also be classified by learning tasks (discriminative and generative learning), learning paradigms (supervised, unsupervised, and reinforcement learning), and diverse applications such as chemistry, computer vision, power systems, logistics, finance, and healthcare.

are applied to classical datasets. A typical example is using neural networks to classify images of cats and dogs.

**CQ Sector:** The **CQ** sector involves using classical ML algorithms on classical processors to analyze quantum data collected from quantum systems. Typical examples include applying classical neural networks to classify quantum states, estimating properties of quantum systems from measurement data, and employing classical regression models to predict outcomes of quantum experiments.

**QC Sector.** The **QC** sector involves developing QML algorithms that run on quantum processors (QPUs) to process classical data. In this context, quantum computing resources are leveraged to enhance or accelerate the analysis of classical datasets. Typical examples include applying QML algorithms, such as quantum neural networks and quantum kernels, to improve pattern recognition in image analysis.

**QQ Sector.** The **QQ** sector involves developing QML algorithms executed on QPUs to process quantum data. In this context, quantum computing resources are leveraged to reduce the computational cost

of analyzing and understanding complex quantum systems. Typical examples include using quantum neural networks for quantum state classification and applying quantum-enhanced algorithms to simulate quantum many-body systems.

The classification above is not exhaustive. As illustrated in Figure 1.2, each sector can be further subdivided based on various learning paradigms, such as discriminative vs. generative learning or supervised, unsupervised, and semi-supervised learning. Additionally, each sector can be further categorized according to different application domains, such as finance, healthcare, and logistics.

#### Remark

The primary focus of this tutorial is on the **QC** and **QQ** sectors. For more details on **CQ**, interested readers can refer to ([Schuld et al., 2015](#); [Dunjko and Briegel, 2018](#); [Carleo et al., 2019](#)).

## 1.2 Progress of Quantum Machine Learning

Huge efforts have been made to the **QC** and **QQ** sectors to determine which tasks and conditions allow QML to offer computational advantages over classical machine learning. In this regard, to provide a clearer understanding of QML’s progress, it is essential to first review recent advancements in quantum computers, the foundational substrate for quantum algorithms.

### 1.2.1 Progress of quantum computers

The novelty and inherent challenges of utilizing quantum physics for computation have driven the development of various computational architectures, giving rise to the formalized concept of circuit-based quantum computers, as discussed in Chapter 1.1.1. In pursuit of this goal, numerous companies and organizations are striving to establish their architecture as the leading approach and to be the first to demonstrate practical utility or quantum advantage on a large-scale quantum device.

Common architectures currently include superconducting qubits (employed by IBM and Google), ion-trap systems (pioneered by IonQ), and Rydberg atom systems (developed by QuEra), each offering distinct advantages ([Cheng et al., 2023](#)). Specifically, superconducting qubits excel in scalability and fast gate operations ([Huang et al., 2020a](#)), while ion-trap systems are

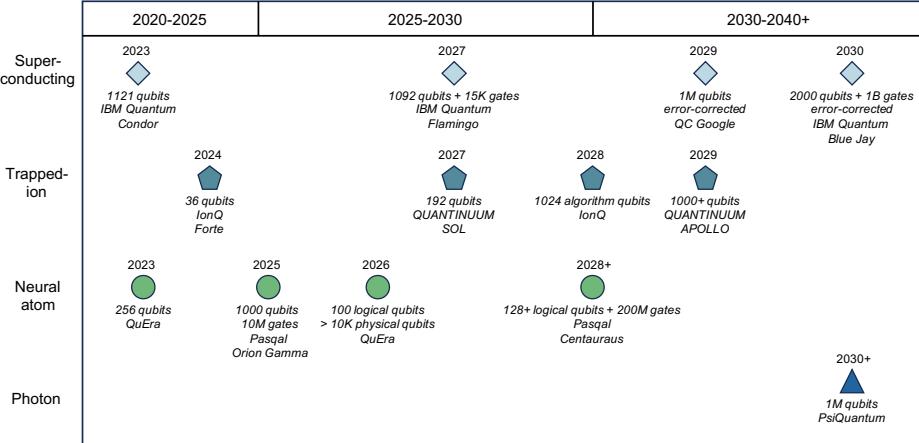


Figure 1.3: Common quantum architectures and roadmaps from different quantum companies.

known for their high coherence times, precise control over individual qubits, and full connectivity of all qubits (Bruzewicz et al., 2019). Moreover, Rydberg atom systems enable flexible qubit connectivity through highly controllable interactions (Morgado and Whitlock, 2021). Besides these architectures, integrated photonic quantum computers are emerging as promising alternatives for robust and scalable quantum computation.

Despite recent advances, today’s quantum computers remain highly sensitive to environmental noise and prone to quantum decoherence, lacking the stability needed for fault-tolerant operation. This results in qubits, quantum gates, and quantum measurements that are inherently imperfect, introducing errors that can lead to incorrect outputs. To capture this stage in quantum computing, John Preskill coined the term “noisy intermediate-scale quantum” (NISQ) era (Preskill, 2018), which describes the current generation of quantum processors. These processors feature up to thousands of qubits, but their capabilities are restricted with error-prone gates and limited coherence times.

In the NISQ era, notable achievements have been made alongside new challenges. Industrial and academic teams, such as those at Google and USTC, have demonstrated quantum advantages on specific sampling tasks, where the noisy quantum computers they fabricated outperform classical computers in computational efficiency (Arute et al., 2019; Wu et al., 2021). However, most quantum algorithms that theoretically offer substantial runtime speedups depend on fault-tolerant, error-free quantum systems—capabilities

that remain beyond the reach of current technology.

At this pivotal stage, the path forward in quantum computing calls for progress on both hardware and algorithmic fronts.

On the hardware side, it is essential to continuously improve qubit count, coherence times, gate fidelities, and the accuracy of quantum measurements across various quantum architectures. Once the number and quality of qubits surpass certain thresholds, *quantum error correction* codes can be implemented (Nielsen and Chuang, 2011), paving the way for fault-tolerant quantum computing (FTQC). Broadly, quantum error correction uses redundancy and entanglement to detect and correct errors without directly measuring the quantum state, thus preserving coherence. Advancements in quantum processors will enable a progression from the **NISQ** era to the early FTQC era, ultimately reaching the fully FTQC era.

On the algorithmic side, two key questions must be addressed:

- (Q1) How can **NISQ** devices be utilized to perform meaningful computations with practical utility?
- (Q2) What types of quantum algorithms can be executed on early fault-tolerant and fully fault-tolerant quantum computers to realize the potential of quantum computing in real-world applications?

Progress on either question could have broad implications. A positive answer to (Q1) would suggest that **NISQ** quantum computers have immediate practical applicability, while advancements in (Q2) would expand the scope and impact of quantum computing as more robust, fault-tolerant systems become feasible. In the following two sections, we examine recent progress in QML concerning these two questions.

### 1.2.2 Progress of quantum machine learning under FTQC

A key milestone in FTQC-based QML algorithms is the quantum linear equations solver introduced by Harrow et al. (2009). Many machine learning models rely on solving linear equations, a computationally intensive task that often dominates the overall runtime due to the polynomial scaling of complexity with matrix size. The HHL algorithm provides a breakthrough by reducing runtime complexity to poly-logarithmic scaling with matrix size, given that the matrix is well-conditioned and sparse. This advancement is highly significant for AI, where datasets frequently reach sizes in the millions or even billions.

The exponential runtime speedup achieved by the HHL algorithm has garnered significant attention from the research community, highlighting the

potential of quantum computing in AI. Following this milestone, a body of work has emerged that employs the quantum matrix inversion techniques developed in HHL (or its variants) as subroutines in the design of various FTQC-based QML algorithms, offering runtime speedups over their classical counterparts (Montanaro, 2016; Dalzell et al., 2023). Notable examples include quantum principal component analysis (Lloyd et al., 2014) and quantum support vector machines (Rebentrost et al., 2014).

Another milestone in FTQC-based QML algorithms is the quantum singular value transformation (QSVT), proposed by Gilyén et al. (2019). QSVT enables polynomial transformations of the singular values of a linear operator embedded within a unitary matrix, offering a unifying framework for various quantum algorithms. It has connected and enhanced a broad range of quantum techniques, including amplitude amplification, quantum linear system solvers, and quantum simulation methods. Compared to the HHL algorithm for solving linear equations, QSVT provides improved scaling factors, making it a more efficient tool for addressing these problems in the context of QML.

In addition to advancements in linear equation solving, another promising line of research in FTQC-based QML focuses on leveraging quantum computing to enhance deep neural networks (DNNs) rather than traditional machine learning models. This research track has two main areas of focus. The first is the acceleration of DNN optimization, with notable examples including the development of efficient quantum algorithms for dissipative differential equations to expedite (stochastic) gradient descent, as well as Quantum Langevin dynamics for optimization (Chen et al., 2023; Liu et al., 2024a). The second area centers on advancing Transformers using quantum computing. In Chapter 5, we will discuss in detail how quantum computing can be employed to accelerate Transformers during the inference stage.

### Remark

However, there are several critical caveats of the HHL-based QML algorithms. First, the assumption of efficiently preparing the quantum states corresponding to classical data runtime is very strong and may be impractical in the dense setting. Second, the obtained result  $\mathbf{x}$  is still in the quantum form  $|\mathbf{x}\rangle$ . Note that extracting one entry of  $|\mathbf{x}\rangle$  into the classical form requires  $O(\sqrt{N})$  runtime, which collapses the claimed exponential speedups. The above two issues amount to the read-in and read-out bottlenecks in QML (Aaronson, 2015). The

last caveat is that the employed strong quantum input model such as quantum random access memory (QRAM) (Giovannetti et al., 2008) leads to an inconclusive comparison. Through exploiting a classical analog of QRAM as the input model, there exist efficient classical algorithms to solve recommendation systems in poly-logarithmic time in the size of input data.

### 1.2.3 Progress of quantum machine learning under NISQ

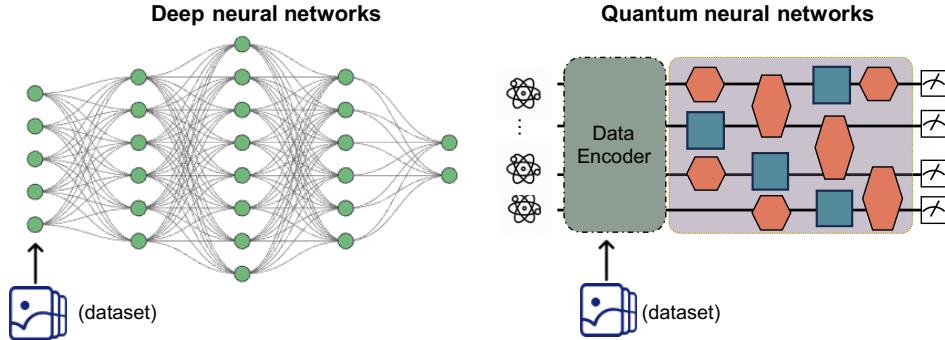
The work conducted by Havlíček et al. (2019) marked a pivotal moment for QML in the NISQ era. This study demonstrated the implementation of quantum kernel methods and quantum neural networks (QNNs) on a 5-qubit superconducting quantum computer, highlighting potential quantum advantages from the perspective of complexity theory. Unlike the aforementioned FTQC algorithms, quantum kernel methods and QNNs are flexible and can be effectively adapted to the limited quantum resources available in the NISQ era. These demonstrations, along with advancements in quantum hardware, sparked significant interest in exploring QML applications using NISQ quantum devices. We will delve into quantum kernel methods and QNNs in Chapter 3 and Chapter 4, respectively.

#### Quantum neural networks (informal)

A quantum neural network (QNN) is a hybrid model that leverages quantum computers to implement trainable models similar to classical neural networks, while using classical optimizers to complete the training process.

As shown in Figure 1.4, the mechanisms of QNNs and deep neural networks (DNNs) are almost the same, whereas the only difference is the way of implementing the trainable model. This difference gives the potential of quantum learning models to solve complex problems beyond the reach of classical neural networks, opening new frontiers in many fields. Roughly speaking, research in QNNs and quantum kernel methods has primarily focused on three key areas: (I) *quantum learning models and applications*, (II) *the adaptation of advanced AI topics to QML*, and (III) *theoretical foundations of quantum learning models*. A brief overview of each category is provided below.

(I) QUANTUM LEARNING MODELS AND APPLICATIONS. This category



**Figure 1.4: Mechanisms of DNNs and QNNs.** Both DNNs and QNNs follow an iterative approach. At each iteration, they take input data, process it through multiple layers, and produce an output prediction. The key difference between DNNs and QNNs is the way of implementing their learning models.

focuses on implementing various DNNs on NISQ quantum computers to tackle a wide range of tasks.

From a model architecture perspective, quantum analogs of popular classical machine learning models have been developed, including quantum versions of multilayer perceptrons (MLPs), autoencoders, convolutional neural networks (CNNs), recurrent neural networks (RNNs), extreme learning machines, generative adversarial networks (GANs), diffusion models, and Transformers. Some of these QNN structures have even been validated on real quantum platforms, demonstrating the feasibility of applying quantum algorithms to tasks traditionally dominated by classical deep learning ([Cerezo et al., 2021a](#); [Li and Deng, 2022](#); [Tian et al., 2023](#)).

From an application perspective, QML models implemented on NISQ devices have been explored across diverse fields, including fundamental science, image classification, image generation, financial time series prediction, combinatorial optimization, healthcare, logistics, and recommendation systems. These applications demonstrate the broad potential of QML in the NISQ era, though achieving full quantum advantage in these areas remains an ongoing challenge ([Bharti et al., 2022](#); [Cerezo et al., 2022](#)).

(II) ADAPTATION OF ADVANCED AI TOPICS TO QML. Beyond model design, advanced topics from AI have been extended to QML, aiming to enhance the performance and robustness of different QML models. Examples include quantum architecture search ([Du et al., 2022a](#)) (the quantum

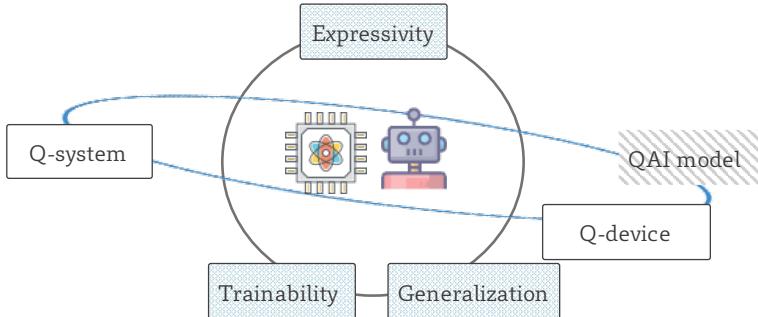


Figure 1.5: The learnability of quantum machine learning models.

equivalent of neural architecture search), advanced optimization techniques (Stokes et al., 2020), and pruning methods to reduce the complexity of quantum models (Sim et al., 2021; Wang et al., 2023a). Other areas of active research include adversarial learning (Lu et al., 2020), continual learning (Jiang et al., 2022), differential privacy (Du et al., 2021a; Watkins et al., 2023), distributed learning (Du et al., 2022b), federated learning (Ren et al., 2023), and interpretability within the context of QML (Pira and Ferrie, 2024). These techniques have the potential to significantly improve the efficiency and effectiveness of QML models, addressing some of the current limitations of NISQ devices.

(III) THEORETICAL FOUNDATIONS. Quantum learning theory (Banchi et al., 2023) has garnered increasing attention, aiming to compare the capabilities of different QML models and to identify the theoretical advantages of QML over classical machine learning models. As shown in Figure 1.5, the learnability of QML models can be evaluated across three key dimensions: expressivity, trainability, and generalization capabilities. Below, we provide a brief overview of each measure.

- Trainability. This area examines how the design of QNNs influences their convergence properties, including the impact of system noise and measurement errors on the ability to converge to local or global minima.
- Expressivity. Researchers investigate how the number of parameters and the structure of QNNs affect the size of the hypothesis space they can represent. A central question is whether QNNs and quantum kernels can efficiently represent functions or patterns that classical neural networks cannot, thereby offering potential quantum advantage.

- Generalization. This focuses on understanding how the gap between training and test error evolves with the size of the dataset, the structure of QNNs or quantum kernels, and the number of parameters. The goal is to determine whether QML models can generalize more effectively than classical models, particularly in the presence of noisy data or when training data is limited.

The combination of advancements in model design, application domains, and theoretical understanding is driving the progress of QML in the NISQ era. Although the field is still in its early stages, the progress achieved thus far provides promising insights into the potential of quantum computing to enhance conventional AI. As quantum hardware continues to evolve, further breakthroughs are expected, potentially unlocking new possibilities for practical QML applications.

#### Remark

It is important to note that QNNs and quantum kernel methods can also be considered FTQC algorithms when executed on fully fault-tolerant quantum computers. The reason these algorithms are discussed in the context of NISQ devices is their flexibility and robustness, making them well-suited to the limitations of current quantum hardware.

#### 1.2.4 A brief review of quantum machine learning

Unlike quantum hardware, where the number of qubits has rapidly scaled from zero to thousands, the development of QML algorithms—and quantum algorithms more broadly—has taken an inverse trajectory, transitioning from FTQC to NISQ devices. This shift reflects the move from idealized theoretical frameworks to practical implementations. The convergence of quantum hardware and QML algorithms, where the quantum resources required by these algorithms become attainable on real quantum computers, enables researchers to experimentally evaluate the power and limitations of various quantum algorithms.

Based on the minimum quantum resources required to complete learning tasks, we distinguish between FTQC algorithms, discussed in Chapter 1.2.2, and NISQ algorithms, including QNNs and quantum kernel methods, in Chapter 1.2.3. FTQC-based QML algorithms necessitate error-corrected quantum computers with tens of billions of qubits—an achievement that remains far from realization. In contrast, QNNs and quantum kernels are

more flexible and can be executed on both NISQ and FTQC devices, depending on the available resources.

As quantum hardware continues to progress, the development of QML algorithms must evolve in tandem. A promising direction is to integrate FTQC algorithms with QNNs and quantum kernel methods, creating new QML algorithms that can be run on current quantum processors while offering enhanced quantum advantages across various tasks.

### 1.3 Organization of This Tutorial

To encourage and enable computer scientists to engage with the rapidly growing field of quantum AI, we provide this hands-on tutorial that revisits QML algorithms from a *computer science perspective*. With this aim, the tutorial is designed to balance theory, practical implementations, and applications, making it suitable for both researchers and practitioners with a background in classical machine learning. The tutorial is divided into the following chapters:

Chapter 2: **BASICS OF QUANTUM COMPUTING.** Before delving into QML, this chapter lays the groundwork by introducing the fundamental concepts of quantum computing. It covers the transition from classical bits to quantum bits, explains quantum circuit models, illustrates how quantum systems interface with classical systems through quantum read-in and read-out mechanisms, and presents some fundamental concepts of quantum linear algebra. By the end of this chapter, you will understand that a solid grasp of linear algebra is all you need to comprehend the basics of quantum computing.

Chapters 3, 4 and 5: **CLASSICAL ML MODELS EXTENDED TO QUANTUM FRAMEWORKS.** Each of these chapters follows a consistent structure, starting with a review of the classical model and progressing to its quantum extension—Quantum kernel methods in Chapter 3, Quantum neural networks in Chapter 4, and Quantum Transformers in Chapter 5. This unified structure enables readers to clearly understand how classical machine learning models can be translated into quantum implementations and how quantum computers may offer computational advantages.

Appendix. The Appendix serves as a supplementary resource, providing a summary of notations and essential mathematical tools that are omitted from the main text for brevity. In particular, it includes basic introduction of concentration inequalities, the Haar measure, and other foundational concepts relevant to the tutorial.

To provide a clear and comprehensive learning experience, each chapter is composed of the following parts:

1. Classical foundations and quantum model construction. Each chapter begins with a review of the classical version of the model, ensuring that readers are well-acquainted with the foundational concepts before exploring their quantum adaptations. After this review, we introduce quantum versions of the models, focusing on implementations based on NISQ, FTQC, or both.
2. Theoretical analysis. There is nothing more practical than a good theory. In this tutorial, each chapter provides a theoretical analysis of the learnability of QML models, focusing on key aspects such as expressivity, trainability, and generalization capabilities.

To ensure a balance between depth and self-consistency, this tutorial provides proof for the most significant theoretical results, as highlighted by **Theorems and Lemmas**. For results that are less central to the main content of this tutorial, we present them as **Facts** and include appropriate references, allowing readers to easily locate the complete proofs if desired.

3. Code implementation. “Talk is cheap, show me the code.” To provide a practical, hands-on learning experience, each chapter includes code implementations using real-world datasets (to be specified). This section walks readers through the process of implementing quantum models on simulated or real quantum hardware. All numerical examples illustrated in this tutorial are available in <https://qml-tutorial.github.io/>, accompanied by Jupyter Notebooks.

Instead of building everything from scratch, we leverage the well-established PennyLane library for implementation (Bergholm et al., 2018). This choice does not imply any specific preference. Other quantum computing libraries, such as Qiskit (Javadi-Abhari et al., 2024), Cirq (Developers, 2024), and TensorFlow Quantum (Broughton et al., 2020), can also be used, offering similar capabilities and flexibility.

4. Frontier topics and future directions. Each chapter concludes with an exploration of cutting-edge topics and emerging challenges in the field. This part highlights open research problems, ongoing developments, and potential future directions for the quantum versions of each model, providing insights into where the field may be headed.

## Chapter 2

# Basics of Quantum Computing

In this chapter, we introduce the fundamental concepts of quantum computation, such as quantum states, quantum circuits, and quantum measurements, along with key topics in quantum machine learning, including quantum read-in, quantum read-out, and quantum linear algebra. These foundational elements are essential for understanding quantum machine learning algorithms and will be repeatedly referenced throughout the subsequent chapters.

This chapter is organized as follows: Chapter 2.1 introduces quantum bits and their mathematical representations; Chapter 2.2 covers quantum circuits, including quantum gates, quantum channels, and quantum measurements; Chapter 2.3 discusses how to encode classical data into quantum systems and extract classical information from quantum states; Chapter 2.4 delves into quantum linear algebra; Chapter 2.5 provides practical coding exercises to reinforce these concepts; and finally, Chapter 2.6 presents recent advancements in efficient quantum read-in and read-out techniques for further exploration.

### 2.1 From Classical Bits to Quantum Bits

In this section, we define quantum bits (qubits) and present the mathematical tools used to describe quantum states. We begin by discussing classical bits and then transition to their quantum counterparts. We recommend interested readers consult the textbook ([Nielsen and Chuang, 2011](#)) for the detailed explanations.

### 2.1.1 Classical bits

In classical computing, a bit is the basic unit of information, which can exist in one of two distinct states: 0 or 1. Each bit holds a definite value at any given time. When multiple classical bits are used together, they can represent more complex information. For instance, a set of three bits can represent  $2^3 = 8$  distinct states, ranging from 000 to 111.

### 2.1.2 Quantum bits (Qubits)

Analogous to the role of ‘bit’ in classical computation, the basic element in quantum computation is the quantum bit (*qubit*). We start by introducing the representation of single-qubit states and then extend this to two-qubit and multi-qubit states.

**Single-qubit state.** A single-qubit state can be represented by a two-dimensional vector with unit length. Mathematically, a qubit state can be written as

$$\mathbf{a} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix} \in \mathbb{C}^2 , \quad (2.1)$$

where  $|\mathbf{a}_1|^2 + |\mathbf{a}_2|^2 = 1$  satisfies the *normalization constraint*. Following conventions in quantum theory, we use Dirac notation to represent vectors (Nielsen and Chuang, 2011), i.e.,  $\mathbf{a}$  is denoted by  $|\mathbf{a}\rangle$  (named ‘ket’) with

$$|\mathbf{a}\rangle = \mathbf{a}_1 |0\rangle + \mathbf{a}_2 |1\rangle , \quad (2.2)$$

where  $|0\rangle \equiv \mathbf{e}_0 \equiv \begin{bmatrix} 1 \\ 0 \end{bmatrix}$  and  $|1\rangle \equiv \mathbf{e}_1 \equiv \begin{bmatrix} 0 \\ 1 \end{bmatrix}$  are two computational (unit) basis states. In this representation, the coefficients  $\mathbf{a}_1$  and  $\mathbf{a}_2$  are referred to as *amplitudes*. The probabilities of obtaining the outcomes 0 or 1 upon measurement of the qubit are given by  $|\mathbf{a}_1|^2$  and  $|\mathbf{a}_2|^2$ , respectively. The normalization constraint ensures that these probabilities always sum to one, as required by the probabilistic nature of quantum mechanics. In addition, the conjugated transpose of  $\mathbf{a}$ , i.e.  $\mathbf{a}^\dagger$ , is denoted by  $\langle \mathbf{a}|$  (named ‘bra’) with

$$\langle \mathbf{a}| = \mathbf{a}_1^* \langle 0| + \mathbf{a}_2^* \langle 1| \in \mathbb{C}^2 , \quad (2.3)$$

where  $\langle 0| \equiv \mathbf{e}_0^\top \equiv [1, 0]$ ,  $\langle 1| \equiv \mathbf{e}_1^\top \equiv [0, 1]$ , and the symbol ‘ $\top$ ’ denotes the transpose operation.

The physical interpretation of coefficients  $\{\mathbf{a}_i\}$  is *probability amplitudes*. Namely, when we intend to extract information from the qubit state  $|\mathbf{a}\rangle$  into the classical form, quantum measurements are applied to this state, where

the probability of sampling the basis  $|0\rangle$  ( $|1\rangle$ ) is  $|\mathbf{a}_1|^2$  ( $|\mathbf{a}_2|^2$ ). Recall that the classical bit only permits the deterministic status with ‘0’ or ‘1’, while the qubit state in Eqn. (2.2) is the *superposition* of the two status ‘ $|0\rangle$ ’ and ‘ $|1\rangle$ ’.

**Remark**

The *quantum superposition* leads to a distinct power between quantum and classical computation, where the former can accomplish certain tasks with provable advantages.

**Two-qubit state.** The two qubits obey the tensor product rule, i.e.,

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \\ \mathbf{x}_2 \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \mathbf{y}_1 \\ \mathbf{x}_1 \mathbf{y}_2 \\ \mathbf{x}_2 \mathbf{y}_1 \\ \mathbf{x}_2 \mathbf{y}_2 \end{bmatrix}, \quad (2.4)$$

which differs from the classical bits yielding the Cartesian product rule.

For instance, let the first qubit follow Eqn. (2.2) and the second qubit state be  $|\mathbf{b}\rangle = \mathbf{b}_1|0\rangle + \mathbf{b}_2|1\rangle$  with  $|\mathbf{b}_1|^2 + |\mathbf{b}_2|^2 = 1$ . The two-qubit state formed by  $|\mathbf{a}\rangle$  and  $|\mathbf{b}\rangle$  is defined as

$$|\mathbf{a}\rangle \otimes |\mathbf{b}\rangle = \mathbf{a}_1 \mathbf{b}_1 |0\rangle \otimes |0\rangle + \mathbf{a}_1 \mathbf{b}_2 |0\rangle \otimes |1\rangle + \mathbf{a}_2 \mathbf{b}_1 |1\rangle \otimes |0\rangle + \mathbf{a}_2 \mathbf{b}_2 |1\rangle \otimes |1\rangle \in \mathbb{C}^4, \quad (2.5)$$

where the computational basis follows  $|0\rangle \otimes |0\rangle \equiv \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$ ,  $|0\rangle \otimes |1\rangle \equiv \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$ ,  $|1\rangle \otimes |0\rangle \equiv \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$ ,  $|1\rangle \otimes |1\rangle \equiv \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$ , and the coefficients satisfy  $\sum_{i=1}^2 \sum_{j=1}^2 |\mathbf{a}_i \mathbf{b}_j|^2 = 1$ .

**Remark**

For ease of notations, the state  $|\mathbf{a}\rangle \otimes |\mathbf{b}\rangle$  can be simplified as  $|\mathbf{ab}\rangle$ ,  $|\mathbf{a}, \mathbf{b}\rangle$ , or  $|\mathbf{a}\rangle |\mathbf{b}\rangle$ . We will *interchangeably* use these notations throughout the tutorial.

**Example 2.1.** A typical example of a two-qubit state is the Bell state, which represents a maximally entangled quantum state of two qubits.

*There are four types of Bell states, expressed as:*

$$\begin{aligned} |\phi^+\rangle &= \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle), \\ |\phi^-\rangle &= \frac{1}{\sqrt{2}}(|00\rangle - |11\rangle), \\ |\psi^+\rangle &= \frac{1}{\sqrt{2}}(|01\rangle + |10\rangle), \\ |\psi^-\rangle &= \frac{1}{\sqrt{2}}(|01\rangle - |10\rangle). \end{aligned} \quad (2.6)$$

*Each Bell state is a superposition of two computational basis states in the four-dimensional Hilbert space.*

**Multi-qubit state.** We now generalize the above two-qubit case to the  $N$ -qubit case with  $N > 2$ . In particular, an  $N$ -qubit state  $|\psi\rangle$  is a  $2^N$ -dimensional vector with

$$|\psi\rangle = \sum_{i=1}^{2^N} \mathbf{c}_i |i\rangle \in \mathbb{C}^{2^N}, \quad (2.7)$$

where the coefficients satisfy the normalization constraint  $\sum_{i=1}^{2^N} |\mathbf{c}_i|^2 = 1$  and the symbol ‘ $i$ ’ of the computational basis  $|i\rangle$  refers to a bit-string with  $i \in \{0, 1\}^N$ . As with the single-qubit case, the physical interpretation of coefficients  $\{\mathbf{c}_i\}$  is probability amplitudes, where the probability to sample the bit-string ‘ $i$ ’ is  $|\mathbf{c}_i|^2$ . When the number of nonzero entries in  $\mathbf{c} = [\mathbf{c}_1, \dots, \mathbf{c}_i, \dots, \mathbf{c}_{2^N}]^\top$  is larger than one, which implies that different bit-strings are coexisting coherently, the state  $|\psi\rangle$  is called *in superposition*.

#### Remark

In quantum computing, a basis state  $|i\rangle$  refers to a computational basis state in the Hilbert space of a quantum system. For an  $N$ -qubit system, the computational basis states are represented as  $|i\rangle \in \{|0\dots0\rangle, |0\dots1\rangle, \dots, |1\dots1\rangle\}$ , where  $i$  is the binary representation of the state index. These states form an orthonormal basis of the  $2^N$ -dimensional Hilbert space, satisfying

$$\langle i|j\rangle = \delta_{ij}, \forall i, j \in [2^N]. \quad (2.8)$$

These basis states are fundamental for representing and analyzing

quantum states, as any arbitrary quantum state can be expressed as a linear combination of these basis states.

Moreover, the size of  $c$  exponentially scales with the number of qubits  $N$ , attributed to the tensor product rule. This exponential dependence is an indispensable factor to achieve quantum supremacy ([Arute et al., 2019](#)), since it is extremely expensive and even intractable to record all information of  $c$  by classical devices for the modest number of qubits, e.g.,  $N > 100$ .

**Entangled multi-qubit state.** A fundamental phenomenon in multi-qubit quantum systems is *entanglement*, which represents a non-classical correlation between quantum systems that cannot be explained by classical physics. As proved by [Jozsa and Linden \(2003\)](#), quantum entanglement is an indispensable component to offer an exponential speed-up over classical computation. A representative example is Shor's algorithm, which utilizes entanglement to attain an exponential speed-up over any classical factoring algorithm. In an entangled quantum state, the state of one qubit cannot be fully described independently of the other qubits, even if they are spatially separated. The formal definition of entanglement for states in Dirac notation is as follows:

**Definition 2.2.** (*Entanglement for States in Dirac Notation*) An  $N$ -qubit state  $|\psi\rangle \in \mathbb{C}^{2^N}$  is entangled if it cannot be expressed as the tensor product of states of its subsystems  $A$  and  $B$ :

$$|\psi\rangle \neq |\psi_a\rangle \otimes |\psi_b\rangle, \quad \forall |\psi_a\rangle \in \mathbb{C}^{2^{N_A}}, |\psi_b\rangle \in \mathbb{C}^{2^{N_B}}, N_A + N_B = N. \quad (2.9)$$

If the state can be expressed in this form, it is referred to as separable.

**Example 2.3.** (*GHZ state*). A typical example of an entangled  $N$ -qubit state is the Greenberger-Horne-Zeilinger (GHZ) state ([Greenberger et al., 1989](#)), which is a generalization of the two-qubit Bell state (see [Example 2.1](#)) to a maximally entangled  $N$ -qubit state. The general form of an  $N$ -qubit GHZ state is:

$$|GHZ_N\rangle = \frac{1}{\sqrt{2}}(|0\rangle^{\otimes N} + |1\rangle^{\otimes N}). \quad (2.10)$$

For  $N = 3$ , the GHZ state is:

$$|GHZ_3\rangle = \frac{1}{\sqrt{2}}(|000\rangle + |111\rangle). \quad (2.11)$$

A key property of the entangled states (e.g., Bell states and GHZ states) is that measuring one qubit determines the outcome of measuring the other qubit, reflecting their strong quantum correlation.

### 2.1.3 Density matrix

Another description of quantum states is through *density matrix* or *density operators*. The reason for establishing density operators instead of Dirac notations arises from the imperfection of physical systems. Specifically, Dirac notations introduced in Chapter 2.1.2 are used to describe ‘ideal’ quantum states (i.e., *pure states*), where the operated qubits are isolated from the environment. Alternatively, when the operated qubits interact with the environment unavoidably, the density operators are employed to describe the behavior of quantum states living in this open system. As such, density operators describe more general quantum states.

Mathematically, an  $N$ -qubit density operator, denoted by  $\rho \in \mathbb{C}^{2^N \times 2^N}$ , presents a mixture of  $m$  quantum pure states  $|\psi_i\rangle \in \mathbb{C}^{2^N}$  with probability  $p_i \in [0, 1]$  and  $\sum_{i=1}^m p_i = 1$ , i.e.,

$$\rho = \sum_{i=1}^m p_i \rho_i, \quad (2.12)$$

where  $\rho_i = |\psi_i\rangle \langle \psi_i| \in \mathbb{C}^{2^N \times 2^N}$  is the outer product of the pure state  $|\psi_i\rangle$ . The outer product of two vectors  $|u\rangle, |v\rangle \in \mathbb{C}^n$  is expressed as

$$|u\rangle \langle v| = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \begin{bmatrix} v_1^* & v_2^* & \cdots & v_n^* \end{bmatrix} = \begin{bmatrix} u_1 v_1^* & u_1 v_2^* & \cdots & u_1 v_n^* \\ u_2 v_1^* & u_2 v_2^* & \cdots & u_2 v_n^* \\ \vdots & \vdots & \ddots & \vdots \\ u_n v_1^* & u_n v_2^* & \cdots & u_n v_n^* \end{bmatrix}, \quad (2.13)$$

where  $u_i$  and  $v_i^*$  are the element of  $|u\rangle$  and the conjugate transpose  $\langle v|$ , respectively.

From the perspective of computer science, the density operator  $\rho$  is just a *positive semi-definite matrix* with trace-preserving, i.e.,  $\mathbf{0} \preceq \rho$  and  $\text{Tr}(\rho) = 1$ .

**Definition 2.4.** (*Positive semi-definite matrix*) A matrix  $A \in \mathbb{C}^{n \times n}$  is positive semi-definite (PSD) if it satisfies the following conditions:

1.  $A$  is Hermitian:  $A = A^\dagger$
2. For any nonzero vector  $|v\rangle \in \mathbb{C}^n$ ,  $\langle v|A|v\rangle \geq 0$ , where  $\langle v|A|v\rangle$  represents the quadratic form of  $A$  with respect to  $|v\rangle$ .

When  $m = 1$ , the density operator  $\rho$  amounts to a pure state with  $\rho = |\psi_1\rangle\langle\psi_1|$ . When  $m > 1$ , the density operator  $\rho$  describes a ‘mixed’ quantum state, where the rank of  $\rho$  is larger than 1. A simple criterion to discriminate the pure states with the mixed states is as follows: the pure state with  $m = 1$  yields  $\text{Tr}(\rho^n) = \text{Tr}(\rho) = 1$  for any  $n \in \mathbb{N}_+$ ; the mixed state with  $m > 1$  satisfies  $\text{Tr}(\rho^n) < \text{Tr}(\rho) = 1$  for any  $n \in \mathbb{N}_+ \setminus \{1\}$ . Similar to the Definition 2.2 for entanglement of pure states, we can define the entanglement of mixed states.

**Definition 2.5.** (*Entanglement for Mixed States*) Let  $\rho$  be a density operator acting on a composite Hilbert space  $\mathcal{H}_A \otimes \mathcal{H}_B$ . The state  $\rho$  is said to be entangled if it cannot be expressed as:

$$\rho = \sum_i p_i \rho_A^{(i)} \otimes \rho_B^{(i)}, \quad (2.14)$$

where  $p_i \geq 0$ ,  $\sum_i p_i = 1$ , and  $\rho_A^{(i)}$  and  $\rho_B^{(i)}$  are density operators on  $\mathcal{H}_A$  and  $\mathcal{H}_B$ , respectively. If  $\rho$  can be written in this form, it is called separable.

**Example 2.6.** (*Density matrix representations*).

(i). Consider the single-qubit pure state  $|\psi\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$ . The corresponding density operator is:

$$\rho = |\psi\rangle\langle\psi| = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Here,  $\text{Tr}(\rho^2) = \text{Tr}(\rho) = 1$ , confirming that it is a pure state.

(ii). Consider the classical probabilistic mixture of  $|0\rangle$  and  $|1\rangle$ , each with equal probability  $p = 0.5$ . The density operator is:

$$\rho = 0.5 |0\rangle\langle 0| + 0.5 |1\rangle\langle 1| = \frac{1}{2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

*In this case,  $\text{Tr}(\rho^2) = 0.5 < \text{Tr}(\rho) = 1$ , indicating it is a mixed state.*

## 2.2 From Digital Logical Circuit to Quantum Circuit Model

To process quantum states, we need to introduce quantum computation, a fundamental model of which is the quantum circuit model. In this section, we will begin with classical computation in Chapter 2.2.1 and transit to details about the quantum circuit model in Chapter 2.2.2, including quantum gates, quantum channel, and quantum measurements.

### 2.2.1 Classical digital logical circuit

Digital logic circuits are the foundational building blocks of classical computing systems. They process classical bits by performing logical operations through logic gates. In this subsection, we introduce the essential components of digital logic circuits and their functionality, followed by a discussion of how these classical circuits relate to quantum circuits.

#### Logic gates

Logic gates are the basic components of a digital circuit. They take binary inputs, represented as 0 or 1, and produce a binary output based on a predefined logical operation. The most common logic gates include:

1. **NOT Gate:** This gate inverts the input bit, i.e., it produces 1 if the input is 0, and vice versa. Its truth table is shown in Table 2.1;

Table 2.1: **Input-output mapping of the NOT gate.**

Input (A)		Output (NOT A)
0		1
1		0

2. **AND Gate:** Produces an output of 1 only if both input bits are 1; otherwise, it outputs 0. The truth table is shown in Table 2.2;
3. **OR Gate:** Outputs 1 if at least one input is 1. The truth table is shown in Table 2.3;

Table 2.2: **Input-output mapping of the AND gate.**

Input (A)	Input (B)	Output (A AND B)
0	0	0
0	1	0
1	0	0
1	1	1

Table 2.3: **Input-output mapping of the OR gate.**

Input (A)	Input (B)	Output (A OR B)
0	0	0
0	1	1
1	0	1
1	1	1

Table 2.4: **Input-output mapping of the XOR gate.**

Input (A)	Input (B)	Output (A XOR B)
0	0	0
0	1	1
1	0	1
1	1	0

4. **XOR Gate:** Produces an output of 1 if the inputs are different, and 0 otherwise. The truth table is shown in Table 2.4.

These logic gates can be combined in various configurations to build more complex circuits capable of performing arbitrary arithmetic operations.

### Circuit design and universality

A classical digital logic circuit is composed of interconnected gates designed to perform specific tasks, such as addition or multiplication. A key property of these circuits is **universality**, meaning any logical function can be implemented using a finite set of gates. For example, the **NAND Gate** (NOT AND) and **NOR Gate** (NOT OR) are universal gates. Any other logical operation can be constructed using only NAND or NOR gates (Leach and Malvino, 1994).

### 2.2.2 Quantum circuit

Classical digital logical circuits provide the essential framework for understanding computation. While classical circuits operate on bits and perform deterministic operations, quantum circuits manipulate qubits and involve probabilistic behavior. The concepts of logic gates, circuit design, and universality lay the groundwork for transitioning to quantum circuits introduced in this subsection.

#### Quantum gate

Recall that the computational toolkit for classical computers is logic gates, e.g., NOT, AND, OR, and XOR, which are applied to the single bit or multiple bits to accomplish computation. Similarly, the computational toolkit for quantum computers (or quantum circuits) is **quantum gate**, which *operates on qubits* introduced in Chapter 2.1.2 to complete the computation. In the following, we will introduce both single-qubit and multi-qubit gates.

**Single-qubit gates.** Single-qubit gates control the evolution of the single-qubit state  $|a\rangle$ . Due to the law of quantum mechanics, the evolved state should satisfy the normalization constraint. The implication of this constraint is that the evolution must be a unitary operation. Concretely, denoted  $U \in \mathbb{C}^{2 \times 2}$  as a linear operator and the evolved state as

$$|\hat{a}\rangle := U|a\rangle = \hat{a}_1|0\rangle + \hat{a}_2|1\rangle \in \mathbb{C}^2, \quad (2.15)$$

the summation of coefficients  $|\hat{a}_1|^2 + |\hat{a}_2|^2 = \langle \hat{a} | \hat{a} \rangle = \langle a | U^\dagger U | a \rangle$  is equal to 1 if and only if  $U$  is unitary with  $U^\dagger U = UU^\dagger = \mathbb{I}_2$ . The symbol ‘ $\dagger$ ’ denotes the conjugate transpose operation. Under the density operator representation, the evolution of  $|a\rangle$  yields

$$\hat{\rho} = U\rho U^\dagger, \quad (2.16)$$

where  $\hat{\rho} = |\hat{a}\rangle\langle\hat{a}|$  and  $\rho = |a\rangle\langle a|$ .

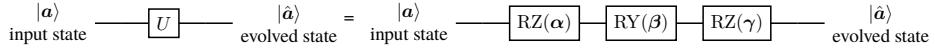
Several common single-qubit gates, including Pauli-X, Pauli-Y, Pauli-Z, Hadamard, and rotational single-qubit gates about the X, Y, and Z axes (RX, RY, RZ), are illustrated in Figure 2.1. According to Theorem 4.1 in ([Nielsen and Chuang, 2011](#)), any unitary operation on a single qubit can be decomposed into a sequence of rotations as:

$$U = \text{RZ}(\alpha)\text{RY}(\beta)\text{RZ}(\gamma), \quad (2.17)$$

where  $\alpha, \beta, \gamma \in [0, 2\pi]$ , up to a global phase shift.

Quantum gate	Matrix form	Circuit representation
Pauli-X (X)	$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$	
Pauli-Y (Y)	$\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$	
Pauli-Z (Z)	$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$	
Hadamard (H)	$\frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$	
Controlled-Z (CZ)	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$	
RX( $\theta$ )	$\begin{pmatrix} \cos(\frac{\theta}{2}) & -i \sin(\frac{\theta}{2}) \\ -i \sin(\frac{\theta}{2}) & \cos(\frac{\theta}{2}) \end{pmatrix}$	
RY( $\theta$ )	$\begin{pmatrix} \cos(\frac{\theta}{2}) & -\sin(\frac{\theta}{2}) \\ \sin(\frac{\theta}{2}) & \cos(\frac{\theta}{2}) \end{pmatrix}$	
RZ( $\theta$ )	$\begin{pmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{i\theta/2} \end{pmatrix}$	
SWAP	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	
Controlled-NOT (CNOT)	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$	

Figure 2.1: **The summarization of quantum gates.** The table contains the abbreviation, the mathematical form, and the graph representation of a set of universal quantum gates.  $i$  represents the imaginary unit.



**Figure 2.2: The evolution of the single-qubit state decomposed into the quantum gates.**

The evolution from  $|a\rangle$  to  $|\hat{a}\rangle$  can be visualized using a quantum circuit diagram, as illustrated in Figure 2.2. The wire in the circuit represents a qubit, which evolves from the initial state  $|a\rangle$  on the left to the final state  $|\hat{a}\rangle$  on the right. Gates are applied sequentially from left to right along the wire.

#### Remark

The circuit model serves as a foundational framework for describing quantum computation due to its *intuitive* and *modular* nature, making it accessible for researchers and practitioners transitioning from classical to quantum computing. First, the circuit model provides a standardized graphical language to represent complex quantum algorithms, enabling clear visualization of the computational flow and interactions among qubits. Second, the modularity of the circuit model allows quantum operations to be easily decomposed into a pre-defined gate set, ensuring compatibility across different quantum hardware architectures.

**Multi-qubit gates.** The evolution of the  $N$ -qubit quantum state can be effectively generalized by the single-qubit case. That is, the unitary operator  $U \in \mathbb{C}^{2^N \times 2^N}$  evolves an  $N$ -qubit state  $|\psi\rangle$  in Eqn. (2.7) as

$$|\hat{\psi}\rangle = U |\psi\rangle \in \mathbb{C}^{2^N}. \quad (2.18)$$

The evolution of  $|\psi\rangle$  under the density operator representation is denoted by  $\hat{\rho} = U \rho U^\dagger$ , where  $\hat{\rho} = |\hat{\psi}\rangle \langle \hat{\psi}|$  and  $\rho = |\psi\rangle \langle \psi|$ .

#### Remark

In the view of computer science, the quantum (logic) gates in Figure 2.1 are well-designed matrices with the following properties. First, all quantum gates are unitary (e.g.,  $XX^\dagger = I_2$ ). Second, X, Y, Z, H gates have the fixed form with size  $2 \times 2$ ; CNOT, CZ, and SWAP gates have the fixed form with size  $4 \times 4$ . Third,  $RX(\theta)$ ,  $RY(\theta)$ ,  $RZ(\theta)$  gates are matrices controlled by a single variable  $\theta$ .

Figure 2.1 includes two significant multi-qubit gates: the controlled-Z (CZ) gate and the controlled-NOT (CNOT) gate. For instance, the CNOT gate operates on two qubits: a *control* qubit (top line) and a *target* qubit (bottom line). If the control qubit is 0, the target qubit remains unchanged; if the control qubit is 1, the target qubit is flipped.

**Example 2.7.** (*State evolved by multi-qubit gates*). Figure 2.3 illustrates the evolution of a 3-qubit state  $|\psi\rangle$  under a multi-qubit circuit consisting of multi-qubit gates. Each wire represents a qubit, and the evolution occurs from left to right. Starting with the initial state  $|\psi\rangle = |000\rangle$ , a Hadamard gate is applied to the first qubit, followed by two CNOT gates: one acting on the first and second qubits, and the other acting on the second and third qubits. The final evolved state, shown on the right, is the GHZ state introduced in Example 2.3, i.e.,  $|\hat{\psi}\rangle = U|\psi\rangle = \frac{1}{\sqrt{2}}(|000\rangle + |111\rangle)$ . The entire unitary operation can be represented as:

$$U = (H \otimes \mathbb{I}_4)(CNOT \otimes \mathbb{I}_2)(\mathbb{I}_2 \otimes CNOT). \quad (2.19)$$

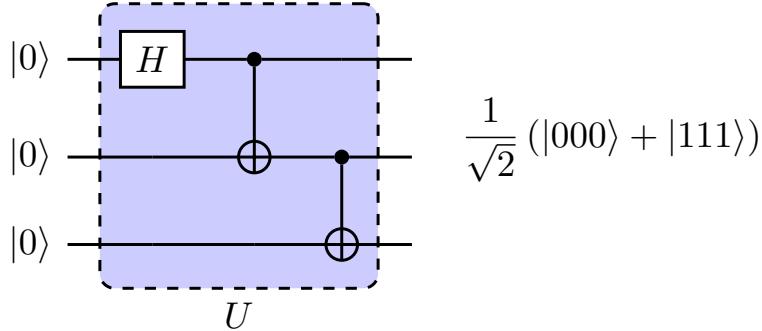


Figure 2.3: The decomposition of the multi-qubit circuit  $U$  in the case of  $N = 3$ .

#### Remark

The CNOT gate plays a pivotal role in quantum computing due to its unique ability to generate entangled states, such as the Bell states and GHZ states presented in Examples 2.1&2.3. Besides, the CNOT gate is one of the most commonly implemented gates on quantum

hardware. Its design and optimization directly impact the fidelity and scalability of quantum systems.

**A universal quantum gate set.** While many single and multi-qubit gates exist, it is sufficient to use a *universal set of gates* to construct any unitary operation. As proved in Chapter 4.5.2 of Ref. (Nielsen and Chuang, 2011), any unitary operator  $U$  in Eqn. (2.18) can be decomposed into the single-qubit and two-qubit gates with a certain arrangement.

**Fact 2.8** (Solovay-Kitaev theorem, (Dawson and Nielsen, 2005)). *Suppose we are given a fixed universal gate set  $\mathcal{G}$ , which generates a dense group  $SU(d)$ . Then any unitary operator  $U \in SU(d)$  can be approximated to an arbitrary precision  $\epsilon > 0$  by a finite sequence of gates from  $\mathcal{G}$ . Formally, there exists a decomposition such that*

$$\left\| U - \prod_{l=1}^L G_l \right\|_{op} \leq \epsilon, \quad G_l \in \mathcal{G}, \quad L \in \mathbb{N}, \quad (2.20)$$

where  $\|\cdot\|_{op}$  is the operator norm which is the largest singular value of a matrix, and  $L$  is the required number of gates that scales as:

$$L = O(\log^c(1/\epsilon)), \quad (2.21)$$

with  $c \approx 4$ .

A commonly used universal gate set includes single-qubit rotations  $RX(\theta)$ ,  $RY(\theta)$ ,  $RZ(\theta)$ , and two-qubit gates such as the CNOT gate. As illustrated in Figure 1.1, any ideal quantum computation can be represented by a unitary operator. This universal gate set provides a practical and foundational toolkit for implementing arbitrary quantum algorithms.

### Quantum channels

Analogous to the unitary operation describing the evolution of quantum states in the closed system, the quantum channel formalizes the evolution of quantum states in the open system. Refer to the textbook (Wilde, 2011) for more details.

Mathematically, every quantum channel  $\mathcal{N}(\cdot)$  can be treated as a linear, completely positive, and trace-preserving map (CPTP map).

**Definition 2.9** (CPTP map). *Denote  $\mathcal{L}(\mathcal{H})$  as the space of square linear operators acting on the Hilbert space  $\mathcal{H}$ . We say  $\mathcal{N}(\cdot)$  is a CPTP map if the following conditions are satisfied:*

- The ‘linearity’ requires for any  $X_A, Y_A \in \mathcal{L}(\mathcal{H}_A)$  and  $a, b \in \mathbb{C}$ ,  $\mathcal{N}(aX_A + bY_A) = a\mathcal{N}(X_A) + b\mathcal{N}(Y_A)$ .
- The definition of completely positive is as follows. A linear map  $\mathcal{N} : \mathcal{L}(\mathcal{H}_A) \rightarrow \mathcal{L}(\mathcal{H}_B)$  is a positive map if  $\mathcal{N}(X_A)$  is positive semi-definite for all positive semi-definite operators  $X_A \in \mathcal{L}(\mathcal{H}_A)$ . Moreover, a linear map  $\mathcal{N} : \mathcal{L}(\mathcal{H}_A) \rightarrow \mathcal{L}(\mathcal{H}_B)$  is completely positive if  $\mathbb{I}_R \otimes \mathcal{N}$  is a positive map for any size of  $R$ .
- The trace preservation requires  $\text{Tr}(\mathcal{N}(X_A)) = \text{Tr}(X_A)$  for any  $X_A \in \mathcal{L}(\mathcal{H}_A)$ .

A quantum channel can be represented by the Choi-Kraus decomposition (Nielsen and Chuang, 2011). Mathematically, let  $\mathcal{L}(\mathcal{H}_A, \mathcal{H}_B)$  denote the space of linear operators taking  $\mathcal{H}_A$  to  $\mathcal{H}_B$ . The Choi–Kraus decomposition of the quantum channel  $\mathcal{N}(\cdot) : \mathcal{L}(\mathcal{H}_A) \rightarrow \mathcal{L}(\mathcal{H}_B)$  is

$$\mathcal{N}(X_A) = \sum_{a=1}^d \mathbf{M}_a X_A \mathbf{M}_a^\dagger \quad (2.22)$$

where  $X_A \in \mathcal{L}(\mathcal{H}_A)$ ,  $M_a \in \mathcal{L}(\mathcal{H}_A, \mathcal{H}_B)$ ,  $\sum_{a=1}^d \mathbf{M}_a^\dagger \mathbf{M}_a = \mathbb{I}_{\dim(\mathcal{H}_A)}$ , and  $d \leq \dim(\mathcal{H}_A)\dim(\mathcal{H}_B)$ . Here  $\dim(\mathcal{H}_*)$  refers to the dimension of the space  $\mathcal{H}_*$ .

We next introduce two common types of quantum channels, which will be broadly employed in the following context to simulate the noise of quantum devices.

The first type is the *depolarizing channel*, which considers the scenario such that the information of the input state can be entirely lost with some probability.

**Definition 2.10** (Depolarization channel). *Given an  $N$ -qubit quantum state  $\rho \in \mathbb{C}^{2^N \times 2^N}$ , the depolarization channel  $\mathcal{N}_p$  acts on a  $2^N$ -dimensional Hilbert space follows*

$$\mathcal{N}_p(\rho) = (1 - p)\rho + p \frac{\mathbb{I}_{2^N}}{2^N}, \quad (2.23)$$

where  $\mathbb{I}_{2^N}/2^N$  refers to the maximally mixed state and  $p$  is a scalar representing the depolarization rate.

**Example 2.11.** (Single-qubit state with depolarization channel). Con-

sider a single-qubit pure state  $\rho = |0\rangle\langle 0|$  with the density matrix

$$\rho = |0\rangle\langle 0| = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. \quad (2.24)$$

When the depolarizing channel  $\mathcal{N}_p$  acts on this state, the output is given by

$$\mathcal{N}_p(\rho) = (1-p)\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \frac{p}{2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 - \frac{p}{2} & 0 \\ 0 & \frac{p}{2} \end{bmatrix}. \quad (2.25)$$

Therefore, the purity is inferred as:

$$\text{Tr}(\mathcal{N}_p^2(\rho)) = 1 - p + \frac{p^2}{2}. \quad (2.26)$$

When  $p = 0$ , the state remains pure and unchanged. When  $0 < p \leq 1$ , the state becomes a mixture of states  $|0\rangle$  and  $|1\rangle$  with  $\text{Tr}(\mathcal{N}_p^2(\rho)) < 1$ . When  $p = 1$ , the state evolves into the maximally mixed state.

The second type is the *Pauli channel*, which serves as a dominant noise source in many computing architectures and as a practical model for analyzing error correction ([Flammia and Wallman, 2020](#)).

**Definition 2.12** (Single-qubit Pauli channel). *Given a quantum state  $\rho \in \mathbb{C}^{2 \times 2}$ , the single-qubit Pauli channel  $\mathcal{N}_{\vec{p}}$  acts on this state follows*

$$\mathcal{N}_{\vec{p}}(\rho) = p_I \rho + p_X X \rho X + p_Y Y \rho Y + p_Z Z \rho Z, \quad (2.27)$$

where  $\vec{p} = (p_I, p_X, p_Y, p_Z)$  and  $p_I + p_X + p_Y + p_Z = 1$ .

Note that for a single-qubit system, the depolarization channel  $\mathcal{N}_p$  is a special Pauli channel with setting  $p_X = p_Y = p_Z = p$ .

**Example 2.13.** (Single-qubit state with Pauli channel). Consider a single-qubit pure state  $\rho = |0\rangle\langle 0|$  with the density matrix:

$$\rho = |0\rangle\langle 0| = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. \quad (2.28)$$

When the Pauli channel  $\mathcal{N}_{\vec{p}}$  acts on this state, the output is given by

$$\mathcal{N}_{\vec{p}}(\rho) = p_I \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + p_X \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} + p_Y \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} + p_Z \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad (2.29)$$

$$= \begin{bmatrix} p_I + p_Z & 0 \\ 0 & p_X + p_Y \end{bmatrix}. \quad (2.30)$$

Let us analyze three special cases for the probability vector  $\vec{p} = (p_I, p_X, p_Y, p_Z)$ :

- **Case 1:** If  $p_X = p_Y = p_Z = p$ , the Pauli channel reduces to the depolarization channel and the prepared state becomes:

$$\mathcal{N}_{\vec{p}}(\rho) = \begin{bmatrix} 1 - 2p & 0 \\ 0 & 2p \end{bmatrix}. \quad (2.31)$$

- **Case 2:** If  $p_Y = p_Z = 0$ , the prepared state becomes:

$$\mathcal{N}_{\vec{p}}(\rho) = \begin{bmatrix} 1 - p_X & 0 \\ 0 & p_X \end{bmatrix}. \quad (2.32)$$

In this scenario, the Pauli channel reduces to the bit-flip channel.

- **Case 3:** For other values of  $\vec{p}$ , the effect of the Pauli channel on the pure state  $|0\rangle$  can be interpreted as a combination of the depolarizing channel and the bit-flip channel.

To generalize the single-qubit Pauli channel to a multi-qubit Pauli channel, we extend the definition to account for the action of Pauli operators on multiple qubits.

**Definition 2.14** (Multi-qubit Pauli channel). *Given a quantum state  $\rho \in \mathbb{C}^{2^N \times 2^N}$  for an  $N$ -qubit system, the multi-qubit Pauli channel  $\mathcal{N}_{\vec{p}}$  acts as*

$$\mathcal{N}_{\vec{p}}(\rho) = \sum_{P \in \mathcal{P}_N} p_P P \rho P^\dagger, \quad (2.33)$$

where  $\mathcal{P}_N = \{I, X, Y, Z\}^{\otimes N}$  denotes the set of all tensor products of the  $N$  single-qubit Pauli operators, and  $p_P$  is the probability of applying the Pauli operator  $P$  with  $\sum_{P \in \mathcal{P}_N} p_P = 1$ .

## Remark

The multi-qubit Pauli channel considers the existence of correlated Pauli noise on different qubits. If each qubit only experiences independent single-qubit Pauli noise, the multi-qubit channel can be written as the tensor product of sing-qubit Pauli channels:

$$\mathcal{N}_{\vec{p}}(\rho) = \otimes_{i=1}^N \mathcal{N}_{\vec{p}_i}(\rho), \quad (2.34)$$

where  $\mathcal{N}_{\vec{p}_i}$  is the single-qubit Pauli channel acting on the  $i$ -th qubit with probabilities  $\vec{p}_i = (p_I, p_X, p_Y, p_Z)$ .

Having acknowledged the motivation and definition of quantum channels, it is natural to ask *what is the relation between quantum channels and quantum gates?* A straightforward observation is that a quantum gate is a special case of a quantum channel. Conversely, the evolution of the quantum state can be built from the unitary operation via the isometric extension (Wilde, 2011). The following theorem shows that any quantum channel arises from a unitary evolution on a larger Hilbert space.

## Remark

According to the Choi-Kraus decomposition, the unitary operator is a special case of a quantum channel. Specifically, when  $d = 1$ , the quantum channel reduces to:

$$\mathcal{N}(X_A) = \mathbf{M}_1 X_A \mathbf{M}_1^\dagger, \quad (2.35)$$

where  $\mathbf{M}_1$  is a unitary operator satisfying  $\mathbf{M}_1^\dagger \mathbf{M}_1 = \mathbb{I}$ . This highlights that all unitary operators are quantum channels, but not all quantum channels are unitary.

**Theorem 2.15.** (Wilde, 2011) Let  $\mathcal{N}(\cdot) : \mathcal{L}(\mathcal{H}_A) \rightarrow \mathcal{L}(\mathcal{H}_B)$  be a quantum channel defined in Eqn. (2.22). Let  $\mathcal{H}_E$  be the Hilbert space of an auxiliary system. Denote the input state as  $\rho$  (i.e., a density operator  $\rho \in \mathbb{C}^{\dim(\mathcal{H}_A) \times \dim(\mathcal{H}_A)}$ ). Then there exists a unitary  $U : \mathcal{L}(\mathcal{H}_A \otimes \mathcal{H}_E) \rightarrow \mathcal{L}(\mathcal{H}_B \otimes \mathcal{H}_E)$  and a normalized vector (i.e., a pure state)  $|\varphi\rangle \in \mathbb{C}^{\dim(\mathcal{H}_E)}$  such that

$$\mathcal{N}(\rho) = \text{Tr}_E \left( U(\rho \otimes |\varphi\rangle\langle\varphi|) U^\dagger \right), \quad (2.36)$$

where  $\text{Tr}_E(\cdot)$  denotes the partial trace over the ancillary Hilbert space  $\mathcal{H}_E$ , and the dimension of  $\mathcal{H}_E$  depends on the rank of the Kraus representation

of  $\mathcal{N}$ .

*Proof sketch of Theorem 2.15.* We extend the system to include an ancillary Hilbert space  $\mathcal{H}_E$ , representing the environment. The combined space  $\mathcal{H}_A \otimes \mathcal{H}_E$  forms a closed physical system, where the evolution of the quantum state can be described by a unitary operator  $U$  acting on  $\mathcal{H}_B \otimes \mathcal{H}_E$ .

To find a feasible unitary  $U$ , we express the quantum channel  $\mathcal{N}$  using its isometric extension (Wilde, 2011), i.e.,

$$\mathcal{N}(\rho) = \text{Tr}_E(V\rho V^\dagger), \quad (2.37)$$

where  $V : \mathcal{H}_A \rightarrow \mathcal{H}_B \otimes \mathcal{H}_E$  is an isometry operator embedding the input state into the larger Hilbert space. For simplicity, assume  $\mathcal{H}_A = \mathcal{H}_B$ . The isometry operator  $V$  can always be embedded into a unitary operator  $U$  acting on  $\mathcal{H}_B \otimes \mathcal{H}_E$ , ensuring that  $U$  captures the reversible evolution of the extended system.

Next, we augment the input state  $\rho$  by introducing an ancillary state  $|\varphi\rangle \in \mathcal{H}_E$ , yielding the combined state  $\rho \otimes |\varphi\rangle\langle\varphi|$ . Substituting this augmented state and the unitary operator  $U$  into the isometric extension in Eqn. (2.37) gives Eqn. (2.36). Theorem 2.15 is thereby proven.  $\square$

The translation between the unitary operation and the quantum channels described by Theorem 2.15 can be visually explained, as shown in Figure 2.4. In this diagram, the first wire corresponds to the original input state  $\rho$ , while the second wire represents the initial state  $|\varphi\rangle$  of the environment. To determine the output of the quantum channel  $\mathcal{N}$  applied to  $\rho$ , an  $\mathcal{N}$ -induced unitary operation  $U$  is performed on the combined system, followed by a partial trace over the environment to discard its information.

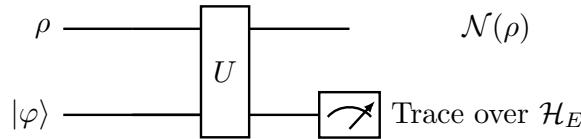


Figure 2.4: **The evolution of quantum states based on Theorem 2.15.**

### Quantum measurements

In addition to quantum gates and quantum channels that manipulate quantum states, another special operation in quantum circuits is measurement.

The aim of quantum measurements is to extract quantum information of the evolved state into the classical form. The quantum circuit diagram, which describes applying a unitary  $U$  followed by the quantum measurement to a single-qubit state  $|\alpha\rangle$ , is shown in Figure 2.5. In particular, both types of measurements are depicted by the ‘meter’ symbol.

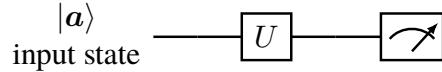


Figure 2.5: The quantum circuit diagram with measurement.

The quantum measurements can be categorized into two types, i.e., *projective measurements* and *positive operator-valued measures* (Preskill, 1999; Nielsen and Chuang, 2011).

The projective measurement, which is also called the von Neumann measurement, is formally described by the Hermitian operator  $A = \sum_i \lambda_i |v_i\rangle\langle v_i|$ , where  $\{\lambda_i\}$  and  $\{|v_i\rangle\}$  refer to the eigenvalues and eigenvectors of  $A$ , respectively. Supported by the Born rule (Nielsen and Chuang, 2011), when the measurement operator  $A \in \mathbb{C}^{2^N \times 2^N}$  is applied to an  $N$ -qubit state  $|\Phi\rangle \in \mathbb{C}^{2^N}$ , the probability of measuring any one of the eigenvalues in  $\{\lambda_i\}$  is

$$\text{Pr}(\lambda_i) = |\langle v_i | \Phi \rangle|^2. \quad (2.38)$$

In the density operator representation, suppose that the state to be measured is  $\rho \in \mathbb{C}^{2^N \times 2^N}$ , the probability of measuring any one of the eigenvalues in  $\{\lambda_i\}$  is

$$\text{Pr}(\lambda_i) = \text{Tr}(\rho |v_i\rangle\langle v_i|). \quad (2.39)$$

Define  $\Pi_i = |v_i\rangle\langle v_i|$  as the  $i$ -th projective operator. The complete set of projective operators  $\{\Pi_i\}$  has the following properties

$$1) \Pi_i \Pi_j = \delta_{ij}; \quad 2) \Pi_i^\dagger = \Pi_i; \quad 3) \Pi_i^2 = \Pi_i; \quad 4) \sum_i \Pi_i = \mathbb{I}_{2^N}. \quad (2.40)$$

A special set of projectors is defined as  $\Pi_i = |i\rangle\langle i|$  for  $\forall i \in [2^N]$ , which measures the probability corresponding to the basis state  $|i\rangle$ . For example, given the single-qubit state  $|\alpha\rangle$  in Eqn. (2.2), the probability to measure the computational basis state  $|i\rangle$  is

$$\text{Pr}(i) = |\langle v_i | \alpha \rangle|^2 = |\alpha_i|^2. \quad (2.41)$$

The second type of quantum measurement is the *positive operator-valued measures (POVM)*. A POVM is described by a collection of positive operators  $0 \preceq E_i$  satisfying  $\sum_i E_i = \mathbb{I}$ . Each positive operator  $E_i$  is associated with an outcome of measurement. Specifically, applying the measurement  $\{E_m\}$  to the state  $|\psi\rangle$ , the probability of outcome  $i$  is given by

$$\Pr(i) = |\langle\psi|E_i|\psi\rangle|^2. \quad (2.42)$$

In the density operator representation, suppose that the state to be measured is  $\rho \in \mathbb{C}^{2^N \times 2^N}$ , the probability of outcome  $i$  is given by

$$\Pr(\lambda_i) = \text{Tr}(\rho E_i). \quad (2.43)$$

We remark that the main difference between projective measurements and POVM elements is that the POVM elements do not have to be orthogonal. Due to this reason, the projective measurement is a special case of the generalized measurement (i.e., with setting  $E_i = \Pi_i^\dagger \Pi_i$ ).

#### Remark

Here we address the accessible information through quantum measurements in the ideal and practical settings. For ease of discussion, suppose that the computation result corresponds to the probability amplitude  $\mathbf{a}_1$  in the single-qubit state  $|\mathbf{a}\rangle = \mathbf{a}_1|0\rangle + \mathbf{a}_2|1\rangle$  in Eqn. (2.2). To extract  $\mathbf{a}_1$  from the quantum state into the classical form, we apply the projective operator  $\Pi_1 = |0\rangle\langle 0|$  to this state. Due to the law of quantum mechanics, after each measurement, the state is collapsed and the measured outcome  $V_i$  can be viewed as a binary random variable with the Bernoulli distribution  $\text{Ber}(\mathbf{a}_i)$ , i.e.,  $\Pr(V_i = 1) = \mathbf{a}_1$  and  $\Pr(V_i = 0) = 1 - \mathbf{a}_1$ . Through applying the measurement  $\Pi_i$  to  $K$  copies of the state  $|\mathbf{a}\rangle$ , the obtained statistics, i.e., the sample mean, is denoted by  $\bar{\mathbf{a}}_1 = \sum_{i=1}^K V_i/K$ . The large number theorem indicates  $\bar{\mathbf{a}}_1 = \mathbf{a}_1$  when  $K \rightarrow \infty$ . However, only the finite number of measurements  $K$  is allowed in practice and thus results in an estimation error.

## 2.3 Quantum Read-in and Read-out protocols

The terms *quantum read-in* and *read-out* refer to the processes of transferring information between classical systems and quantum systems. These are

fundamental steps in the workflow of quantum machine learning shown in Figure 1.1, responsible for loading data and extracting results.

Quantum read-in and read-out pose significant bottlenecks in leveraging quantum computing to address classical computational tasks. As emphasized in (Aaronson, 2015), while quantum algorithms can offer exponential speed-ups in specific problem domains, these advantages can be negated if the processes of loading classical data into quantum systems (read-in) or extracting results from quantum systems (read-out) are inefficient. Specifically, the high-dimensional nature of quantum states and the constraints on measurement precision often lead to overheads that scale poorly with problem size. These challenges underscore the importance of optimizing quantum read-in and read-out protocols to realize the full potential of quantum computing. Below is a detailed introduction to quantum read-int and read-out protocols, including the basic concept and several typical algorithms.

### 2.3.1 Quantum read-in protocols

Quantum read-in refers to the process of encoding classical information into quantum systems that can be manipulated by a quantum computer, which can be regarded as the *classical-to-quantum mapping*. It acts as a bridge to utilize quantum algorithms to solve classical problems in quantum computing. Here, we will introduce several typical encoding methods, including basis encoding, amplitude encoding, angle encoding, and quantum random access memory. Some easy-to-use demonstrations are provided in Chapter 2.5.

#### Basis encoding

Basis encoding is a basic method for processing classical data that can be represented in binary form. Given a classical binary vector  $\mathbf{x} = (\mathbf{x}_0, \dots, \mathbf{x}_i, \dots, \mathbf{x}_{N-1}) \in \{0, 1\}^N$ , this encoding technique maps the vector directly into a quantum computational basis state as follows:

$$|\psi\rangle = |\mathbf{x}_0, \dots, \mathbf{x}_{N-1}\rangle. \quad (2.44)$$

In this process,  $N$  qubits are required to represent a binary vector of length  $N$ . To prepare the corresponding quantum state  $|\psi\rangle$ , an  $X$  gate is applied to each qubit where the corresponding bit value is 1. The overall quantum state preparation can be expressed as:

$$|\psi\rangle = \bigotimes_{i=0}^{N-1} X^{\mathbf{x}_i} |0\rangle^{\otimes N},$$

where  $|0\rangle^{\otimes N}$  represents an initial state of all qubits set to  $|0\rangle$ , and  $X^{\mathbf{x}_i}$  means applying the  $X$  gate to the  $i$ -th qubit only if  $\mathbf{x}_i = 1$ .

**Example 2.16.** (*Basis encoding*). Consider encoding the integer 6, which has the binary representation  $\mathbf{x} = (1, 1, 0)$ . The corresponding quantum state is  $|110\rangle$ . This state can be implemented by applying  $X$  gates to the first and second qubits, as shown in Figure 2.6.

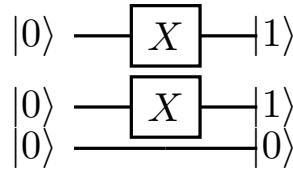


Figure 2.6: Example of basis encoding for the integer 6.

### Amplitude encoding

Amplitude encoding is a technique that maps classical data into the amplitudes of a quantum state. Given a vector  $\mathbf{x} = (\mathbf{x}_0, \dots, \mathbf{x}_i, \dots, \mathbf{x}_{2^N-1}) \in \mathbb{C}^{2^N}$  containing complex values, we first apply  $L_2$  normalization to obtain a normalized vector

$$\hat{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}, \quad (2.45)$$

where  $\|\mathbf{x}\|_2$  is the Euclidean norm. This ensures that the normalized vector  $\hat{\mathbf{x}}$  satisfies  $\sum_{i=0}^{2^N-1} |\hat{\mathbf{x}}_i|^2 = 1$ . The corresponding quantum state is then expressed as

$$|\psi\rangle = \sum_{i=0}^{2^N-1} \hat{\mathbf{x}}_i |i\rangle \quad (2.46)$$

with  $|i\rangle$  representing the  $N$ -qubit computational basis states.

**Example 2.17.** (*Amplitude encoding*). Consider encoding a normalized vector  $\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_1) \in \mathbb{C}^2$  into the quantum state  $|\psi\rangle = \mathbf{x}_0 |0\rangle + \mathbf{x}_1 |1\rangle$ . This can be achieved by applying a rotation gate  $U = R_Y(\theta)$  to the initial state  $|0\rangle$ , where  $\theta = 2 \arccos(\mathbf{x}_0)$ .

Amplitude encoding is highly efficient because it allows an exponentially large vector of length  $2^N$  to be represented using only  $N$  qubits. However, preparing this quantum state requires constructing a unitary transformation  $U$  such that  $|\psi\rangle = U|0\rangle^{\otimes N}$ . Efficiently finding such transformations can be challenging and is an active area of research (see Chapter 2.6 for the discussions).

### Angle encoding

Basis encoding and amplitude encoding are fundamental techniques for mapping classical data to quantum states, but each comes with distinct resource costs. Basis encoding requires a number of qubits equal to the dimensionality of the binary representation of classical data and necessitates minimal gate operations for state preparation. In contrast, amplitude encoding is highly compact in terms of qubits, using only the logarithmic of the data dimensionality, but it involves a significant gate complexity.

To address this limitation, an alternative is *angle encoding*. The core idea of angle encoding is to embed classical data into a quantum state through rotation angles.

Given a real-valued vector  $\mathbf{x} = (\mathbf{x}_0, \dots, \mathbf{x}_i, \dots, \mathbf{x}_{N-1}) \in \mathbb{R}^N$ , the encoded quantum state can be represented as:

$$|\psi\rangle = \bigotimes_{i=0}^{N-1} R_\sigma(\mathbf{x}_i) |0\rangle^{\otimes N} = \bigotimes_{i=0}^{N-1} \exp\left(-i\frac{\mathbf{x}_i}{2}\sigma\right) |0\rangle^{\otimes N}, \quad (2.47)$$

where  $\sigma \in \{X, Y, Z\}$  denotes a Pauli operator, as defined in Figure 2.1. Since Pauli rotation gates are  $2\pi$ -periodic, it is essential to scale each element  $\mathbf{x}_i$  into the range  $[0, \pi)$  to ensure that different values are encoded into distinct quantum states.

A key advantage of angle encoding is its ability to introduce nonlinearity. By mapping classical data into the parameters of quantum rotation gates, angle encoding leverages trigonometric functions to naturally capture nonlinear relationships. This property is particularly important in quantum machine learning, where nonlinearity is essential for models to learn complex patterns in data, such as non-linearly separable decision boundaries.

### Quantum Random Access Memory (QRAM)

Basis encoding, amplitude encoding, and angle encoding are generally designed to encode a single item of data at one time, which makes it challenging to process complicated classical datasets. The QRAM ([Giovannetti et al.](#),

2008), analogous to classical RAM, aims to simultaneously store, address, and access multiple quantum states.

QRAM consists of two types of qubits: data qubits for storing classical data and address qubits for addressing. Given a classical dataset  $\mathcal{D} = \{\mathbf{x}^{(j)}\}_{j=0}^{M-1}$  with  $M$  training examples, assume we separately encode each data item into a quantum state  $|\mathbf{x}^{(j)}\rangle_d$  using one of the encoding methods above. The QRAM can be constructed as follows: (1) Prepare an  $N_a$ -qubit address register where  $N_a = \lceil \log_2(M) \rceil$ ; (2) Associate each data state  $|\mathbf{x}^{(j)}\rangle_d$  with corresponding address state  $|j\rangle_a$ . The whole dataset is therefore encoded into a quantum state of the form

$$|\mathcal{D}\rangle = \sum_{j=0}^{M-1} \frac{1}{\sqrt{M}} |j\rangle_a |\mathbf{x}^{(j)}\rangle_d. \quad (2.48)$$

#### Remark

The subscript  $d$  in  $|\mathbf{x}^{(j)}\rangle_d$  indicates that this quantum state resides in the data register, differentiating it from address qubits, which are denoted with the subscript  $a$  (e.g.,  $|j\rangle_a$ ). This convention helps to distinguish between the roles of data and address qubits in QRAM operations.

**Example 2.18. (QRAM Encoding).** Consider a dataset  $\mathcal{D} = \{2, 3\}$ . Using basis encoding, each sample is first converted into a two-qubit quantum state:  $\{|10\rangle_d, |11\rangle_d\}$ . Each data state is then assigned an address state,  $|0\rangle_a$  for the first state  $|10\rangle_d$  and  $|1\rangle_a$  for the second state  $|11\rangle_d$ . The resulting QRAM-encoded state is:

$$|\mathcal{D}\rangle = \frac{1}{\sqrt{2}} (|0\rangle_a |10\rangle_d + |1\rangle_a |11\rangle_d). \quad (2.49)$$

The corresponding quantum circuit for implementing this state is shown in Fig. 2.7.

QRAM allows the dataset  $\mathcal{D}$  to be stored in a coherent quantum superposition, enabling simultaneous access to all data items through the entanglement of address and data qubits. While QRAM is theoretically powerful, its practical implementation remains a significant challenge due to the need for a large number of qubits and quantum operations (see Chapter 2.6 for the discussion).

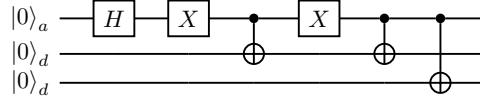


Figure 2.7: Example of QRAM encoding for the dataset  $\mathcal{D} = \{2, 3\}$ .

### 2.3.2 Quantum read-out protocols

Quantum read-out refers to the process of translating the quantum state resulting from a quantum computation into classical data, enabling further processing, interpretation, or optimization in classical systems. This process can be considered the inverse operation of quantum read-in, representing a quantum-to-classical mapping.

Based on the completeness of the information obtained during the read-out process, quantum read-out protocols can be broadly categorized into two types, i.e., *full information and partial information read-out protocols*. These protocols enable tailored read-out processes that match the requirements of different quantum applications, ranging from tomography to optimization and machine learning tasks.

#### Full information read-out protocol

The full information read-out protocol is used to completely reconstruct the quantum state, which is used to fully understand the quantum system's behavior. The most general approach to implementing this protocol is through quantum state tomography (QST) ([Vogel and Risken, 1989](#)).

QST involves performing quantum measurements, gathering measurement statistics, and using classical post-processing to reconstruct the quantum state. In what follows, we introduce two reconstruction techniques broadly used in QST, i.e., linear inversion ([Qi et al., 2013](#)) and maximum likelihood estimation (MLE) ([Hradil, 1997](#)).

*QST with linear inversion.* Linear inversion is a straightforward method to reconstruct the quantum state from measurement data by directly solving linear systems of equations. Let  $\rho$  be the explored quantum state and  $\{E_i\}$  be a set of measurements. According to the Born rule, the probability of measurement outcome  $i$  is given by

$$\Pr(E_i|\rho) = \text{Tr}(\rho E_i). \quad (2.50)$$

In practice,  $\Pr(E_i|\rho)$  is not directly accessible but is approximated by the frequency  $p_i$  of measurement outcome  $i$  over multiple measurements. By the

law of large numbers, as the number of measurements increases,  $p_i$  converges to the true probability  $\Pr(E_i|\rho)$ . Collecting measurements across all bases, we obtain a linear system

$$\begin{bmatrix} \text{Tr}(\rho E_0) \\ \text{Tr}(\rho E_1) \\ \vdots \end{bmatrix} = \begin{bmatrix} \vec{E}_0^\dagger \cdot \vec{\rho} \\ \vec{E}_1^\dagger \cdot \vec{\rho} \\ \vdots \end{bmatrix} = A\vec{\rho} \approx \mathbf{p} = \begin{bmatrix} p_0 \\ p_1 \\ \vdots \end{bmatrix}, \quad (2.51)$$

where  $\vec{E}$  and  $\vec{\rho}$  refer to the vector representations of matrices  $E_i$  and  $\rho$ , respectively. The vector representation of a matrix is obtained by stacking its columns into a single-column vector. For example, the vector representation of a  $2 \times 2$  identity matrix is  $\vec{\mathbb{I}}_2 = [1, 0, 0, 1]^T$ . The matrix  $A$  is constructed such that each row corresponds to the vector representation of the measurement operator, i.e.,  $A = [\vec{E}_0^\dagger; \vec{E}_1^\dagger; \dots]$ . The vector  $\mathbf{p}$  contains the measured frequencies  $p_i$ .

Assuming the measurements are tomographically complete, i.e.,  $\{E_i\}$  forms a basis for the system's Hilbert space, we can reconstruct  $\rho$  by solving the following linear systems of equations, i.e.,

$$\vec{\rho} = (A^T A)^{-1} A^T \mathbf{p}. \quad (2.52)$$

A common strategy is to use Pauli operators as measurement bases  $\{E_i\}$ . The density matrix  $\rho$  of an  $N$ -qubit system can be expanded in terms of the Pauli basis as:

$$\rho = \frac{1}{2^N} \sum_{i=0}^{4^N - 1} c_i P_i, \quad c_i \in \mathbb{R}, \quad P_i \in \{I, X, Y, Z\}^{\otimes N}, \quad (2.53)$$

where the coefficients  $c_i$  represent projections of  $\rho$  onto the Pauli basis, calculated as:

$$c_i = \text{Tr}(\rho P_i). \quad (2.54)$$

To fully reconstruct  $\rho$ , the quantum state must theoretically be measured in all  $4^N - 1$  Pauli bases to estimate each  $c_i$ .

### Remark

The Pauli basis consists of four Hermitian matrices  $I$ ,  $X$ ,  $Y$  and  $Z$  introduced in Figure 2.1. These operators form a complete basis for the space of  $2 \times 2$  complex matrices. For  $N$ -qubit systems, the tensor products of these single-qubit operators span the space of  $2^N \times 2^N$

complex matrix. This makes the Pauli basis essential for representing quantum states, observables, and their transformations.

A key limitation of linear inversion is that it does not guarantee a valid density matrix, as the estimated quantum state may lack properties such as positive semi-definiteness in Definition 2.4, especially with limited measurements.

*Maximum Likelihood Estimation (MLE).* To ensure physical constraints on the quantum state during reconstruction, MLE is introduced. MLE reconstructs  $\rho$  by maximizing the likelihood of observing the measurement outcomes, subject to the constraints that  $\rho$  is Hermitian, positive semi-definite, and trace one. The likelihood function is given by

$$L(\rho) = \prod_i \text{Tr}(\rho E_i)^{p_i}. \quad (2.55)$$

Reconstructing  $\rho$  then reduces to solving the following optimization problem

$$\arg \max_{\rho'} L(\rho'), \quad \text{s.t.} \quad \rho' \succeq 0, \quad \rho' = \rho'^\dagger, \quad \text{Tr}(\rho') = 1. \quad (2.56)$$

Solving this typically requires iterative numerical optimization, which can be computationally intensive.

#### Remark

A common challenge across all quantum state tomography (QST) methods, including linear inversion and MLE, is the *exponential computational cost* with respect to the number of qubits. Specifically, the number of parameters to reconstruct grows exponentially with the system size, making QST methods feasible only for small-qubit systems in practice. This limitation underscores the need for scalable approaches to quantum state characterization in larger quantum systems.

#### Partial information read-out protocol

The partial information read-out protocol focuses on extracting specific, task-relevant information from a quantum state without reconstructing the entire density matrix. This protocol is efficient and can be applied to comprehend large-qubit systems. According to the type of collected information,

current partial read-out protocols can be categorized into three classes, i.e., sampling, expectation value estimation, and shadow tomography.

*Sampling.* Sampling involves repeatedly measuring the quantum state in the computational basis to estimate the probability distribution over bit-strings. Given a state  $|\psi\rangle$ , the probability of observing a specific computational basis  $|i\rangle$  is given by

$$\Pr(i) = |\langle\psi|i\rangle|^2. \quad (2.57)$$

The empirical frequency of each outcome from repeated measurements provides an estimate of  $\Pr(i)$ . Sampling is particularly useful in the following applications

- Sampling over complicated distributions. Quantum states can represent complex probability distributions that are difficult to sample classically. Quantum sampling allows efficient exploration of these distributions for specific applications, such as probabilistic modeling and Markov chain Monte Carlo ([Layden et al., 2023](#)).
- Optimization problems. Sampling can identify high-probability bit-strings in quantum algorithms, such as the Quantum Approximate Optimization Algorithm ([Farhi et al., 2014](#)) and Grover search ([Grover, 1996](#)), where these bitstrings often correspond to optimal or near-optimal solutions.
- Verification. Sampling facilitates the comparison of a quantum circuit's output with theoretical expectations or desired distributions, helping to verify the quantum systems ([Boixo et al., 2018; Bouland et al., 2019; Arute et al., 2019](#)).

*Expectation value estimation.* For a wide class of quantum computation problems, such as quantum chemistry and quantum many-body physics, the computation outcome refers to the estimation of the expectation values of certain observables on the evolved quantum state ([Kandala et al., 2017a; Tilly et al., 2022](#)).

An observable  $O \in \mathbb{C}^{2^N \times 2^N}$  mentioned here is a Hermitian operator that represents a measurable physical quantity. For an  $N$ -qubit system,  $O$  can be expressed in terms of a Pauli basis expansion, i.e.,

$$O = \sum_{i=1}^{4^N} \alpha_i P_i, \quad P_i \in \{\mathbb{I}_2, X, Y, Z\}^{\otimes N}, \quad \alpha_i \in \mathbb{R}. \quad (2.58)$$

where  $P_i$  denotes the  $i$ -th  $N$ -qubit Pauli string.

The expectation value of an observable  $O$  with respect to an  $N$ -qubit state  $\rho$  is

$$\langle O \rangle = \text{Tr}(\rho O). \quad (2.59)$$

Substituting the Pauli expansion of  $O$ , the expectation value is expressed as the weighted sum of the expectation values of each Pauli basis term due to the linearity of the trace operation, i.e.,

$$\langle O \rangle = \sum_{i=1}^{4^N} \alpha_i \text{Tr}(\rho P_i) \equiv \sum_{i=1}^{4^N} \alpha_i \langle P_i \rangle. \quad (2.60)$$

To estimate the expectation value of each individual Pauli term  $P_i$ , the quantum state  $\rho$  must be measured on the basis of the eigenstates of  $P_i$ . The measurement outcome is then associated with the corresponding eigenvalue of  $P_i$ . Notably, the eigenstates and eigenvalues of  $P_i$  can be derived from the eigenstates and eigenvalues of its constituent single-qubit Pauli operators  $P_{ij}$ :

- Eigenvalues. The eigenvalues of  $P_i$  are the product of the eigenvalues of each single-qubit Pauli operator  $P_{ij}$ , i.e.,  $P_i = \otimes_{j=1}^N P_{ij}$ . For example, if the eigenvalues of  $P_{ij}$  are  $\pm 1$ , then the eigenvalues of  $P_i$  are products of these individual eigenvalues and remain in  $\{\pm 1\}$ .
- Eigenstates. The eigenstates of  $P_i$  are the tensor products of the eigenstates of the single-qubit Pauli operators  $P_{ij}$ . If  $|\lambda_{ijk}\rangle$  is one of the eigenstate of  $P_{ij}$ , then the corresponding eigenstate of  $P_i$  is  $\otimes_{j=1}^N |\lambda_{ijk}\rangle$ .

This structure allows  $P_i$  to be analyzed in terms of its simpler single-qubit components, significantly simplifying the process of determining the measurement basis for expectation value estimation. By repeating the measurements  $M$  times and obtaining the corresponding measurement results  $\{r_j\}_{j=1}^M$ , the statistical value of  $\langle P_i \rangle$  can be estimated by

$$\langle \hat{P}_i \rangle = \frac{1}{M} \sum_{j=1}^M r_j. \quad (2.61)$$

The expectation value of the observable  $O$  is therefore statistically estimated by  $\langle \hat{O} \rangle = \sum_{i=0}^{K-1} \alpha_i \langle \hat{P}_i \rangle$ .

## Remark

A key step in the process is to measure the quantum system in the basis of the eigenstates of  $P_i$ . If  $P_i$  is diagonal in the computational basis (e.g., a tensor product of Pauli-Z operators), we can directly measure the state without additional operations. Otherwise (e.g., for Pauli-X or Pauli-Y operators), we need to apply a unitary transformation to rotate the quantum state into the desired basis. Specifically, when measuring in the Pauli-X basis (i.e.,  $|+\rangle$  and  $|-\rangle$ ), a Hadamard gate  $H$  is applied to the state  $\rho$ , i.e.,

$$\rho' = H\rho H. \quad (2.62)$$

When measuring in the Pauli-Y basis (i.e.,  $\frac{|0\rangle+i|1\rangle}{\sqrt{2}}$  and  $\frac{|0\rangle-i|1\rangle}{\sqrt{2}}$ ), a phase gate  $S = \sqrt{Z}$  followed by a Hadamard gate  $H$  is applied, i.e.,

$$\rho' = S^\dagger H\rho HS. \quad (2.63)$$

Measuring the state  $\rho'$  in the computational basis is equivalent to measuring the state  $\rho$  in the corresponding Pauli basis.

*Shadow tomography.* Performing full QST requires an exponential number of copies of the quantum state, making it impractical for systems beyond a small number of qubits. Instead of reconstructing the complete density matrix, shadow tomography [Aaronson \(2018\)](#) focuses on efficiently obtaining specific properties of a quantum state, such as the expectation values of many observables.

**Definition 2.19** (Shadow tomography, [Aaronson \(2018\)](#)). *Given an unknown  $D$ -dimensional quantum state  $\rho$ , as well as  $M$  observables  $O_1, \dots, O_M$ , output real numbers  $b_1, \dots, b_M$  such that  $|b_i - \text{Tr}(O_i\rho)| \leq \epsilon$  for all  $i$ , with success probability at least  $1 - \delta$ . Do this via a measurement of  $\rho^{\otimes k}$ , where  $k = k(D, M, \epsilon, \delta)$  is as small as possible.*

[Aaronson \(2018\)](#) proved that the shadow tomography problem can be solved using a *polylogarithmic* number of copies of states in terms of the dimension  $D$  and number  $M$  of observables. This result demonstrates that it is possible to estimate the expectation values of exponentially many observables for a quantum state of exponential dimension using only a polynomial number of measurements.

The central idea of shadow tomography is to create a compact measurement classical representation, or “shadow”, of a quantum state that encodes sufficient information to estimate many properties of the state. Building on this concept, [Huang et al. \(2020b\)](#) proposed a more practical and efficient approach, termed *classical shadow*, which uses randomized measurements to construct this classical representation. The classical shadow approach consists of the following steps:

1. Randomized measurements. Perform random unitary transformations on the quantum state and measure the transformed state in the computational basis. These random transformations can be drawn from specific ensembles, such as Clifford gates or local random rotations, which ensure that the measurement outcomes capture the essential properties of the quantum state.
2. Classical shadow construction. Using the measurement results, construct a classical shadow of the quantum state. This compact representation encodes the quantum state in a way that allows for the efficient estimation of properties.
3. Property estimation. Use the classical shadow to compute the desired properties of the quantum state, such as expectation values of specific observables, subsystem entropies, or fidelities with known states.

Shadow tomography requires exponentially fewer measurements compared to full quantum state tomography, making it a practical solution for large-scale quantum systems. Moreover, the shadow of a quantum state serves as a versatile representation, enabling the efficient estimation of various properties such as expectation values, entanglement measures, and subsystem correlations.

## 2.4 Quantum Linear Algebra

We next introduce quantum linear algebra, a potent toolbox for designing various FTQC-based algorithms introduced in Chapter [1.2.2](#). For clarity, we start with the definition of block encoding in Chapter [2.4.1](#), which is about how to implement a matrix on the quantum computer. Based on this, we introduce some basic arithmetic rules for block encodings in Chapter [2.4.2](#), like the multiplication, linear combination, and the Hadamard product. Finally, in Chapter [2.4.3](#), we introduce the quantum singular value transformation

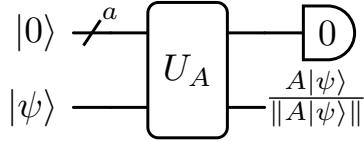


Figure 2.8: **Quantum circuit for block encoding.**

method, which enables one to implement functions onto singular values of block-encoded matrices.

#### 2.4.1 Block encoding

For many computational problems, such as solving linear equations, we need to deal with a non-unitary matrix  $A$ . However, remember that quantum gates as discussed in Chapter 2.2 are unitaries. Therefore, if we want to solve these problems on quantum computers, it is essential to consider how to encode the matrix  $A$  into a unitary. This challenge can be addressed by the block encoding technique.

**Definition 2.20** (Block encoding, [Gilyén et al. \(2019\)](#)). *Suppose that  $A$  is an  $N$ -qubit operator,  $\alpha, \varepsilon \geq 0$  and  $a \in \mathbb{N}$ . Then we say that the  $(a+N)$ -qubit unitary  $U$  is an  $(\alpha, a, \varepsilon)$ -block-encoding of  $A$  if*

$$\|A - \alpha(\langle 0 |^{\otimes a} \otimes \mathbb{I}_{2^N})U(|0\rangle^{\otimes a} \otimes \mathbb{I}_{2^N})\| \leq \varepsilon. \quad (2.64)$$

Here,  $\|\cdot\|$  represents the spectral norm, i.e., the largest singular value of the matrix.

The circuit implementation of the block encoding is illustrated in Figure 2.8. The scaled matrix  $A/\alpha$  interacts with the state  $|\psi\rangle$  if the first qubit registers are measured as  $|0\rangle$ . By definition, we have  $\alpha \geq \|A\|$  and any unitary  $U$  is an  $(1, 0, 0)$ -block encoding of itself.

**Fact 2.21.** (*Block encoding via the linear combination of unitaries (LCU) method, [Gilyén et al. \(2019\)](#)*). *Suppose that  $A$  can be written in the form*

$$A = \sum_k \alpha_k U_k, \quad (2.65)$$

where  $\{\alpha_k\}$  are real numbers and  $U_k$  are some easily prepared unitaries such as Pauli strings. Then, the LCU method allows us to have the access to two

*unitaries, i.e.,*

$$U_{\text{SEL}} = \sum_k |k\rangle\langle k| \otimes U_k, \quad (2.66)$$

$$U_{\text{PREP}} : |0\rangle \rightarrow \frac{1}{\sqrt{\|\vec{\alpha}\|_1}} \sum_k \sqrt{\alpha_k} |k\rangle, \quad (2.67)$$

where  $\vec{\alpha} = (\alpha_1, \alpha_2, \dots)$ .

After simple mathematical analysis, one can obtain  $U = (U_{\text{PREP}}^\dagger \otimes \mathbb{I}_{2^N})U_{\text{SEL}}(U_{\text{PREP}} \otimes \mathbb{I}_{2^N})$  is a  $(\|\vec{\alpha}\|_1, m, 0)$ -block-encoding of  $A$ . Here,  $\mathbb{I}_{2^N}$  is the identity operator of  $N$ -qubit size and  $\|\cdot\|_1$  denotes the  $\ell_1$  norm of a given vector.

Similar to the definition of block encoding, we can also define the state preparation encoding.

**Definition 2.22** (State preparation encoding [Guo et al. \(2024a\)](#)). *We say a unitary  $U_\psi$  is an  $(\alpha, a, \epsilon)$ -state-encoding of an  $N$ -qubit quantum state  $|\psi\rangle$  if*

$$\||\psi\rangle - \alpha(\langle 0^a | \otimes I)U_\psi |0^{a+N}\rangle\|_\infty \leq \epsilon, \quad (2.68)$$

where  $\|\cdot\|_\infty$  denotes the infinity norm of the given vector.

More straightforwardly, the  $(\alpha, a, \epsilon)$ -state-encoding  $U_\psi$  prepares the state

$$U_\psi |0\rangle |0\rangle = \frac{1}{\alpha} |0\rangle |\psi'\rangle + \sqrt{1 - \alpha^2} |1\rangle |\text{bad}\rangle,$$

where  $\||\psi'\rangle - |\psi\rangle\|_\infty \leq \epsilon$  and  $|\text{bad}\rangle$  is an arbitrary quantum state. One can further prepare the state  $|\psi'\rangle$  by using  $\mathcal{O}(\alpha)$  times of amplitude amplification ([Brassard et al., 2002](#)). The state preparation encoding can be understood as a specific case of the block encoding, i.e., it is the block encoding of a  $\mathbb{C}^{2^N \times 1}$  matrix.

## 2.4.2 Basic arithmetic for block encodings

Now we introduce some arithmetic rules for block encoding unitaries. The following two facts describe the product and linear combination rules of block encoding unitaries, respectively.

**Fact 2.23** (Product of block encoding, [Gilyén et al. \(2019\)](#)). *If  $U$  is an  $(\alpha, a, \delta)$ -block encoding of an  $N$ -qubit operator  $A$ , and  $V$  is a  $(\beta, b, \varepsilon)$ -block encoding of an  $N$ -qubit operator  $B$ , then  $(\mathbb{I}_{2^b} \otimes U)(\mathbb{I}_{2^a} \otimes V)$  is an  $(\alpha\beta, a + b, \alpha\varepsilon + \beta\delta)$ -block-encoding of  $AB$ . Here,  $\mathbb{I}_{2^a}$  is the identity operator of  $a$ -qubit size.*

**Fact 2.24** (Linear combination of block encoding, Gilyén et al. (2019)). *Let  $A = \sum_k x_k A_k$  be an  $s$ -qubit operator with  $\beta \geq \|\vec{x}\|_1$  and  $\varepsilon_1 > 0$ , where  $\vec{x}$  is the vector of coefficients. Suppose we have access to*

$$P_L |0\rangle = \sum_k c_k |k\rangle, \quad (2.69)$$

$$P_R |0\rangle = \sum_k d_k |k\rangle, \quad (2.70)$$

$$W = \sum_k |k\rangle\langle k| \otimes U_k + \left( (\mathcal{I}_s - \sum_k |k\rangle\langle k|) \otimes \mathcal{I}_a \otimes \mathcal{I}_b \right), \quad (2.71)$$

where  $\sum_k |\beta c_k^* d_k - x_k| \leq \varepsilon_1$  and  $U_k$  is an  $(\alpha, a, \varepsilon_2)$ -block-encoding of  $A_k$ . Then we can implement an  $(\alpha\beta, a+b, \alpha\varepsilon_1 + \beta\varepsilon_2)$ -block-encoding of  $A$  by using one time of  $W$ ,  $P_L$ , and  $P_R$ .

These results can be verified via direct computation. Another arithmetic rule broadly employed in quantum machine learning is the Hadamard product, a.k.a, the element-wise product. The following lemma exhibits how to achieve this operation via the block encoding framework.

**Lemma 2.25** (Hadamard product of the block encoding unitaries, Guo et al. (2024a)). *With  $N \in \mathbb{N}$ , consider two matrices  $A, B \in \mathbb{C}^{2^N \times 2^N}$ , and assume that we have an  $(\alpha, a, \delta)$ -encoding  $U_A$  of matrix  $A$  and  $(\beta, b, \epsilon)$ -encoding  $U_B$  of matrix  $B$ , then we can construct an  $(\alpha\beta, a+b+N, \alpha\epsilon + \beta\delta)$ -encoding of matrix  $A \circ B$  corresponding to the Hadamard product of  $A$  and  $B$ .*

*Proof sketch of Lemma 2.25.* For simplicity, we only consider the perfect case, i.e., no errors. Refer to Ref. (Guo et al., 2024a) for the proof details under the more general cases.

The intuition for achieving the Hadamard product is that all the needed elements can be found in the tensor product, i.e.,

$$\begin{aligned} & (\langle 0^{a+b} | \otimes \mathbb{I}_{2^{2N}})(\mathbb{I}_{2^b} \otimes U_A \otimes \mathbb{I}_{2^N})(\mathbb{I}_{2^a} \otimes U_B \otimes \mathbb{I}_{2^N})(|0^{a+b}\rangle \otimes \mathbb{I}_{2^{2N}}) \\ &= \frac{A \otimes B}{\alpha\beta}. \end{aligned} \quad (2.72)$$

To this end, the question is reduced to finding proper permutation unitaries that can shift the required elements to the correct positions to achieve the Hadamard product. Denote  $P' = \sum_{i=0}^{d-1} |i\rangle\langle i| \otimes |0\rangle\langle i|$ . As proved by Zhao et al. (2021), the tensor product of  $A$  and  $B$  can be reformulated to the Hadamard product via  $P$ , i.e.,

$$P'(A \otimes B)P'^{\dagger} = (A \circ B) \otimes |0\rangle\langle 0|.$$

However,  $P'$  is not a unitary. Instead, we consider  $P = \sum_{i,j=0}^{d-1} |i\rangle\langle i| \otimes |i \oplus j\rangle\langle j|$ , which can be easily constructed by using  $N$  CNOT gates, i.e., one CNOT gate between each pair of qubits consisting of one qubit from the first register and the corresponding qubit from the second register. By direct computation, we have

$$(\mathbb{I}_{2^N} \otimes \langle 0^N |)P(A \otimes B)P^\dagger(\mathbb{I}_{2^N} \otimes |0^N\rangle) = A \circ B. \quad (2.73)$$

Therefore, by direct computation, one can verify that  $(P \otimes \mathbb{I}_{2^{a+b}})(\mathbb{I}_{2^b} \otimes U_A \otimes \mathbb{I}_{2^N})(\mathbb{I}_{2^a} \otimes U_B \otimes \mathbb{I}_{2^N})(P^\dagger \otimes \mathbb{I}_{2^{a+b}})$  is the desired block encoding.  $\square$

### 2.4.3 Quantum singular value transformation

Now we understand how to implement matrices on the quantum computer and some arithmetic methods among these matrices, here we further introduce how one can implement matrix functions on the quantum computer. The method is called the quantum singular value transformation (QSVT), which is a powerful framework that can unify most known quantum algorithms.

For machine learning applications, we mostly deal with the real matrices. The computational cost of QSVT for the real matrix case is summarized in the following theorem. For a matrix  $A$ , consider its singular value decomposition  $A = \sum_i \sigma_i |\psi_i\rangle\langle\phi_i|$ . Given a function  $P(x)$ , we use  $P^{(SV)}(A)$  to represent  $P^{(SV)}(A) = \sum_i P(\sigma_i) |\psi_i\rangle\langle\phi_i|$ .

**Fact 2.26** (Quantum singular value transformation - Real matrix case, Gilyén et al. (2019)). *Suppose that  $U_A$  is an  $(\alpha, a, \varepsilon)$ -block-encoding of a real matrix  $A$ . If  $\delta \geq 0$  and  $P : \mathbb{R} \rightarrow \mathbb{C}$  is a  $d$ -degree polynomial satisfying that*

$$\text{for all } x \in [-1, 1] : |P(x)| \leq \frac{1}{4}, \quad (2.74)$$

*then there is a quantum circuit  $\tilde{U}$ , which is an  $(1, a+3, 4d\sqrt{\varepsilon/\alpha} + \delta)$ -block-encoding of  $P^{(SV)}(A/\alpha)$ , and consists of  $d$  applications of  $U_A$  and  $U_A^\dagger$  gates. Further, the description of such a circuit can be computed classically in time  $\mathcal{O}(\text{poly}(d, \log(1/\delta)))$ .*

Notice that if the block-encoded matrix  $A$  is Hermitian, the singular value transformation is equivalent to the eigenvalue transformation. In this case, we can directly implement the matrix function via QSVT.

In the following, we introduce some applications of QSVD method. There are several important applications, however, in this tutorial, we focus on introducing some applications that can be applied to machine-learning related tasks. The first is matrix inversion, which has been widely used in traditional machine learning methods such as principal component analysis. Note that for a general matrix, we actually mean to implement the Moore-Penrose pseudoinverse, i.e., the inverse of all singular values.

**Lemma 2.27** (Matrix inversion, simplified [Gilyén et al. \(2019\)](#)). *Let  $U_A$  be a  $(1, a, 0)$ -block encoding of matrix  $A$ . Further, for simplicity, assume the nonzero singular values of  $A$  are lower bounded by  $\delta > 0$ . Let  $0 \leq \epsilon \leq \delta \leq \frac{1}{2}$ . One can construct a  $(1/\delta, a+2, \epsilon)$ -block encoding of  $A^{-1}$  by using  $\tilde{\mathcal{O}}(\frac{1}{\delta} \log(\frac{1}{\epsilon}))$  times of  $U_A$  and  $U_A^\dagger$ .*

*Proof sketch of Lemma 2.27.* This can be achieved by finding a good polynomial approximation for the function  $1/x$ . One can not find such a polynomial on the whole interval  $[-1, 1]$ , however, such a polynomial exists on the interval  $[-1, -\delta] \cup [\delta, 1]$  for some  $\delta > 0$ .  $\square$

The second application of QSVD is the nonlinear amplitude transformation. As mentioned in Chapter 2.3.1, to deal with classical data, there are several ways to encode them into quantum. Here, we focus on the amplitude encoding case described in Chapter 2.3.1, especially for the real amplitudes. The nonlinear transformation is achieved by combining the diagonal block encoding and QSVD.

**Fact 2.28** (Diagonal block encoding of amplitudes, [Guo et al. \(2024b\)](#); [Rattew and Rebentrost \(2023\)](#)). *Given a state preparation unitary  $U_\psi$  of an  $N$ -qubit state  $|\psi\rangle = \sum_{j=1}^{2^N} \psi_j |j\rangle$ , where  $\{\psi_j\}$  are real,  $\|\psi\|_2 = 1$ , one can construct an  $(1, N+2, \epsilon)$ -encoding of the diagonal matrix  $A = \text{diag}(\psi_1, \dots, \psi_d)$  with  $\mathcal{O}(N)$  circuit depth and  $\mathcal{O}(1)$  times of controlled- $U$  and controlled- $U^\dagger$ .*

As a straightforward generalization, one can replace the state preparation unitary with the general state preparation encoding, mentioned in Definition 2.22. By constructing the block encoding of amplitudes, one can implement many functions onto these amplitudes via QSVD. A direct application is performing the neural network on the quantum computer, as will be detailed in Chapter 5.2.

## 2.5 Code Demonstration

This section provides code implementations for key techniques introduced earlier, including quantum read-in strategies and block encoding, to give readers the opportunity to practice and deepen their understanding.

### 2.5.1 Read-in implementations

This subsection demonstrates toy examples of implementing data encoding methods in quantum computing, as discussed in earlier sections. Specifically, we cover basis encoding, amplitude encoding, and angle encoding from Chapter 2.3.1. These examples aim to provide readers with hands-on experience in applying quantum data encoding techniques.

#### Basis encoding

PennyLane provides built-in support for basis encoding through its ‘BasisEmbedding’ function. Below is the Python code demonstrating the basis encoding for the integer 6.

```

1 import pennylane as qml
2
3 dev = qml.device("default.qubit", range(3))
4 @qml.qnode(dev)
5 def circuit(x):
6     qml.BasisEmbedding(x, range(3))
7     return qml.state()
8
9 # Call the function
10 circuit(6)
```

#### Amplitude encoding

PennyLane offers built-in support for amplitude encoding via the ‘AmplitudeEmbedding’ function. Below is a Python example demonstrating amplitude encoding for a randomly generated complex vector.

```

1 import pennylane as qml
2 import numpy as np
3
4 # Number of qubits
5 n_qubits = 8
6
```

```

7 # Define a quantum device with 8 qubits
8 dev = qml.device("default.qubit", wires=n_qubits)
9
10 @qml.qnode(dev)
11 def circuit(x):
12     qml.AmplitudeEmbedding(features=x, wires=range(
13         n_qubits), normalize=True, pad_with=0.)
14     return qml.state()
15
16 # Generate a random complex vector of length 2^n_qubits
17 x_real = np.random.normal(loc=0, scale=1.0, size=2**n_qubits)
18 x_imag = np.random.normal(loc=0, scale=1.0, size=2**n_qubits)
19 x = x_real + 1j * x_imag
20
21 # Execute the circuit to encode the vector as a quantum
22     state
23 circuit(x)

```

### Angle encoding

PennyLane provides built-in support for angle encoding via the ‘AngleEmbedding’ function. Below is a Python example demonstrating angle encoding for a randomly generated real vector.

```

1 import pennylane as qml
2 import numpy as np
3
4 # Number of qubits
5 n_qubits = 8
6
7 # Define a quantum device with 8 qubits
8 dev = qml.device("default.qubit", wires=n_qubits)
9
10 @qml.qnode(dev)
11 def circuit(x):
12     qml.AngleEmbedding(features=x, wires=range(n_qubits),
13         rotation="X")
14     return qml.state()
15
16 # Generate a random real vector of length n_qubits
17 x = np.random.uniform(0, np.pi, (n_qubits))

```

```

18 # Execute the circuit to encode the vector as a quantum
19     state
  circuit(x)

```

### 2.5.2 Block encoding

Here, we provide an example of how we may construct a block encoding. We construct the block encoding via the linear combination, as fact 2.21. We use PennyLane to keep consistency, yet there are many other platforms that are available as well. Please note that it is time-consuming to do the Pauli decomposition (for an  $N$ -qubit matrix, it takes time  $\mathcal{O}(N4^N)$ ), so we suggest not trying a large matrix with this method.

```

1 import numpy as np
2 import pennylane as qml
3 import matplotlib.pyplot as plt
4
5 a = 0.36
6 b = 0.64
7
8 # matrix to be decomposed
9 A = np.array(
10     [[a, 0, 0, b],
11      [0, -a, b, 0],
12      [0, b, a, 0],
13      [b, 0, 0, -a]])
14
15
16 # decompose the matrix into sum of Pauli strings
17 LCU = qml.pauli_decompose(A)
18 LCU_coeffs, LCU_ops = LCU.terms()
19
20 # normalized square roots of coefficients
21 alphas = (np.sqrt(LCU_coeffs) / np.linalg.norm(np.sqrt(
22     LCU_coeffs)))
23
24 dev = qml.device("default.qubit", wires=3)
25
26 # unitaries
27 ops = LCU_ops
28 # relabeling wires: 0 --> 1, and 1 --> 2
29 unitaries = [qml.map_wires(op, {0: 1, 1: 2}) for op in
    ops]

```

```

30 @qml.qnode(dev)
31 def lcu_circuit(): # block_encode
32     # PREP
33     qml.StatePrep(alphas, wires=0)
34
35     # SEL
36     qml.Select(unitaries, control=0)
37
38     # PREP_dagger
39     qml.adjoint(qml.StatePrep(alphas, wires=0))
40     return qml.state()
41
42 print(np.real(np.round(output_matrix,2)))

```

## 2.6 Bibliographic Remarks

We end this chapter by discussing the recent advancements in efficiently implementing fundamental components of quantum computing. For clarity, we begin with a brief discussion of advanced quantum read-in and read-out protocols, which are crucial for efficiently loading and extracting classical data in the pipeline of quantum machine learning. Next, we review the latest progress in quantum linear algebra.

### 2.6.1 Advanced quantum read-in protocols

Although conventional read-in protocols offer feasible solutions for encoding classical data into quantum computers, they typically face two key challenges that limit their broad applicability for solving practical learning problems. To address these limitations, initial efforts have been made to develop more advanced quantum read-in protocols.

*Challenge I: high demand for quantum resources.* Encoding methods like amplitude encoding and basis encoding presented in Chapter 2.3.2 generally suffer from high quantum resource requirements. While amplitude encoding is highly compact in terms of qubit requirements, the trade-off is the requirement of an exponential number of quantum gates with the data size to prepare an exact amplitude-encoded state. In contrast, while basis encoding can be implemented with a small number of quantum gates, it requires a large number of qubits proportional to the input size. The high demand for either quantum gates or qubit counts makes these basic encoding strategies infeasible for practical use.

*Challenge II: insufficient nonlinearity.* While quantum mechanics is inherently linear, most practical machine learning models require nonlinearity to capture complex data patterns effectively. Conventional encoding methods like angle encoding introduce some degree of nonlinearity; however, the representational power remains limited due to the linear nature of quantum operations and limited circuit depth.

For Challenge I, a practical alternative is the approximate amplitude encoding (AAE) ([Nakaji et al., 2022](#)). Instead of implementing exact amplitude encoding, AAE trains a parameterized quantum circuit with a constrained depth to approximate the desired quantum state with high fidelity. The training process optimizes the fidelity between the target state and the approximate state, ensuring that the representation error remains within a small bound.

For Challenge II, techniques like *data re-uploading* ([Pérez-Salinas et al., 2020](#)) have been developed. *Data re-uploading* involves feeding the same classical data into the quantum circuit multiple times, interspersed with trainable quantum operations. By alternating data encoding with trainable transformations, this approach allows the quantum model to capture non-linear relationships more effectively without requiring additional qubits. Additionally, neural quantum embedding ([Hur et al., 2024](#)) has been proposed, which leverages classical deep learning techniques to learn optimal quantum embeddings, effectively separating non-linearly separable classes of data.

To address both Challenges I & II, hybrid encoding strategies have been introduced to leverage the respective advantages of each encoding method. For instance, basis-amplitude encoding combines basis encoding for discrete random variables with amplitude encoding for high-precision probabilities, effectively encoding both categorical and continuous features without requiring additional qubits ([Schuld et al., 2018](#)). Another widely used strategy involves classical preprocessing methods for high-dimensional data, such as principal component analysis (PCA) ([Abdi and Williams, 2010](#)), to reduce input dimensionality before applying quantum encoding. This preprocessing step reduces the overall quantum resource requirements while preserving relevant information.

In addition to fixed encoding strategies, learning-based approaches have emerged to dynamically adjust data encoding for specific tasks. For example, [Lloyd et al. \(2020\)](#) achieves task-specific quantum embeddings by incorporating learnable parameters into the encoding layers, which are optimized to maximize class separability in Hilbert space. This technique is

analogous to classical metric learning. Following this routine, a quantum few-shot embedding framework (Liu et al., 2022) has been proposed to encode classical data into quantum states, which can be generalized to the downstream quantum machine learning tasks. These methods enable quantum circuits to adapt their encodings dynamically, improving efficiency and performance.

### 2.6.2 Advanced quantum read-out protocols

Conventional quantum read-out protocols often face significant challenges, including high computational overhead and resource inefficiencies. Below, we discuss the primary challenges and discuss solutions in two quantum read-out protocols: QST and observable estimation.

*Challenge I: High computational overhead of QST.* QST aims to reconstruct the density matrix of a quantum state, but this becomes computationally infeasible as the system size increases. This is because the required number of measurements and the classical memory grows exponentially with the number of qubits.

*Challenge II: Resource inefficiency in observable estimation.* The required number of measurements for observable estimation grows linearly with the number of Pauli terms in the observable. For observables where the number of Pauli terms substantially increases with the system size, the measurement cost becomes prohibitive.

For Challenge I, the key idea is to focus on representing only a subspace of the quantum space, effectively capturing task-relevant properties while reducing the computational cost. For example, in many QML algorithms, such as the HHL algorithm for solving linear systems (Harrow et al., 2009) and quantum singular value decomposition (Rebentrost et al., 2018a), the solution state exists within the row or column space of the input matrix. When the input matrix is low-rank, state tomography can be obtained efficiently (Zhang et al., 2021a) as the linear combination of a complete basis chosen from the input matrix. Besides, an effective technique is matrix product state (MPS) tomography (Lanyon et al., 2017; Orús, 2019), which leverages the fact that many practical quantum states, such as those in Ising models or low-entanglement systems, can be efficiently represented with a reduced number of parameters. By focusing on states with limited entanglement, MPS tomography reconstructs the state using only a polynomial number of measurements with the qubit counts.

Another promising approach is the use of neural networks to parameterize quantum states. Neural quantum states allow for the efficient repre-

sentation and reconstruction of density matrices, particularly for complex or high-dimensional quantum systems. For instance, Restricted Boltzmann Machines (Fischer and Igel, 2012) and Transformer (Vaswani, 2017) have been applied to approximate the probability of measurement outcome and density matrices (Torlai et al., 2018; Schmale et al., 2022; Wang et al., 2022a; Zhao et al., 2023). These approaches are particularly effective for systems that are difficult to capture using traditional methods.

For Challenge II, a measurement reduction technique can be applied by exploiting the commutativity of Pauli operators. When multiple Pauli terms commute, they can be measured simultaneously within the same measurement basis, significantly reducing the total measurement cost (Kandala et al., 2017b; Verteletskyi et al., 2020). This approach has been widely adopted in hybrid quantum-classical algorithms, such as variational quantum Eigensolvers (VQE) (Cerezo et al., 2021a), where Hamiltonians are decomposed into sums of Pauli terms. Grouping commuting terms into clusters allows for efficient measurement strategies while preserving accuracy.

In addition to measurement grouping, adaptive measurement strategies further improve resource allocation during expectation value estimation. The key observation is that not all Pauli terms contribute equally to the total observable—terms with higher variance require more measurement shots for reliable estimation, while low-variance terms can be measured with fewer shots. Building on this insight, adaptive shot allocation techniques (Rubin et al., 2018; Arrasmith et al., 2020; Qian et al., 2024) dynamically distribute measurement resources across Pauli terms based on their statistical properties and achieve more accurate estimations with a finite measurement budget.

### 2.6.3 Advanced quantum linear algebra

Quantum linear algebra, based on the block encoding and quantum singular value transformation framework, has proven its power for the design of quantum algorithms. Compared to the traditional subroutines like quantum phase estimation and quantum arithmetic (Kitaev, 1995; Ruiz-Perez and Garcia-Escartin, 2017), quantum linear algebra can exponentially improve the dependency on precision (Gilyén et al., 2019). However, a major drawback is that it can only deal with the singular values of block-encoded matrices.

A natural consideration is to generalize the singular value transformation to the eigenvalue transformation. One strong motivation from the application aspect for this is to solve the differential equations on the quantum

computer (Liu et al., 2021a; Childs et al., 2021; An et al., 2021; Jin et al., 2022; Shang et al., 2024). This remains an active research field. Quantum eigenvalue processing, proposed by (Low and Su, 2024), focuses on matrices with real spectra and Jordan forms, in which they prepare the Faber history state to achieve efficient eigenvalue transformation over the complex plane. (An et al., 2023, 2024) shows that simulating a general class of non-unitary dynamics can be achieved by the linear combination of Hamiltonian simulation (LCHS).

Another approach is to broaden the range of functions that can be implemented by quantum linear algebra. Quantum phase processing, proposed by (Wang et al., 2023b), can directly apply arbitrary trigonometric transformations to eigenphases of a unitary operator. Similar results have been independently obtained by Motlagh and Wiebe (2024). In addition, Rossi and Chuang (2022) investigates how to implement multivariate functions. For the application, a representative example is the multivariate state preparation achieved by (Mori et al., 2024), enabling the amplitude encoding of classical multivariate data.

In Chapter 2.4, we introduce the concept of diagonal block encoding, which can convert a state preparation unitary into a block encoding. As the efficient construction of block encodings is a prerequisite for achieving end-to-end quantum advantage, an important research direction is to investigate which types of matrices can be efficiently prepared. By leveraging state-of-the-art techniques in quantum state preparation (Zhang et al., 2022a; Sun et al., 2023) and the linear combination of unitaries (Childs and Wiebe, 2012), it is possible to efficiently construct block encodings for certain classes of matrices (Guseynov and Liu, 2024; Guseynov et al., 2024). Additionally, explicit constructions have been explored for specific types of sparse matrices (Camps et al., 2023).



## Chapter 3

# Quantum Kernel Methods

The fundamental goal of machine learning (ML) algorithms is to learn the underlying feature representations embedded in the training data, allowing data points to be effectively modeled using simple models like linear classifiers. **Kernel methods** are a powerful approach to achieving this by enabling non-linear patterns to be captured in a computationally efficient manner. In kernel methods, a kernel function is defined as the inner product between the high-dimensional feature representations of data points. These feature representations are generated by a hidden feature map that transforms the original data into a higher-dimensional space where complex patterns become easier to identify and model. The kernel function, therefore, serves as a measure of similarity between data points in this transformed space.

The effectiveness of kernel methods heavily depends on the hidden feature map's ability to capture relevant patterns in the data. The more effectively this feature map can reveal the underlying structure, the better the kernel method's performance in learning and generalizing from the data. However, classical kernel methods are inherently limited by the types of patterns they can recognize, as these are constrained by classical computational frameworks. Essentially, classical models excel at detecting patterns they are specifically designed to recognize but may struggle with patterns that deviate from this framework.

In contrast, quantum mechanics is known for generating complex, non-intuitive patterns that are often beyond the reach of classical algorithms. Quantum systems can produce statistical correlations that are computationally challenging—or even impossible—for classical computers to replicate. This suggests that employing quantum circuits as hidden feature maps

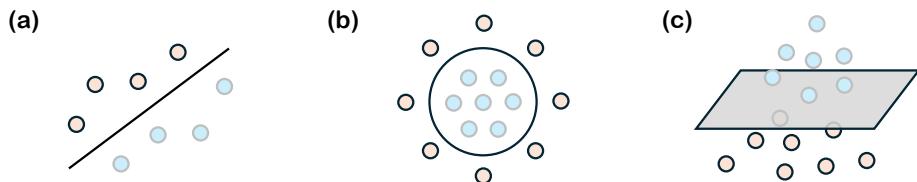
could enable the detection of patterns that are difficult or impractical for classical models to capture. By leveraging quantum circuits, we can potentially access new regions of the feature space, leading to improved pattern recognition capabilities and, consequently, better learning performance.

These insights motivate the development of **quantum kernel methods**, where both the hidden feature map and the kernel function are implemented on a quantum computer. By harnessing the unique properties of quantum mechanics, such as superposition and entanglement, quantum kernel methods have the potential to surpass their classical counterparts in specific machine learning tasks, particularly those involving highly complex or subtle patterns. This could result in more powerful models with enhanced generalization capabilities.

In this chapter, we provide a step-by-step explanation of the transition from classical kernel machines to quantum kernel machines in Chapter 3.1 and Chapter 3.2. Moreover, we discuss the theoretical foundation of quantum kernel machines in Chapter 3.3 from the aspects of expressivity and generalization of quantum kernel machines. Finally, we demonstrate simple yet illustrative code implementations on MNIST dataset.

## 3.1 Classical Kernel Machines

### 3.1.1 Motivation of kernel methods



**Figure 3.1: Various distributions of data points.** The left and middle panels show the cases where data points can and cannot be separated by a straight line. The right panel shows that the kernel function could map the linearly inseparable data points into the high dimensional linearly separable data points.

Before delving into kernel machines, it is essential to first understand the motivation behind kernel methods. In many machine learning tasks, particularly in classification, the goal is to find a decision boundary that best

separates different classes of data. When the data is linearly separable, this boundary can be represented as a straight line (in 2D), a plane (in 3D), or a hyperplane (in higher dimensions), as illustrated in Figure 3.1(a). Mathematically, given an input space  $\mathcal{X} \subset \mathbb{R}^d$  with  $d \geq 1$  and a target or output space  $\mathcal{Y} = \{+1, -1\}$ , we consider a training dataset  $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$  where each data point  $\mathbf{x}^{(i)} \in \mathcal{X}$  is associated with a label  $y^{(i)} \in \mathcal{Y}$ . For the dataset to be linearly separable, there must exist a vector  $\mathbf{w} \in \mathbb{R}^d$  and a bias term  $b \in \mathbb{R}$  such that

$$\forall i \in [n], \quad y^{(i)}(\mathbf{w}^\top \mathbf{x}^{(i)} + b) \geq 0, \quad (3.1)$$

where  $\mathbf{w}^\top \mathbf{x}^{(i)}$  represents the inner product of vectors  $\mathbf{w}$  and  $\mathbf{x}^{(i)}$ . This means that a hyperplane defined by  $(\mathbf{w}, b)$  can perfectly separate the two classes.

However, in real-world scenarios, data is often not linearly separable, as shown in Figure 3.1(b). The decision boundary required to separate classes may be curved or highly complex. Traditional linear models struggle with such non-linear data because they are inherently limited to creating only linear decision boundaries. This limitation highlights the need for more flexible approaches.

To address the challenge of non-linear data, one effective strategy is to transform the input data into a higher-dimensional space where the data may become linearly separable. This transformation is known as *feature mapping*, denoted by

$$\phi : \mathbf{x} \rightarrow \phi(\mathbf{x}) \in \mathbb{R}^D, \quad (3.2)$$

where the original input space  $\mathcal{X}$  is mapped to a higher-dimensional feature space  $\mathbb{R}^D$  with  $D \geq d$ . The idea is that, in this higher-dimensional space, complex patterns in the original data can be more easily identified using linear models.

However, explicitly computing the feature map  $\phi(\mathbf{x})$  in Eqn. (3.2) can be computationally expensive, especially if the feature space is high-dimensional or even infinite-dimensional. Fortunately, many machine learning algorithms for tasks like classification or regression depend primarily on the inner product between data points, which will be explained in Chapter 3.1.2. In the feature space, this inner product is given by  $\langle \phi(\mathbf{x}^{(i)}), \phi(\mathbf{x}^{(j)}) \rangle$ .

### Remark

Throughout the whole tutorial, we interchangeably use  $\mathbf{a}^\top \mathbf{b}$ ,  $\mathbf{a} \cdot \mathbf{b}$ ,  $\langle \mathbf{a}, \mathbf{b} \rangle$ , and  $\langle \mathbf{a} | \mathbf{b} \rangle$  to denote the inner production of two vectors  $\mathbf{a}$  and

**b.**

To circumvent the computational cost of explicitly calculating the feature map, we can use a *kernel function*. A kernel function  $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$  is defined as

$$k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \langle \phi(\mathbf{x}^{(i)}), \phi(\mathbf{x}^{(j)}) \rangle. \quad (3.3)$$

This allows us to compute the inner product in the higher-dimensional feature space indirectly, without ever having to compute  $\phi(\mathbf{x})$  explicitly. This approach is commonly known as the *kernel trick*.

By using the kernel function directly within algorithms, we avoid the computational overhead of working in a high-dimensional space. The collection of kernel values for a dataset forms the kernel matrix (or Gram matrix), where each entry is given by  $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ . This matrix is central to many kernel-based algorithms, as it captures the pairwise similarities between all training data points.

To illustrate the kernel trick, let's consider a simple example in a  $d$ -dimensional input space, where  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_d)^\top$ . Suppose we use the polynomial kernel of degree 2, defined as

$$\begin{aligned} k(\mathbf{x}, \mathbf{z}) &= (\mathbf{x}^\top \mathbf{z})^2 \\ &= (\mathbf{x}_1 \mathbf{z}_1 + \dots + \mathbf{x}_d \mathbf{z}_d)^2 \\ &= \sum_{i=1}^d \sum_{j=1}^d \mathbf{x}_i \mathbf{z}_i \mathbf{x}_j \mathbf{z}_j \\ &= [\mathbf{x}_1^2, \dots, \mathbf{x}_d^2, \sqrt{2}\mathbf{x}_1\mathbf{x}_2, \dots, \sqrt{2}\mathbf{x}_d\mathbf{x}_{d-1}] \\ &\quad [\mathbf{z}_1^2, \dots, \mathbf{z}_d^2, \sqrt{2}\mathbf{z}_1\mathbf{z}_2, \dots, \sqrt{2}\mathbf{z}_d\mathbf{z}_{d-1}]^\top \\ &= \phi(\mathbf{x})^\top \phi(\mathbf{z}). \end{aligned} \quad (3.4)$$

Here, we see that the feature mapping, which comprises all second-order terms, takes the form as

$$\phi(\mathbf{x}) = [\mathbf{x}_1^2, \dots, \mathbf{x}_d^2, \sqrt{2}\mathbf{x}_1\mathbf{x}_2, \dots, \sqrt{2}\mathbf{x}_d\mathbf{x}_{d-1}]^\top. \quad (3.5)$$

Notably, directly computing the kernel function  $(\mathbf{x}^\top \mathbf{z})^2$  for a large  $d$  is much more efficient than explicitly calculating the feature map  $\phi(\mathbf{x})$  and then taking the inner product  $\phi(\mathbf{x})^\top \phi(\mathbf{z})$ . Specifically, using the kernel function only requires  $\mathcal{O}(d)$  time, since it involves computing the dot product in the original input space  $\mathbb{R}^d$ . In contrast, if we were to explicitly compute the

transformed feature vectors  $\phi(\mathbf{x})$  and their inner product, the time complexity could increase to  $\mathcal{O}(D)$ , where  $D$  is the dimensionality of the feature space after mapping. For this example of a polynomial kernel with degree 2,  $D$  can grow to  $\mathcal{O}(d^2)$ . This demonstrates the computational efficiency of using the kernel trick.

**Remark**

Throughout this manuscript, we use the notations  $\mathcal{O}$  and  $\Omega$  to represent the asymptotic upper and lower bounds, respectively, on the growth rate of a term, ignoring constant factors and lower-order terms.

### 3.1.2 Dual representation

To understand why many machine learning algorithms rely primarily on the inner products between data points, we need to introduce the concept of the *dual representation*. In essence, many linear parametric models used for regression or classification can be re-cast into an equivalent dual form, where the kernel function evaluated on the training data emerges naturally.

Let's start with a linear regression model with the training dataset  $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n$ , where the parameters are determined by minimizing a regularized sum-of-squares error function

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^n (\mathbf{w}^\top \phi(\mathbf{x}^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2} \mathbf{w}^\top \mathbf{w}, \quad (3.6)$$

where  $\mathbf{w}^\top$  refers to the transpose of model parameters  $\mathbf{w}$ ,  $\phi(\mathbf{x}^{(i)})$  represents the feature mapping of the input  $\mathbf{x}^{(i)}$ , and  $\lambda \geq 0$  is the regularization factor that helps prevent overfitting.

To find the optimal  $\mathbf{w}$ , we set the gradient of  $\mathcal{L}(\mathbf{w})$  with respect to  $\mathbf{w}$  to zero, i.e.,

$$\frac{\partial \mathcal{L}(\mathbf{w})}{\partial \mathbf{w}} = \sum_{i=1}^n (\mathbf{w}^\top \phi(\mathbf{x}^{(i)}) - y^{(i)}) \phi(\mathbf{x}^{(i)}) + \lambda \mathbf{w} = 0, \quad (3.7)$$

From this, we see that the solution for  $\mathbf{w}$  can be expressed as a linear combination of the training data's feature vectors

$$\mathbf{w} = -\frac{1}{\lambda} \sum_{i=1}^n (\mathbf{w}^\top \phi(\mathbf{x}^{(i)}) - y^{(i)}) \phi(\mathbf{x}^{(i)}) = \sum_{i=1}^n \mathbf{a}^{(i)} \phi(\mathbf{x}^{(i)}) := \Phi^\top \mathbf{a}, \quad (3.8)$$

where  $\Phi = [\phi(\mathbf{x}^{(1)}), \dots, \phi(\mathbf{x}^{(n)})]^\top$  is the design matrix, whose  $i$ -th row is given by  $\phi(\mathbf{x}^{(i)})^\top$ . Here, the coefficients  $\mathbf{a}^{(i)}$  are functions of  $\mathbf{w}$ , defined as

$$\mathbf{a}^{(i)} = -\frac{1}{\lambda}(\mathbf{w}^\top \phi(\mathbf{x}^{(i)}) - y^{(i)}). \quad (3.9)$$

Thus, instead of directly optimizing  $\mathbf{w}$ , we can reformulate the problem in terms of the parameter vector  $\mathbf{a}$ , giving rise to a dual representation. By substituting  $\mathbf{w} = \Phi^\top \mathbf{a}$  into the original objective function  $\mathcal{L}(\mathbf{w})$  in Eqn. (3.6), we obtain

$$\mathcal{L}(\mathbf{a}) = \frac{1}{2}\mathbf{a}^\top \Phi \Phi^\top \Phi \Phi^\top \mathbf{a} - \mathbf{a}^\top \Phi \Phi^\top \mathbf{y} + \frac{1}{2}\mathbf{y}^\top \mathbf{y} + \frac{\lambda}{2}\mathbf{a}^\top \Phi \Phi^\top \mathbf{y}, \quad (3.10)$$

where  $\mathbf{y} = (y^{(1)}, \dots, y^{(n)})^\top$  denotes the vector representation of  $n$  training labels. We define the kernel matrix  $K = \Phi \Phi^\top$ , where each element is given by

$$K_{ij} = \phi(\mathbf{x}^{(i)})^\top \phi(\mathbf{x}^{(j)}) = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}), \quad (3.11)$$

using kernel function  $k(\mathbf{x}, \mathbf{x}')$  defined by Eqn. (3.3). The objective function in terms of  $\mathbf{a}$  simplifies to

$$\mathcal{L}(\mathbf{a}) = \frac{1}{2}\mathbf{a}^\top K^2 \mathbf{a} - \mathbf{a}^\top K \mathbf{y} + \frac{1}{2}\mathbf{y}^\top \mathbf{y} + \frac{\lambda}{2}\mathbf{a}^\top K \mathbf{y}, \quad (3.12)$$

Setting the gradient of  $\mathcal{L}(\mathbf{a})$  with respect to  $\mathbf{a}$  to zero give us

$$\mathbf{a} = (K + \lambda \mathbb{I}_n)^{-1} \mathbf{y}, \quad (3.13)$$

where  $\mathbb{I}_n$  is the identity matrix of size  $n \times n$ .

Now, using this dual formulation, we can derive the prediction for a new input  $\mathbf{x}$ . Substituting  $\mathbf{w} = \Phi^\top \mathbf{a}$  in Eqn. (3.8), the prediction of  $\mathbf{x}$  is given by

$$y(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}) = \langle \Phi^\top \mathbf{a}, \phi(\mathbf{x}) \rangle = \mathbf{k}(\mathbf{x})^\top (K + \lambda \mathbb{I}_n)^{-1} \mathbf{y}, \quad (3.14)$$

where  $\mathbf{k}(\mathbf{x}) \in \mathbb{R}^n$  is a vector with elements  $\mathbf{k}_i(\mathbf{x}) = k(\mathbf{x}^{(i)}, \mathbf{x}) = \phi(\mathbf{x}^{(i)})^\top \phi(\mathbf{x})$ . This shows that the dual formulation allows us to express the solution entirely in terms of the kernel function  $k(\mathbf{x}, \mathbf{x}')$ , rather than explicitly working with the feature map  $\phi(\mathbf{x})$ . This is particularly advantageous because it enables us to work in high-dimensional or even infinite-dimensional feature spaces implicitly.

In the dual formulation, we determine the parameter vector  $\mathbf{a}$  by inverting an  $n \times n$  matrix, compared to the original parameter space formulation which requires inverting a  $d \times d$  matrix to determine  $\mathbf{w}$ . Although this may

not seem advantageous when  $n > d$ , the true benefit of the dual formulation lies in its ability to leverage the kernel trick. By expressing the solution in terms of the kernel function, we avoid the explicit computation of the feature vectors  $\phi(\mathbf{x})$ . This allows us to implicitly utilize feature spaces of very high, or even infinite, dimensionality, enabling the model to capture complex, non-linear relationships in the data without the associated computational cost.

#### Remark

We standardize the notation used throughout this chapter to help readers follow the content more easily. The kernel function is represented by the lowercase letter  $k$ , or with subscripts  $k_Q$  and  $k_C$ . The kernel matrix is denoted by the capital letter  $K$ , or with subscripts  $K_Q$  and  $K_C$ . Additionally, we use the bold lowercase letter  $\mathbf{k}(\mathbf{x})$  to represent the vector of kernel values, where each element is given by  $\mathbf{k}_j(\mathbf{x}) = k(\mathbf{x}^{(j)}, \mathbf{x})$ , corresponding to the training points  $\mathbf{x}^{(j)} \in \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\}$ .

### 3.1.3 Kernel construction

To utilize the kernel trick in machine learning algorithms, it is essential to construct valid kernel functions. One approach is to start with a feature mapping  $\phi(\mathbf{x})$  and then derive the corresponding kernel. For a one-dimensional input space, the kernel function is defined as

$$k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \phi(\mathbf{x}') = \sum_{i=1}^D \langle \phi_i(\mathbf{x}), \phi_i(\mathbf{x}') \rangle, \quad (3.15)$$

where  $\phi_i(\mathbf{x})$  are the basis functions of the feature map.

Alternatively, kernels can be constructed directly without explicitly defining a feature map. In this case, we must ensure that the chosen function is a valid kernel, meaning it corresponds to an inner product in some (possibly infinite-dimensional) feature space. The validity of a kernel function is guaranteed by Mercer's condition.

**Fact 3.1** (Mercer's condition). *Let  $\mathcal{X} \subset \mathbb{R}^d$  be a compact set and let  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  be a continuous and symmetric function. Then,  $k$  admits a uniformly convergent expansion of the form*

$$k(\mathbf{x}, \mathbf{x}') = \sum_{i=0}^{\infty} a_i \langle \phi_i(\mathbf{x}), \phi_i(\mathbf{x}') \rangle \quad (3.16)$$

with  $a_i > 0$  if and only if for any square-integrable function  $c$ , the following condition holds:

$$\int_{\mathcal{X}} \int_{\mathcal{X}} c(\mathbf{x})c(\mathbf{x}')k(\mathbf{x}, \mathbf{x}')d\mathbf{x}d\mathbf{x}' \geq 0. \quad (3.17)$$

Mercer's condition is crucial because it ensures that the optimization problem for algorithms like support vector machines (SVM) remains convex (Mohri, 2018), guaranteeing convergence to a global minimum. A condition equivalent to Mercer's condition (under the assumptions of the theorem) is that the kernel  $k(\cdot, \cdot)$  be positive definite symmetric. This property is more general since it does not require any assumption about  $\mathcal{X}$ .

**Definition 3.2** (Positive definite symmetric kernels). *A kernel  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is said to be positive definite symmetric (PDS) if for any  $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\} \subset \mathcal{X}$ , the matrix  $K = [k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})]_{ij} \in \mathbb{R}^{n \times n}$  is symmetric positive semidefinite (SPSD).*

In other words, a kernel matrix  $K$  associated with a PDS kernel function will always be SPSD, ensuring that the corresponding optimization problem remains well-behaved.

Below, we present several commonly used positive definite symmetric kernels.

**Example 3.3** (Polynomial kernels). *A polynomial kernel of degree  $m \in \mathbb{N}$  with a constant  $c > 0$  is defined as*

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}' + c)^m, \quad \forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d. \quad (3.18)$$

*This kernel maps the input space to a higher-dimensional space of dimension  $\binom{d+m}{m}$ . For instance, in a two-dimensional input space ( $d = 2$ ) and with  $m = 2$ , the kernel expands as follows*

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}_1 \mathbf{x}'_1 + \mathbf{x}_2 \mathbf{x}'_2 + c)^2 = \begin{bmatrix} \mathbf{x}_1^2 \\ \mathbf{x}_2^2 \\ \sqrt{2} \mathbf{x}_1 \mathbf{x}_2 \\ \sqrt{2c} \mathbf{x}_1 \\ \sqrt{2c} \mathbf{x}_2 \\ c \end{bmatrix} \cdot \begin{bmatrix} \mathbf{x}'_1^2 \\ \mathbf{x}'_2^2 \\ \sqrt{2} \mathbf{x}'_1 \mathbf{x}'_2 \\ \sqrt{2c} \mathbf{x}'_1 \\ \sqrt{2c} \mathbf{x}'_2 \\ c \end{bmatrix} \quad (3.19)$$

*In other words, this kernel corresponds to an inner product in a higher-dimensional space of dimension 6.*

Thus, the features corresponding to a second-degree polynomial are the original features ( $\mathbf{x}_1$  and  $\mathbf{x}_2$ ), as well as products of these features, and the constant feature. More generally, the features associated with a polynomial kernel of degree  $m$  are all the monomials of degree at most  $m$  based on the original features.

**Example 3.4** (Gaussian kernels). *The Gaussian kernel (or Radial Basis Function, RBF) is one of the most widely used kernels, defined as*

$$\forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d, \quad k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|}{2\sigma^2}\right), \quad (3.20)$$

where  $\sigma > 0$  controls the width of the Gaussian function.

Gaussian kernels are particularly effective in capturing complex non-linear patterns due to their infinite-dimensional feature space.

**Example 3.5** (Sigmoid kernels). *The sigmoid kernel over  $\mathbb{R}^d$  is defined as:*

$$k(\mathbf{x}, \mathbf{x}') = \tanh(a \cdot (\mathbf{x} \cdot \mathbf{x}') + b), \quad \forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d, \quad (3.21)$$

where  $a, b > 0$  are constants, and  $\tanh(c) = \frac{e^c - e^{-c}}{e^c + e^{-c}}$  is the hyperbolic tangent function, which squashes an arbitrary constant  $c \in \mathbb{R}$  to a value between  $-1$  and  $1$ .

This kernel is related to neural networks, as it resembles the activation function commonly used in multi-layer perceptrons as will be introduced in the next chapter. Using the sigmoid kernel with support vector machines results in a model similar to a simple neural network.

#### Remark

Support Vector Machines (SVMs) are a well-known algorithm that heavily relies on kernel methods and are primarily used for classification tasks. The objective of an SVM is to identify the optimal hyperplane that separates data points from different classes with the *maximum margin*. The margin is defined as the distance between the hyperplane and the closest data points from each class, known as support vectors. SVMs can be applied to both linear and non-linear

classification problems. For non-linear cases, kernel functions are employed to map the data into higher-dimensional spaces, enabling the separation of data that is not linearly separable in the original space. For a detailed introduction to SVMs, please refer to [Mohri \(2018\)](#).

## 3.2 Quantum Kernel Machines

### 3.2.1 Motivations for quantum kernel machines

To effectively introduce quantum kernel machines, it is essential to recognize the limitations of classical kernel machines. As discussed in Chapter 3.1, classical kernel machines rely on manually tailored feature mappings, such as polynomials or radial basis functions. However, these mappings may fail to capture the complex patterns behind the dataset. Quantum kernel machines emerge as a promising alternative, as they perform feature mapping using quantum circuits, enabling them to explore exponentially larger feature spaces that are otherwise infeasible for classical computation.

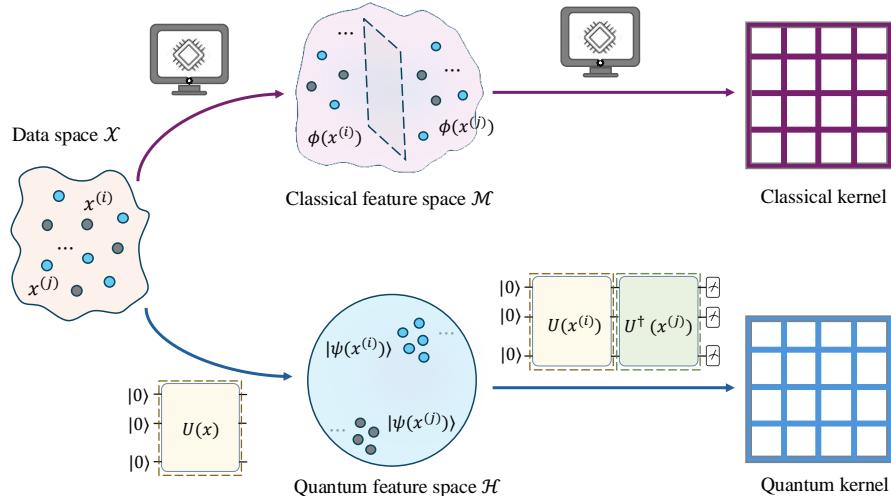


Figure 3.2: **The paradigm of classical and quantum kernels.** Both of the classical and quantum kernels embed the data points from data space  $\mathcal{X}$  into high-dimensional space, and then compute the kernel as the inner product of feature maps. The quantum kernel leverages quantum circuits to achieve this goal, as indicated by the blue color.

Another crucial characteristic of quantum kernels is that they can be effectively implemented on near-term quantum devices, making them a practical tool for exploring the utility of near-term quantum technologies.

### 3.2.2 Quantum feature maps and quantum kernel machines

The key difference between quantum kernel machines and classical kernel machines lies in how the feature mapping is performed. In the quantum context, a feature map refers to the injective encoding of classical data  $\mathbf{x} \in \mathbb{R}^d$  into a quantum state  $|\phi(\mathbf{x})\rangle = U(\mathbf{x})|\phi\rangle$  on an  $N$ -qubit quantum register, where  $U(\mathbf{x})$  refers to the physical operation or quantum circuit that depends on the data  $\mathbf{x}$ . This feature map is implemented on a quantum computer and produces quantum states, which are referred to as quantum feature maps.

**Definition 3.6** (Quantum feature map). *Given an  $N$ -qubit quantum system initialized in state  $|\psi\rangle$ , let  $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$  be classical data. The quantum feature map is defined as the mapping*

$$\begin{aligned}\phi : \mathcal{X} &\rightarrow \mathcal{F}, \\ \phi(\mathbf{x}) &= |\phi(\mathbf{x})\rangle \langle \phi(\mathbf{x})| = \rho(\mathbf{x}),\end{aligned}\tag{3.22}$$

where  $\mathcal{F}$  is the space of complex-valued  $2^N \times 2^N$  matrices equipped with the Hilbert-Schmidt inner product  $\langle \rho, \sigma \rangle = \text{Tr}(\rho\sigma)$  for  $\rho, \sigma \in \mathcal{F}$ . In addition, the state  $|\phi(\mathbf{x})\rangle$  can be implemented by applying a data-encoding quantum circuit  $U(\mathbf{x})$  introduced in Chapter 2.3.1 on an initial state  $|\psi\rangle$ , leading to the expression of  $|\phi(\mathbf{x})\rangle = U(\mathbf{x})|\psi\rangle$ .

Recall that one way of constructing kernels is adopting the inner product of the defined feature mappings. Using the Hilbert-Schmidt inner product from Definition 3.6, the quantum kernel is defined as follows.

**Definition 3.7** (Quantum Kernel). *Let  $\phi$  be a quantum feature map over the domain  $\mathcal{X}$ . The quantum kernel  $k_Q$  is the inner product between two quantum feature maps  $\rho(\mathbf{x})$  and  $\rho(\mathbf{x}')$  for data points  $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ ,*

$$\begin{aligned}k_Q : \mathcal{X} \times \mathcal{X} &\rightarrow \mathbb{R}, \\ k_Q(\mathbf{x}, \mathbf{x}') &= \text{Tr}(\rho(\mathbf{x})\rho(\mathbf{x}')) = |\langle \phi(\mathbf{x}) | \phi(\mathbf{x}') \rangle|^2.\end{aligned}\tag{3.23}$$

To justify the term ‘kernel’, we need to show that the quantum kernel is indeed a *positive definite function*. A quantum kernel can be expressed as the product of a complex-valued kernel  $\hat{k}_Q(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}) | \phi(\mathbf{x}') \rangle \in \mathbb{C}$  and

its complex conjugate  $\hat{k}_Q(\mathbf{x}, \mathbf{x}')^* = \langle \phi(\mathbf{x}) | \phi(\mathbf{x}') \rangle^* = \langle \phi(\mathbf{x}') | \phi(\mathbf{x}) \rangle$ . Since the product of two kernels is known to be a valid kernel, it suffices to show that  $\hat{k}_Q(\mathbf{x}, \mathbf{x}')$  is a valid complex-valued kernel and satisfies positive definiteness. For any  $\mathbf{x}^{(i)} \in \mathcal{X}$ ,  $i = 1, \dots, n$ , and any coefficients  $c_i \in \mathbb{C}$ , we have

$$\begin{aligned} \sum_{i,j} c_i c_j^* (\hat{k}_Q(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})) &= \sum_{i,j} c_i c_j^* \langle \phi(\mathbf{x}^{(i)}) | \phi(\mathbf{x}^{(j)}) \rangle \\ &= \left( \sum_i c_i \langle \phi(\mathbf{x}^{(i)}) | \right) \left( \sum_j c_j^* | \phi(\mathbf{x}^{(j)}) \rangle \right) \\ &= \left\| \sum_i c_i^* | \phi(\mathbf{x}^{(i)}) \rangle \right\|^2 \geq 0. \end{aligned} \quad (3.24)$$

This inequality confirms that  $\hat{k}_Q(\mathbf{x}, \mathbf{x}')$  satisfies Mercer's condition to be a valid kernel as illustrated in Eqn. (3.17). Therefore, the quantum kernel  $k_Q(\mathbf{x}, \mathbf{x}')$  is also a valid kernel.

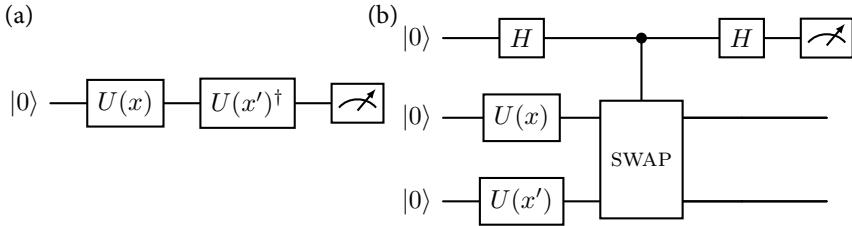


Figure 3.3: **Two methods for computing the inner product of the kernel.** (a) Loschmidt echo test. (b) Swap test.

The inner product between quantum states can be efficiently estimated on quantum computers using techniques such as Loschmidt echo test (Kusumoto et al., 2021) and SWAP test (Blank et al., 2020). Both methods correspond to distinct quantum circuit architectures, as illustrated in Figure 3.3.

One key merit of quantum kernels is that their derivation does not require the explicit representation of the quantum feature maps. Instead, it relies only on the construction of the associated quantum circuits. This aligns with the essence of kernel methods: while feature mappings can be computationally complex, the kernel function itself must remain efficient to evaluate.

Below, we outline the core steps for constructing a quantum kernel.

### General construction rules of quantum kernels

There are three steps to construct a quantum kernel:

1. Quantum feature map construction. Design a data-dependent quantum circuit  $U(\mathbf{x})$  to encode classical input data  $\mathbf{x}$  into the amplitudes or parameters of a quantum state  $|\phi(\mathbf{x})\rangle = U(\mathbf{x})|\psi\rangle$  where the initial state  $|\psi\rangle$  is typically  $|0\rangle^{\otimes N}$ .
2. Kernel evaluation. The quantum kernel is typically defined as the inner product of quantum states corresponding to two data points. Mathematically, this can be expressed as  $k_Q(\mathbf{x}, \mathbf{x}') = |\langle\phi(\mathbf{x})|\phi(\mathbf{x}')\rangle|^2$ .
3. Post-processing. After executing the quantum circuit for different input pairs, measure the output and calculate the kernel matrix. This matrix will then be used in machine learning models, such as SVM.

Below is a simple example of a quantum kernel with an angle encoding feature map introduced in Chapter 2.3.1.

**Example 3.8** (Single-qubit kernel). Consider an embedding that encodes a scalar input  $x \in \mathbb{R}$  into the quantum state of a single qubit. The embedding is implemented by the Pauli-X rotation gate  $RX(x) = e^{-ix\sigma_x/2}$ , where  $\sigma_x$  is the Pauli-X operator. The quantum feature map is then given by  $\phi : x \rightarrow \rho(x) = |\phi(x)\rangle\langle\psi(x)|$  with

$$\begin{aligned} |\phi(x)\rangle &= e^{-ix\sigma_x/2} |0\rangle \\ &= (\cos(x)\mathbb{I} - i \sin(x)\sigma_x) |0\rangle \\ &= \cos(x) |0\rangle - i \sin(x) |1\rangle, \end{aligned} \tag{3.25}$$

and hence the quantum kernel yields

$$k(x, x') = \left| \cos\left(\frac{x}{2}\right) \cos\left(\frac{x'}{2}\right) + \sin\left(\frac{x}{2}\right) \sin\left(\frac{x'}{2}\right) \right|^2 = \cos\left(\frac{x-x'}{2}\right)^2,$$

which is a translation invariant squared cosine kernel.

### 3.2.3 Relation between quantum and classical kernel machines

An intuitive way to understand the connections and differences between classical and quantum kernel machines is by comparing their fundamental components. As illustrated in Figure 3.2, both types of kernel machines involve four fundamental components: input, feature mapping, kernel matrix, and the computation process. Table 3.1 summarizes how these components are implemented in classical and quantum kernel machines.

Table 3.1: Comparison between classical and quantum kernel machines.

	Classical Kernel	Quantum Kernel
Input	classical data $\{\mathbf{x}^{(i)}\}_{i=1}^n \in \mathbb{R}^d$	classical data $\{\mathbf{x}^{(i)}\}_{i=1}^n \in \mathbb{R}^d$
Feature	Real vector $\phi(\mathbf{x}) \in \mathbb{R}^D$	Complex vector $ \phi(\mathbf{x})\rangle \in \mathbb{C}^{2^N}$
Kernel	$n$ -dimensional real matrix $K_C$	$n$ -dimensional real matrix $K_Q$
Computation	Digital logical circuits $\phi(\mathbf{x})$	Quantum circuits $U(\mathbf{x}) \psi\rangle$

The main distinctions between classical and quantum kernel machines lie in the computation processes for feature mapping and kernel matrix construction, as outlined below.

- *Classical versus quantum feature maps.* Quantum feature maps encode data into quantum states, resulting in exponentially large complex-valued vectors  $|\phi(\mathbf{x})\rangle \in \mathbb{C}^{2^N}$ , whereas classical feature maps operate in finite-dimensional real-valued spaces  $\phi(\mathbf{x}) \in \mathbb{R}^D$ . Although quantum feature maps can theoretically be simulated on classical computers by separating real and imaginary parts, this simulation becomes computationally infeasible as the number of qubits grows. Even feature maps generated by shallow quantum circuits are hard to simulate efficiently on classical hardware, demonstrating the inherent computational complexity of quantum feature maps.
- *Classical versus quantum kernels.* The kernel function is determined by the feature mapping, but its computational properties differ significantly between classical and quantum methods. A key merit of kernel methods is that they allow the use of complex feature maps while maintaining efficient kernel evaluations. Quantum kernels leverage quantum circuits to compute the inner product of quantum states, enabling the recognition of intricate patterns that classical kernels fail to capture. If a quantum kernel is computationally hard to eval-

ate classically, it offers a significant advantage by exploiting quantum computing's ability to process complex data representations efficiently.

#### Remark

The efficiency discussed here refers to the computational time within the respective classical or quantum frameworks. Specifically,

- *Classical Efficiency*: Determined by the depth of digital logical circuits used for feature mapping and kernel computation.
- *Quantum Efficiency*: Determined by the depth of quantum circuits required to achieve the same tasks.

A computation process is considered efficient if it can be completed in polynomial time relative to the problem size in its corresponding framework (classical or quantum).

#### 3.2.4 Concrete examples of quantum kernels

To better understand the concept of a quantum kernel, let's examine the kernels associated with common information encoding strategies used in quantum machine learning. It is important to note that some kernels cannot be efficiently computed on classical computers (Liu et al., 2021b). While such results are significant, the question of which quantum kernels are practically useful for real-world problems remains an open challenge.

In the following examples, we will first review the various encoding strategies introduced in Chapter 2.3.1, and then present the corresponding quantum kernels.

**Example 3.9** (Quantum kernel with basis encoding). *Given a classical binary vector  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_d) \in \{0, 1\}^d$ , the quantum feature mapping related to basis encoding refers to*

$$|\phi(\mathbf{x})\rangle = |\mathbf{x}_1, \dots, \mathbf{x}_d\rangle, \quad (3.26)$$

*and the induced quantum kernel yields*

$$k(\mathbf{x}, \mathbf{x}') = |\langle \phi(\mathbf{x}) | \phi(\mathbf{x}') \rangle|^2 = \delta_{\mathbf{x}\mathbf{x}'}, \quad (3.27)$$

*where  $\delta_{\mathbf{x}\mathbf{x}'} = 1$  if  $\mathbf{x} = \mathbf{x}'$  and otherwise 0.*

The basis encoding requires  $N = d$  qubits. This kernel function is a very strict similarity measure on input space, and arguably not the best choice of data encoding for quantum machine learning tasks.

**Example 3.10** (Quantum kernel with amplitude encoding). *Given a vector  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_d) \in \mathbb{R}^d$ , the quantum feature mapping related to amplitude encoding refers to*

$$|\phi(\mathbf{x})\rangle = \sum_{i=1}^d \frac{\mathbf{x}_i}{\|\mathbf{x}\|_2} |i\rangle, \quad (3.28)$$

where  $\|\mathbf{x}\|_2$  is the Euclidean norm. The related quantum kernel is given by

$$k(\mathbf{x}, \mathbf{x}') = |\langle \mathbf{x} | \mathbf{x}' \rangle|^2 = |\langle \mathbf{x}, \mathbf{x}' \rangle|^2. \quad (3.29)$$

The amplitude encoding requires  $N = \lceil \log 2(d) \rceil$  qubits. This encoding strategy leads to an identity feature map, which can be implemented by a non-trivial quantum circuit (for obvious reasons also known as “arbitrary state preparation”), which takes time  $\mathcal{O}(d)$  in the worst case. Besides, this quantum kernel does not add much power to a linear model in the original feature space, and it is more of interest for theoretical investigations that want to eliminate the effect of the feature map.

**Example 3.11** (Quantum kernel with angle encoding). *Given a vector  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_d) \in \mathbb{R}^d$ , the quantum feature mapping related to angle encoding refers to*

$$|\phi(\mathbf{x})\rangle = W_d e^{-i\mathbf{x}_d G_d} W_d \cdots W_2 e^{-i\mathbf{x}_2 G_2} W_1 |0\rangle^{\otimes d}, \quad (3.30)$$

where  $W_0, \dots, W_d$  are arbitrary unitary evolutions, and  $G_i$  is  $d_i \leq d$ -dimensional Hermitian operator called the generating Hamiltonian.

For a special case in which  $W_i = \mathbb{I}$  and  $G_i$  refers to the Pauli-X operators  $\sigma_x$  acting on the  $i$ -th qubit, the quantum feature mapping refers to

$$|\phi(\mathbf{x})\rangle = \bigotimes_{i=1}^d \exp\left(-i \frac{\mathbf{x}_i}{2} \sigma_x\right) |0\rangle^{\otimes d}, \quad (3.31)$$

and the related quantum kernel is given by

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= \prod_{i=1}^d \left| \sin(\mathbf{x}_i) \sin(\mathbf{x}'_i) + \cos(\mathbf{x}^{(i)}) \cos(\mathbf{x}'_i) \right|^2 \\ &= \prod_{i=1}^d |\cos(\mathbf{x}_i - \mathbf{x}'_i)|^2. \end{aligned} \quad (3.32)$$

The angle encoding requires  $d$ -qubit, mapping the classical data to  $2^d$ -dimensional Hilbert space. One merit of angle encoding is introducing non-linearity, which is crucial for transforming low-dimensional, non-linearly separable data into higher-dimensional, linearly separable representations—a property essential for effective kernel-based machine learning. Additionally, angle encoding is well-suited for implementation on modern devices featuring limited qubits and circuit depth, making it practical for exploring the practical utility of near-term quantum computers.

The quantum kernels related to different data encoding strategies have a resemblance to kernels from the classical machine learning literature. This means that sometimes up to an absolute square value, they can be identified with standard kernels such as the polynomial or Gaussian kernel. For the special case of angle encoding, the resemblance to classical kernels is because the employed quantum circuit does not employ any entangled quantum gates such that it can be simulated classically. We now discuss the general form of the quantum kernels induced by quantum feature maps from angle encoding in Eqn. (3.30). We focus on the simplified case that each input  $\mathbf{x}^{(i)}$  is only encoded once and that all the encoding Hamiltonians are the same, i.e.,  $G_1 = \dots = G_d = G$ .

**Theorem 3.12** (Fourier representation of the quantum kernel). *Let  $\mathcal{X} = \mathbb{R}^d$  and  $U(\mathbf{x})$  be a quantum circuit that encodes the data inputs  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_d) \in \mathcal{X}$  into a  $d$ -qubit quantum state  $|\phi(\mathbf{x})\rangle$  via gates of the form  $e^{-i\mathbf{x}_i G}$  for  $i = 1, \dots, d$ . Without loss of generality,  $G$  is assumed to be a  $m \leq 2^d$ -dimensional diagonal operator with spectrum  $\lambda_1, \dots, \lambda_m$ . Between such data-encoding gates, and before and after the entire encoding circuit, arbitrary unitary evolutions  $W_1, \dots, W_{d+1}$  can be applied, so that*

$$U(\mathbf{x}) = W_{d+1} e^{-i\mathbf{x}_d G_d} W_d \cdots W_2 e^{-i\mathbf{x}_1 G_1} W_1. \quad (3.33)$$

The quantum kernel  $k_Q(\mathbf{x}, \mathbf{x}')$  can be written as

$$k_Q(\mathbf{x}, \mathbf{x}') = \sum_{\mathbf{s}, \mathbf{t} \in \Omega} e^{i\mathbf{s}\mathbf{x}} e^{i\mathbf{t}\mathbf{x}'} c_{\mathbf{s}\mathbf{t}}, \quad (3.34)$$

where  $\Omega \subset \mathbb{R}^d$ , and  $c_{\mathbf{s}\mathbf{t}} \in \mathbb{C}$ . For every  $\mathbf{s}, \mathbf{t} \in \Omega$ , we have  $-\mathbf{s}, -\mathbf{t} \in \Omega$  and  $c_{\mathbf{s}\mathbf{t}} = c_{-\mathbf{s}-\mathbf{t}}^*$ , which guarantees that the quantum kernel is real-valued.

*Proof sketch of Theorem 3.12.* The assumption that the generator  $G$  is diagonal could be made without loss of generality because one can diagonalize Hermitian operators as  $G = V e^{-i\mathbf{x}_i \Sigma} V^\dagger$  with

$$e^{-i\mathbf{x}_i \Sigma} = \begin{pmatrix} e^{-i\mathbf{x}_i \lambda_1} & 0 & \cdots & 0 \\ 0 & e^{-i\mathbf{x}_i \lambda_2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & e^{-i\mathbf{x}_i \lambda_m} \end{pmatrix}, \quad (3.35)$$

where  $V^\dagger$  refers to the conjugate transpose of the matrix  $V$ ,  $\lambda_1, \dots, \lambda_m$  are the eigenvalues of  $G$ . Formally,  $V, V^\dagger$  can be absorbed into the arbitrary circuits  $W_{i+1}$  and  $W_i$  before and after the encoding gate. In this regard, the quantum kernel can be written down as the inner product between the feature state of the specific forms in Eqn. (3.30), i.e.,

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= |\langle \phi(\mathbf{x}') | \phi(\mathbf{x}) \rangle| \\ &= \left| \langle \mathbf{0} | W_1^\dagger (e^{-i\mathbf{x}'_1 \Sigma})^\dagger \cdots (e^{-i\mathbf{x}'_d \Sigma})^\dagger W_{d+1}^\dagger W_{d+1} e^{-i\mathbf{x}_d \Sigma} \cdots e^{-i\mathbf{x}_1 \Sigma} W_1 | \mathbf{0} \rangle \right|^2 \\ &= \left| \langle \mathbf{0} | W_1^\dagger (e^{-i\mathbf{x}'_1 \Sigma})^\dagger \cdots (e^{-i\mathbf{x}'_d \Sigma})^\dagger e^{-i\mathbf{x}_d \Sigma} \cdots e^{-i\mathbf{x}_1 \Sigma} W_1 | \mathbf{0} \rangle \right|^2 \\ &= \left| \sum_{j_1, \dots, j_d=1}^m \sum_{k_1, \dots, k_d=1}^m e^{-i(\lambda_{j_1} \mathbf{x}_1 - \lambda_{k_1} \mathbf{x}'_1 + \dots + \lambda_{j_d} \mathbf{x}_d - \lambda_{k_d} \mathbf{x}'_d)} \right. \\ &\quad \times \left. \left( W_1^{(1k_1)} \cdots W_d^{(k_{d-1}k_d)} \right)^* W_d^{(j_d j_{d-1})} \cdots W_1^{(j_1 1)} \right|^2 \\ &= \left| \sum_j \sum_{\mathbf{k}} e^{-i(\Lambda_j \mathbf{x} - \Lambda_{\mathbf{k}} \mathbf{x}')} (\omega_{\mathbf{k}})^* \omega_j \right|^2 \\ &= \sum_j \sum_{\mathbf{k}} \sum_{\mathbf{h}} \sum_l e^{-i(\Lambda_j - \Lambda_l) \mathbf{x}} e^{i(\Lambda_{\mathbf{k}} - \Lambda_{\mathbf{h}}) \mathbf{x}'} (\omega_{\mathbf{k}} \omega_{\mathbf{h}})^* \omega_j \omega_l, \end{aligned} \quad (3.36)$$

Here, the scalars  $W_i^{(ab)}$  refer to the element  $\langle a | W_i | b \rangle$  of the unitary operator  $W_i$ , the bold multi-index  $\mathbf{j}$  summarizes the set  $(j_1, \dots, j_d)$  and  $\Lambda_j$  is a vector containing the eigenvalues selected by the multi-index (and similarly for  $\mathbf{k}, \mathbf{h}, \mathbf{l}$ ).

We can now summarize all terms where  $\Lambda_j - \Lambda_l = \mathbf{s}$  and  $\Lambda_k - \Lambda_h = \mathbf{t}$ , in other words where the differences of eigenvalues amount to the same vectors  $\mathbf{s}, \mathbf{t}$ . Then

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= \sum_{\mathbf{s}, \mathbf{t} \in \Omega} e^{-i\mathbf{s}\mathbf{x}} e^{i\mathbf{t}\mathbf{x}'} \sum_{\mathbf{j}, \mathbf{l}: \Lambda_j - \Lambda_l = \mathbf{s}} \sum_{\mathbf{k}, \mathbf{h}: \Lambda_k - \Lambda_h = \mathbf{t}} \omega_j \omega_l (\omega_k \omega_h)^* \\ &= \sum_{\mathbf{s}, \mathbf{t} \in \Omega} e^{-i\mathbf{s}\mathbf{x}} e^{i\mathbf{t}\mathbf{x}'} c_{\mathbf{s}\mathbf{t}}. \end{aligned} \quad (3.37)$$

The frequency set  $\Omega$  contains all vectors  $\{\Lambda_j - \Lambda_l\}$  with  $\Lambda_j = (\lambda_{j_1}, \dots, \lambda_{j_d})$  and  $\lambda_{j_1}, \dots, \lambda_{j_d} \in [1, \dots, m]$ .  $\square$

We summarize the various strategies for the construction of quantum feature mappings and quantum kernels in Table 3.2.

Table 3.2: Overview of typical data encoding strategies and their quantum kernels. The input domain is assumed to be the  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_d) \in \mathcal{X} \subset \mathbb{R}^d$ .

Encoding	Qubits	Dimension	Quantum Kernel $k(\mathbf{x}, \mathbf{x}')$
Basis encoding	$d$	$2^d$	$\delta_{\mathbf{x}, \mathbf{x}'}$
Amplitude encoding	$\lceil \log_2(d) \rceil$	$d$	$ \mathbf{x}^\dagger \mathbf{x}' ^2$
Angel encoding	$d$	$2^d$	$\prod_{k=1}^d  \cos(\mathbf{x}_k - \mathbf{x}'_k) ^2$
General angle encoding	$d$	$2^d$	$\sum_{\mathbf{s}, \mathbf{t} \in \Omega} e^{-i\mathbf{s}\mathbf{x}} e^{i\mathbf{t}\mathbf{x}'} c_{\mathbf{s}\mathbf{t}}$

### Remark

After obtaining the quantum kernel matrix  $K_Q$  for a given training dataset  $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n$ , we can use it to perform regression or classification tasks in a manner similar to the classical kernel methods. In particular, as discussed in Chapter 3.1.2, consider the linear regression model given by

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^n \left( \mathbf{w}^\top \cdot \phi_Q(\mathbf{x}^{(i)}) - y^{(i)} \right)^2 + \frac{\lambda}{2} \mathbf{w}^\top \cdot \mathbf{w}, \quad (3.38)$$

where  $\phi_Q(\mathbf{x}^{(i)})$  denotes the quantum feature mapping related to the quantum kernel  $k_Q$ ,  $\mathbf{w}$  denotes the model parameters, and  $\lambda \geq 0$  is the regularization factor. Using the quantum kernel matrix, we can express the dual representation of the linear model given in Eqn. (3.14) to predict the output for a new input as

$$y(\mathbf{x}) = \mathbf{k}_Q(\mathbf{x})^\top \cdot (K_Q + \lambda \mathbb{I}_n)^{-1} \mathbf{y}, \quad (3.39)$$

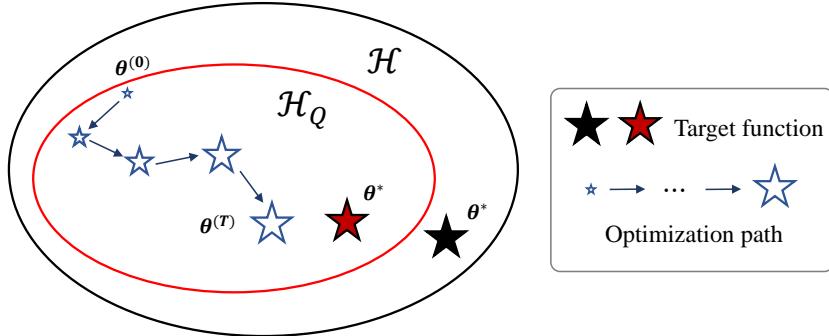


Figure 3.4: **The expressivity and generalization ability of quantum kernels.** Expressivity concerns the size of the hypothesis space  $\mathcal{H}_Q$  represented by quantum kernels, where  $\mathcal{H}$  refers to the whole hypothesis space. Generalization ability considers the learned hypothesis that could predict the unseen data accurately features a small distance with the target concept.

where  $\mathbf{y} = (y^{(1)}, \dots, y^{(n)})$  refers to the label vector,  $\mathbf{k}_Q(\mathbf{x})$  is a vector with elements  $\mathbf{k}_Q^{(i)}(\mathbf{x}) = k_Q(\mathbf{x}^{(i)}, \mathbf{x})$ .

### 3.3 Theoretical Foundations of Quantum Kernel Machines

In this section, we take a step further to explore the theoretical foundations of quantum kernels. Specifically, we focus on two crucial aspects: the *expressivity* and *generalization* properties of quantum kernel machines. As shown in Figure 3.4, these two aspects are essential for understanding the potential advantages of quantum kernels over classical learning approaches and their inherent limitations. For ease of understanding, this section emphasizes the fundamental concepts necessary for evaluating the power and limitation of quantum kernels instead of exhaustively reviewing all theoretical results.

The outline of this chapter is as follows. In Chapter 3.3.1, we will discuss the expressivity of quantum kernels, which refers to the diversity of feature spaces that quantum kernels can represent. The achieved insights will help identify tasks that are particularly well-suited for quantum kernels. Then, in Chapter 3.3.2, we will examine the potential advantage of quantum kernels

in terms of generalization error compared to all classical kernel machines. This analysis highlights their ability to accurately predict labels or values for unseen data.

### 3.3.1 Expressivity of quantum kernel machines

Quantum kernels, as discussed in Chapter 3.2, are constructed by explicitly defining quantum feature mappings. In this context, the expressivity of quantum kernel machines refers to the types of functions that quantum feature mappings can approximate and the kinds of correlations that quantum kernels can effectively model.

Following the conventions of Gil-Fuster et al. (2024), we demonstrate that *any kernel function can be approximated using finitely deep quantum circuits* by showing that the associated feature mapping can also be approximated using quantum circuits. This conclusion rests on two key theoretical foundations: Mercer's feature space construction and the universality of quantum circuits. Together, these principles establish the theoretical feasibility of realizing any kernel function as a quantum kernel.

It is important to note that if exact mathematical equality were required, Mercer's construction would demand an infinite-dimensional Hilbert space, which in turn would require quantum computers with infinitely many qubits—an impractical scenario. However, in practical applications, we are more interested in approximating functions to a certain level of precision rather than achieving exact evaluations. This perspective makes it feasible to implement the corresponding quantum feature mappings using a finite number of qubits. The following theorem confirms that any kernel function can be approximated as a quantum kernel to arbitrary precision with finite computational resources (We defer the proof details at the end of this subsection).

**Theorem 3.13** (Approximate universality of finite-dimensional quantum feature maps). *Let  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  be a kernel function. Then, for any  $\varepsilon \geq 0$  there exists  $N \in \mathbb{N}$  and a quantum feature mapping  $\rho_N$  onto the Hilbert space of quantum states of  $N$  qubits such that*

$$|k(\mathbf{x}, \mathbf{x}') - 2^N \text{Tr}(\rho_N(\mathbf{x})\rho_N(\mathbf{x}')') + 1| < \varepsilon \quad (3.40)$$

for almost all  $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ .

Theorem 3.13, instead of discussing the  $\varepsilon$ -approximation of quantum kernels in the form  $|k(\mathbf{x}, \mathbf{x}') - \text{Tr}(\rho_N(\mathbf{x})\rho_N(\mathbf{x}')')| < \varepsilon$ , introduces additional multiplicative and additive factors, expressed as  $|k(\mathbf{x}, \mathbf{x}') - 2^N \text{Tr}(\rho_N(\mathbf{x})\rho_N(\mathbf{x}')') + 1| < \varepsilon$ .

$1| < \varepsilon$ . These additional factors, explained below, do not impede the universality of the theorem.

Moreover, the statement that Eqn. (3.40) holds for almost all  $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$  stems from measure theory. It signifies that the inequality is valid “except on sets of measure zero,” or equivalently “with probability 1.” In other words, while adversarial instances of  $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$  may exist for which the inequality does not hold, such instances are so sparse that the probability of encountering them when sampling from the relevant probability distribution is zero.

Last, Theorem 3.13 establishes that any kernel function can be approximated as a quantum kernel up to a multiplicative and an additive factor using a finite number of qubits.

Before presenting the proof of this theorem, let us first introduce Algorithm 1, which maps classical vectors to quantum states. These quantum states can then be used to evaluate Euclidean inner products as quantum kernels. Then, we demonstrate Lemma 3.14 and Lemma 3.15, which separately formalize the correctness and runtime complexity of Algorithm 1, as well as establish the relationship between the Euclidean inner product of encoded real vectors and the Hilbert-Schmidt inner product of the corresponding quantum states.

---

**Algorithm 1** Classical to quantum embedding (C2QE)

---

**Input:** a unit vector with 1-norm  $\mathbf{r} \in \ell_1^d$ .

**Output:** Quantum state  $\rho_{\mathbf{r}} \propto \mathbb{I} + \sum_{i=1}^d \mathbf{r}_i P_i$ . ▷ See Lemma 3.14.

- 1: Set  $N = \lceil \log_4(d+1) \rceil$ .
  - 2: Pad  $\mathbf{r}$  with zeros until its length is  $4^N - 1$ .
  - 3: Draw  $i \in \{1, \dots, 4^N - 1\}$  with probability  $|\mathbf{r}_i|$ .
  - 4: Prepare  $\rho_i = \frac{1}{2^N}(\mathbb{I} + \text{sign}(\mathbf{r}_i)P_i)$ .
  - 5: **return**  $\rho_i$ .
- 

The output of Algorithm 1,  $\frac{1}{2^N}(\mathbb{I} \pm P)$ , is a single (pure) eigenstate of a Pauli operator  $P$  with eigenvalue  $\pm 1$ . However, since Line 3 involves sampling an index  $i \in \{1, \dots, 4^N - 1\}$ , Algorithm 1 is inherently random, and the resulting quantum state is a classical mixture of pure states.

**Lemma 3.14** (Correctness and runtime of Algorithm 1). *Let  $\mathbf{r} \in \ell_1^d \subset \mathbb{R}^d$  be a unit vector with respect to the 1-norm, i.e.,  $\|\mathbf{r}\|_1 = 1$ . Take  $N = \lceil \log_4(d+1) \rceil$  and pad  $\mathbf{r}$  with zeros until its length is  $4^N - 1$ . Let  $(P_i)_{i=1}^{4^N-1}$  be the set of all Pauli matrices on  $N$  qubits, excluding the identity. Then*

*Algorithm 1* prepares the following state as a classical mixture

$$\begin{aligned} \rho(\cdot) : \ell_1^d &\rightarrow \text{Herm}(2^N), \\ \mathbf{r} \mapsto \rho_{\mathbf{r}} &= \frac{\mathbb{I} + \sum_{i=1}^{4^N-1} \mathbf{r}_i P_i}{2^N}. \end{aligned} \quad (3.41)$$

The total runtime complexity  $t$  of Algorithm 1 fulfills  $t \leq \mathcal{O}(\text{poly}(d))$ .

*Proof of Lemma 3.14.* The proof begins by expanding the state as follows

$$\frac{\mathbb{I} + \sum_{i=1}^{4^N-1} \mathbf{r}_i P_i}{2^N} = \frac{1}{2^N} \left( \sum_{i=1}^{4^N-1} |\mathbf{r}_i| \mathbb{I} + \sum_{i=1}^{4^N-1} \mathbf{r}_i P_i \right), \quad (3.42)$$

where the first equality follows that  $\|\mathbf{r}\|_1 = 1$  and  $\mathbf{r} \in \mathbb{R}^{4^N-1}$ . Rewriting the above equation using  $\text{sign}(\mathbf{r}_i)$  yields

$$\begin{aligned} \frac{\mathbb{I} + \sum_{i=1}^{4^N-1} \mathbf{r}_i P_i}{2^N} &= \frac{1}{2^N} \left( \sum_{i=1}^{4^N-1} |\mathbf{r}_i| \mathbb{I} + \sum_{i=1}^{4^N-1} |\mathbf{r}_i| \text{sign}(\mathbf{r}_i) P_i \right) \\ &= \frac{1}{2^N} \sum_{i=1}^{4^N-1} |\mathbf{r}_i| (\mathbb{I} + \text{sign}(\mathbf{r}_i) P_i) \succeq 0. \end{aligned} \quad (3.43)$$

Here, it is used that  $\sum_i |\mathbf{r}_i| = \|\mathbf{r}\|_1 = 1$  and  $\mathbb{I} \pm P_i \geq 0$  for all Pauli operators  $P_i$ . Notice that efficiently preparing  $\mathbb{I} + P_i$  can be achieved by rotating each qubit's  $|0\rangle$  basis state to the corresponding Pauli basis and flipping the necessary qubits individually. Since this state is a convex combination of quantum states, it can be efficiently prepared by mixing, when the number of terms is polynomial.  $\square$

**Lemma 3.15** (Euclidean inner products). *Let  $\mathbf{r}, \mathbf{r}' \in \mathbb{R}^d$  be unit vectors with respect to the 1-norm, i.e.,  $\|\mathbf{r}\|_1 = \|\mathbf{r}'\|_1 = 1$ . For the states  $\rho_{\mathbf{r}}, \rho'_{\mathbf{r}'}$  produced in Algorithm 1, the following identity holds*

$$\langle \mathbf{r}, \mathbf{r}' \rangle = 2^N \text{Tr}(\rho_{\mathbf{r}} \rho'_{\mathbf{r}'}) - 1. \quad (3.44)$$

*Proof of Lemma 3.15.* The proof utilizes the following principles: (1) The trace is linear, and the trace of a tensor product equals the product of traces. (2) All Pauli words are traceless except for the identity, and each Pauli operator is its own inverse. Hence, the product of distinct Pauli operators is also traceless.

Expanding the trace of  $\rho_{\mathbf{r}}\rho_{\mathbf{r}'}$ , we have

$$\text{Tr}(\rho_{\mathbf{r}}\rho_{\mathbf{r}'}) = \text{Tr}\left(\frac{1}{4^N}\left(\mathbb{I} + \sum_{j=1}^{4^N-1} \mathbf{r}_j P_j\right)\left(\mathbb{I} + \sum_{k=1}^{4^N-1} \mathbf{r}'_k P_k\right)\right), \quad (3.45)$$

which could be simplified as

$$\text{Tr}(\rho_{\mathbf{r}}\rho_{\mathbf{r}'}) = \frac{1}{4^N} \text{Tr}\left(\mathbb{I} + \sum_{j=1}^{4^N-1} \mathbf{r}_j \mathbf{r}'_j P_j^2 + \sum_{k \neq j}^{4^N-1} \mathbf{r}_j \mathbf{r}'_k P_j P_k\right). \quad (3.46)$$

Using the properties of Pauli operators, the trace becomes

$$\text{Tr}(\rho_{\mathbf{r}}\rho_{\mathbf{r}'}) = \frac{1}{4^N} \left( \text{Tr}(\mathbb{I}) + \text{Tr}\left(\sum_{j=1}^{4^N-1} \mathbf{r}_j \mathbf{r}'_j \mathbb{I}\right) \right) = \frac{1 + \langle \mathbf{r}, \mathbf{r}' \rangle}{2^N}. \quad (3.47)$$

This completes the proof.  $\square$

Lemma 3.15 clarifies the origin of the extra factors in Theorem 3.13. In particular, the  $2^N$  multiplicative factor is unproblematic, as  $N \leq \mathcal{O}(\log(d))$  and the methods are designed to scale polynomially with  $d$ . Moreover, the quantum state  $\rho_{\mathbf{r}}$  is generally mixed but can be efficiently prepared. The mapping is injective but not surjective.

With these results in place, we now present the proof of Theorem 3.13.

*Proof of Theorem 3.13.* The proof follows from a corollary of Mercer's theorem and the universality of quantum computing. First, by a direct corollary of the Mercer's Theorem (i.e., Fact 3.1) which states that an arbitrary kernel  $k$  admits a uniformly convergent expansion of the form in Eqn. (3.16), it is ensured that there exists a finite-dimensional feature map  $\Phi_m : \mathcal{X} \rightarrow \mathbb{R}^m$  such that

$$|k(\mathbf{x}, \mathbf{x}') - \langle \Phi_m(\mathbf{x}), \Phi_m(\mathbf{x}') \rangle| < \varepsilon. \quad (3.48)$$

Without loss of generality, it is assumed that  $\|\Phi_m(\mathbf{x})\| = 1$  for all  $\mathbf{x} \in \mathcal{X}$ . The quantum state  $\rho_{\Phi_m}$  can then be prepared, which requires  $\lceil \log_4(m+1) \rceil$  qubits. By preparing two such states—one for  $\Phi_m(\mathbf{x})$  and one for  $\Phi_m(\mathbf{x}')$ —their inner product can be computed as the Hilbert-Schmidt inner product of the quantum states, as shown in Lemma 3.15. This leads to

$$\langle \Phi_m(\mathbf{x}), \Phi_m(\mathbf{x}') \rangle = 2^N \text{Tr}(\rho_{\Phi_m(\mathbf{x})}\rho_{\Phi_m(\mathbf{x}')}^*) - 1. \quad (3.49)$$

For reference, it is noted that  $\text{Tr}(\rho_{\Phi_m(\mathbf{x})}\rho_{\Phi_m(\mathbf{x}')})$  can be computed using the SWAP test to an additive precision determined by the number of measurement shots. This allows us to approximate the result efficiently to any desired polynomial additive precision. Consequently, it follows that

$$|k(\mathbf{x}, \mathbf{x}') - 2^N \text{Tr}(\rho_{\Phi_m(\mathbf{x})}\rho_{\Phi_m(\mathbf{x}')}) + 1| < \varepsilon, \quad (3.50)$$

for almost all  $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ . This completes the proof.  $\square$

We remark that Theorem 3.13 does not aim to demonstrate any quantum advantage but rather establishes the ultimate expressivity of quantum kernels. The theorem guarantees the existence of a quantum kernel using a finite number of qubits but does not address how quickly the number of required qubits grows with increasing computational complexity of the kernel function  $k$  or with decreasing approximation error  $\varepsilon > 0$ . The number of qubits  $N$  will depend on certain properties of the kernel  $k$  and the approximation error  $\varepsilon$ . For instance, if the required number of qubits scales exponentially with these parameters, Theorem 3.13 would have limited practical utility. Similarly, the time required to find such a quantum kernel approximation-independent of the memory and runtime requirements for preparing the feature vectors and computing their inner product—must also be considered.

#### Remark

Although Theorem 3.13 establishes that all kernel functions can be realized as quantum kernels, there may still exist kernel functions that cannot be *realized efficiently* as quantum kernels. This observation requires us to identify quantum kernels that can be computed efficiently on quantum computers, i.e., in polynomial time.

### 3.3.2 Generalization of quantum kernel machines

Generalization, which quantifies the ability of learning models (both classical and quantum) to predict unseen data, is a critical metric for evaluating the quality of a learning model. Due to its importance, here we analyze the potential advantages of quantum kernels in terms of the generalization ability.

For comprehensive, in this section, we first elucidate the generalization error bounds for general kernel machines, establishing a unified framework for a fair comparison between quantum kernels and classical kernels. Subsequently, we introduce a geometry metric to assess the potential quantum

advantage of quantum kernels with respect to generalization error for a fixed amount of training data.

### Generalization error bound for kernel machines

We begin by reviewing the optimal learning models based on the specified kernel machines which could be either classical or quantum, as discussed in Chapter 3.1.2. Suppose we have obtained  $n$  training examples  $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n$  with  $\mathbf{x}^{(i)} \in \mathbb{R}^d$  and  $y^{(i)} = f(\mathbf{x}^{(i)}) \in \mathbb{R}$ , where  $f$  is the target function. After training on this data, there exists a machine learning algorithm that outputs  $h(\mathbf{x}) = \mathbf{w}^\dagger \phi(\mathbf{x})$ , where  $\phi(\mathbf{x}) \in \mathbb{C}^D$  refers to the hidden feature map corresponding the classical/quantum kernel function  $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = K_{ij} = \phi(\mathbf{x}^{(i)}) \cdot \phi(\mathbf{x}^{(j)})$ . More precisely, considering the mean square error as the loss function for such a task, we have

$$\mathcal{L}(\mathbf{w}, \mathbf{x}) = \lambda \mathbf{w}^\dagger \mathbf{w} + \sum_{i=1}^n (\mathbf{w}^\dagger \phi(\mathbf{x}^{(i)}) - y^{(i)})^2, \quad (3.51)$$

where  $\lambda \geq 0$  is the regularization parameters for avoiding over-fitting.

Then the optimal parameters for optimizing this loss function refer to

$$\mathbf{w}^* = \arg \min_{\mathbf{w} \in \Theta} \mathcal{L}(\mathbf{w}, \mathbf{x}). \quad (3.52)$$

As discussed in Chapter 3.1.2, the optimal solution  $\mathbf{w}^*$  in Eqn. (3.52) has the explicit form of

$$\mathbf{w}^* = \Phi^\dagger (K + \lambda \mathbb{I}_n)^{-1} \mathbf{y} = \sum_{i=1}^n \sum_{j=1}^n \phi(\mathbf{x}^{(i)}) ((K + \lambda \mathbb{I}_n)^{-1})_{ij} y^{(j)}, \quad (3.53)$$

where  $\mathbf{y} = [y^{(1)}, \dots, y^{(n)}]^\top$  refers to the vector of labels and  $K \in \mathbb{R}^{n \times n}$  is the kernel matrix, and the second equality follows that  $\Phi = [\phi(\mathbf{x}^{(1)}), \dots, \phi(\mathbf{x}^{(n)})]^\dagger$ . Moreover, the norm of the optimal parameters has a simple form for the case of  $\lambda \rightarrow 0$ , i.e.,

$$\|\mathbf{w}^*\|_2^2 = \mathbf{y}^\top K^{-1} \mathbf{y}. \quad (3.54)$$

We now expose the prediction error of these learning models, i.e.,

$$\epsilon_{\mathbf{w}^*}(\mathbf{x}) = |f(\mathbf{x}) - (\mathbf{w}^*)^\dagger \phi(\mathbf{x})|, \quad (3.55)$$

which is uniquely determined by the kernel matrix  $K$  and the hyper-parameter  $\lambda$  as shown in Eqn. (3.53). In particular, we will focus on discussing the upper bound on the expected prediction error, which is the sum of training error and generalization error.

Prediction, training, and generalization error

In the context of learning theory, the upper bound of the expected prediction error defined in Eqn. (3.55) (a.k.a, expected risk) is achieved by separately analyzing the upper bounds of the training error (a.k.a, empirical risk) and the generalization error, i.e.,

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \epsilon_{\mathbf{w}^*}(\mathbf{x}) = \underbrace{\frac{1}{n} \sum_{i=1}^n \epsilon_{\mathbf{w}^*}(\mathbf{x}^{(i)})}_{\text{Training error}} + \underbrace{\mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \epsilon_{\mathbf{w}^*}(\mathbf{x}) - \frac{1}{n} \sum_{i=1}^n \epsilon_{\mathbf{w}^*}(\mathbf{x}^{(i)})}_{\text{Generalization error}}. \quad (3.56)$$

This decomposition stems from the fact that data distribution  $\mathcal{D}$  is inaccessible in most scenarios.

We now will separately give a rough derivation of the upper bound of training error and generalization error, which present the necessary steps for the derivation for a clear exposition and omit the specific details that could be found in Huang et al. (2021a).

- *Training error.* Employing the convexity of function and Jensen's inequality, the training error yields

$$\frac{1}{n} \sum_{i=1}^n \epsilon_{\mathbf{w}^*}(\mathbf{x}^{(i)}) \leq \sqrt{\frac{1}{n} \sum_{i=1}^n ((\mathbf{w}^*)^\dagger \phi(\mathbf{x}^{(i)}) - y^{(i)})^2}. \quad (3.57)$$

Moreover, combining with the expression for the optimal  $\mathbf{w}^*$  given in Eqn. (3.53), we can obtain the upper bound of training error in terms of the kernel matrix  $K$  and hyper-parameter  $\lambda$ , i.e.,

$$\frac{1}{n} \sum_{i=1}^n \epsilon_{\mathbf{w}^*}(\mathbf{x}^{(i)}) \leq \sqrt{\frac{\lambda^2 \mathbf{y}^\top (K + \lambda \mathbb{I}_n)^{-2} \mathbf{y}}{n}}. \quad (3.58)$$

We can see that when  $\lambda = 0$  and  $K$  are invertible, the training error is zero. However, the hyper-parameter is usually set as  $\lambda > 0$  in practice.

- *Generalization error.* The derivation of generalization error is more complicated than training error, which involves a basic theorem in statistic and learning theory as presented below.

**Fact 3.16** (Theorem 3.3, Mohri (2018)). *Let  $\mathcal{G}$  be a family of function mappings from a set  $\mathcal{Z}$  to  $[0, 1]$ . Then for any  $\delta > 0$ , with probability at least  $1 - \delta$*

over identical and independent draw of  $n$  samples from  $\mathcal{Z} : \mathbf{z}^{(1)}, \dots, \mathbf{z}^{(n)}$ , we have for all  $g \in \mathcal{G}$ ,

$$\mathbb{E}_{\mathbf{z}} g(\mathbf{z}) \leq \frac{1}{n} \sum_{i=1}^n g(\mathbf{z}^{(i)}) + 2\mathbb{E}_{\sigma} \left[ \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(\mathbf{z}^{(i)}) \right] + 3\sqrt{\frac{\log(2/\delta)}{2n}}, \quad (3.59)$$

where  $\sigma_1, \dots, \sigma_n$  are in independent and uniform random variables over  $\{1, -1\}$ .

For kernel functions defined in Eqn. (3.55), the set  $\mathcal{Z}$  refers to the space of input vector with  $\mathbf{z}^{(i)} = \mathbf{x}^{(i)}$  drawn from some input distribution. Each function  $g$  would be equal to  $\epsilon_{\mathbf{w}}/\alpha$  for some  $\mathbf{w}$ , where  $\epsilon_{\mathbf{w}}$  is defined in Eqn. (3.55) and  $\alpha$  is a normalization factor such that the range of  $\epsilon_{\mathbf{w}}/\alpha$  is  $[0, 1]$ . Without loss of generality, we assume that  $\alpha = 1$ . For any specific parameter  $\mathbf{w}$ , consider the special case of  $\mathcal{G}$  with setting  $\mathcal{G}_{\mathbf{w}} = \{\epsilon_{\mathbf{v}} \mid \forall \|\mathbf{v}\| \leq \|\mathbf{w}\|\}$ . Then we have the upper bound of generalization error for the optimal parameter,

$$\begin{aligned} & \mathbb{E}_{\mathbf{x}} \epsilon_{\mathbf{w}^*}(\mathbf{x}) - \frac{1}{n} \sum_{i=1}^n \epsilon_{\mathbf{w}^*}(\mathbf{x}^{(i)}) \\ & \leq 2\mathbb{E}_{\sigma} \left[ \sup_{\|\mathbf{v}\| \leq \|\mathbf{w}^*\|} \frac{1}{n} \sum_{i=1}^n \sigma_i \epsilon_{\mathbf{v}}(\mathbf{x}^{(i)}) \right] + 3\sqrt{\frac{\log(2\|\mathbf{w}^*\|/\delta)}{2n}}. \end{aligned} \quad (3.60)$$

Moreover, applying Talagrand's contraction lemma (Mohri, 2018) to the first term on the right-hand side, we have

$$\begin{aligned} \mathbb{E}_{\sigma} \left[ \sup_{\|\mathbf{v}\| \leq \|\mathbf{w}^*\|} \frac{1}{n} \sum_{i=1}^n \sigma_i \epsilon_{\mathbf{v}}(\mathbf{x}^{(i)}) \right] & \leq \mathbb{E}_{\sigma} \left[ \sup_{\|\mathbf{v}\| \leq \|\mathbf{w}^*\|} \frac{1}{n} \sum_{i=1}^n \sigma_i (\mathbf{w}^*)^\dagger \phi(\mathbf{x}^{(i)}) \right] \\ & \leq \sqrt{\frac{\|\mathbf{w}^*\|^2}{n}}, \end{aligned} \quad (3.61)$$

where the first inequality follows that  $\epsilon_{\mathbf{v}}(\mathbf{x}^{(i)})$  is Lipschitz continuous with respect to  $(\mathbf{w}^*)^\dagger \cdot \phi(\mathbf{x}^{(i)})$  with Lipschitz constant 1, the second inequality follows direct algebra operation. For the detailed simplification processes, refer to Lemma 1 of Huang et al. (2021a).

In conjunction with Eqn. (3.60), Eqn. (3.61), and the expression of the optimal parameter  $\mathbf{w}^*$  given in Eqn. (3.53), we can reach the final upper

bound of generalization error in terms of the kernel matrix, i.e.,

$$\begin{aligned} & \mathbb{E}_{\mathbf{x}} \epsilon_{\mathbf{w}^*}(\mathbf{x}) - \frac{1}{n} \sum_{i=1}^n \epsilon_{\mathbf{w}^*}(\mathbf{x}^{(i)}) \\ & \leq 5 \cdot \frac{\mathbf{y}^\top (K + \lambda \mathbb{I}_n)^{-1} K (K + \lambda \mathbb{I}_n)^{-1} \mathbf{y}}{n} + 3 \sqrt{\frac{\log(2/\delta)}{2n}}. \end{aligned} \quad (3.62)$$

For the case of  $\lambda = 0$ , the first term in the generalization error bound has a simple form of  $5 \cdot \mathbf{y}^\top K^{-1} \mathbf{y}/n$ .

By obtaining the upper bound of training and generalization error, we can directly get the prediction error of the learning model for a specific kernel matrix. We summarize these three errors below.

#### Remark

The upper bound of the prediction error for kernel methods defined in Eqn. (3.55) refers to

$$\begin{aligned} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \epsilon_{\mathbf{w}^*}(\mathbf{x}) & \leq \mathcal{O} \left( \underbrace{\sqrt{\frac{\lambda^2 \mathbf{y}^\top (K + \lambda \mathbb{I}_n)^{-2} \mathbf{y}}{n}}}_{\text{Training error}} + \right. \\ & \quad \left. \underbrace{\sqrt{\frac{\mathbf{y}^\top (K + \lambda \mathbb{I}_n)^{-1} K (K + \lambda \mathbb{I}_n)^{-1} \mathbf{y}}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}} \right), \end{aligned} \quad (3.63)$$

Generalization error

where  $K$  is a specific kernel related to the learning models,  $\mathbf{y} = [y^{(1)}, \dots, y^{(n)}]$  refers to the label vector of  $n$  training data. For the special case of  $\lambda = 0$ , the training error is zero, and the prediction error reduce to the generalization error with a simple form

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \epsilon_{\mathbf{w}^*}(\mathbf{x}) \leq \mathcal{O} \left( \sqrt{\frac{\mathbf{y}^\top K^{-1} \mathbf{y}}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right). \quad (3.64)$$

We remark that the derived upper bound of the prediction error applies to both classical and quantum kernels, as we have not imposed any restrictions on the kernel matrix  $K$  during the derivation.

#### Quantum kernels with prediction advantages

Using the above theoretical results of generalization error for general kernel machines, we now elucidate how to access the potential quantum advantage

of quantum kernels. For a clear understanding, we focus on the case of  $\lambda = 0$  in which the prediction error bound has a simple form of  $\mathcal{O}(\sqrt{\mathbf{y}^\top K^{-1} \mathbf{y}}/n + \sqrt{\log(1/\delta)/n})$  as shown in Eqn. (3.64). In particular, this bound has a key dependence on two quantities, namely (1) the size of training data  $n$ ; (2) the kernel-dependent term  $\mathbf{y}^\top K^{-1} \mathbf{y}$ , which we denote as

$$s_K(\mathbf{y}) = \mathbf{y}^\top K^{-1} \mathbf{y}, \quad (3.65)$$

in the following discussion for simplification.

The dependence on  $n$  reflects the role of data to improve prediction performance. On the other hand, the quantity  $s_K(\mathbf{y})$  is equal to the model complexity of the trained function  $h(\mathbf{x}) = (\mathbf{w}^*)^\dagger \cdot \phi(\mathbf{x})$ , where  $s_K(\mathbf{y}) = \|\mathbf{w}^*\|^2 = (\mathbf{w}^*)^\dagger \cdot \mathbf{w}^*$  after training. A smaller value of  $s_K(\mathbf{y})$  implies better generalization to new data  $\mathbf{x}$  sampled from the distribution  $\mathcal{D}$ . Intuitively,  $s_K(\mathbf{y})$  measures whether the closeness between  $\mathbf{x}^{(i)}$  and  $\mathbf{x}^{(j)}$  defined by the kernel function  $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$  matches well with the closeness of the labels  $y^{(i)}$  and  $y^{(j)}$ , recalling that a larger kernel value indicates two points are closer.

Based on the above discussion, we are now in the position to analyze the potential advantage of quantum kernel machines. Given a set of training data  $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n$ , let  $\mathcal{Q}$  and  $\mathcal{C}$  be the class of **Quantum** and **Classical** kernels respectively that can be efficiently evaluated on quantum and classical computers for any given  $\mathbf{x}$ . In order to formally evaluate the potential for quantum prediction advantage generally, one must take the quantum kernel  $K_Q \in \mathcal{Q}$  to satisfy the following two conditions:

- $K_Q$  is hard to compute classically for any given  $\mathbf{x}$ .
- According to Eqn. (3.65), the quantity  $s_Q(\mathbf{y})$  related to the quantum kernel  $K_Q$  must be the minimal over all efficient classical models, namely  $s_Q(\mathbf{y}) \leq s_C(\mathbf{y})$  for any  $K_C \in \mathcal{C}$  with  $s_C(\mathbf{y})$  being the  $K_C$  related quantity.

From the second condition, we can see that the potential advantage for the quantum kernel  $K_Q$  to predict better than a classical kernel  $K_C$  depends on the largest possible separation between  $s_Q(\mathbf{y})$  and  $s_C(\mathbf{y})$  for a dataset. Huang et al. (2021a) define a geometry metric, namely **asymmetric geometric difference**, to characterize this separation for a fixed training dataset, which is given by

$$g_{CQ} = g(K_C || K_Q) = \sqrt{\left\| \sqrt{K_Q}(K_C^{-1})\sqrt{K_Q} \right\|_\infty}, \quad (3.66)$$

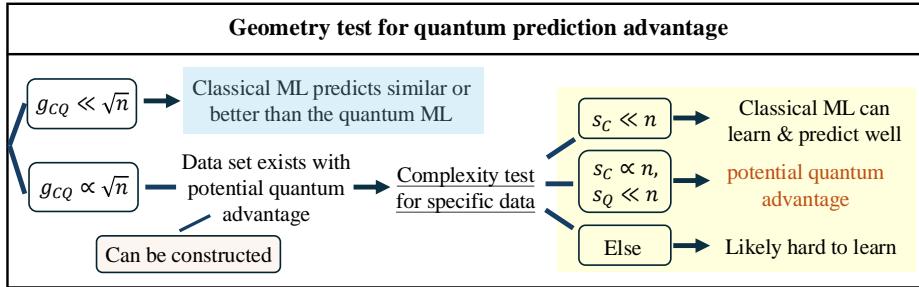


Figure 3.5: A flowchart for understanding the potential for quantum prediction advantage (Adapted from [Huang et al. \(2021a\)](#)).

where  $\|\cdot\|_\infty$  is the spectral norm of the resulting matrix and we assume  $\text{Tr}(K_Q) = \text{Tr}(K_C) = n$ . The geometric difference  $g(K_C||K_Q)$  can be computed on a classical computer by performing a singular value decomposition of the  $n \times n$  matrices  $K_C$  and  $K_Q$  in time at most order  $n^3$ .

Figure 3.5 presents a detailed flowchart for evaluating the potential quantum prediction advantage using the defined geometric difference  $g_{CQ}$  in a machine learning task. The input consists of  $n$  data samples, along with both quantum and classical methods, each associated with its respective kernel. The tests are conducted as a function of  $n$  to highlight the role of data size in determining the potential for a prediction advantage.

First, the geometric quantity  $g_{CQ}$  is evaluated, which quantifies the potential for a separation between quantum and classical predictions, without yet considering the actual function to be learned. Specifically, a large value of  $g_{CQ} \propto \sqrt{n}$  suggests the possibility of a quantum prediction advantage. If the test is passed, an adversarial dataset that saturates this limit can be constructed. In particular, there exists a dataset with  $s_C = g_{CQ}^2 s_Q$ , where the quantum model exhibits superior prediction performance, as will be described in the subsequent context.

Subsequently, to incorporate the provided data, a label-specific test can be performed using the model complexities  $s_C$  and  $s_Q$ . For quantum kernels and classical learning models, when  $s_Q \ll n$  and  $s_C \propto n$ , a prediction advantage for quantum models is possible, as supported by the generalization bound in Eqn. (3.64). In contrast, if  $g_{CQ}$  is small such as  $g_{CQ} \ll \sqrt{n}$ , the classical learning model will likely have a similar or better model complexity  $s_C(\mathbf{y})$  compared to the quantum model. In this case, the classical model's prediction performance will be competitive or superior, and the classical model would likely be preferred.

### Construction of dataset with maximal quantum advantage

We now elucidate how to explicitly construct such a data set to enable the maximal separation between the model complexity of quantum kernels and classical kernels, as indicated by the geometry test in Figure 3.5. To separate between quantum and classical models related to kernel matrix  $K_Q$  and  $K_C$ , we consider that the ratio between  $s_C$  and  $s_Q$  is as large as possible for a particular choice of targets  $y^{(1)}, \dots, y^{(n)}$ . This could be achieved by solving the optimization problem

$$\min_{\mathbf{y} \in \mathbb{R}^n} \frac{s_C}{s_Q} = \min_{\mathbf{y} \in \mathbb{R}^n} \frac{\mathbf{y}^\top K_C^{-1} \mathbf{y}}{\mathbf{y}^\top K_Q^{-1} \mathbf{y}}, \quad (3.67)$$

which has an exact solution given by a generalized eigenvalue problem. The solution is given by  $\mathbf{y} = \sqrt{K_Q} \mathbf{v}$ , where  $\mathbf{v}$  is the eigenvector of  $\sqrt{K_Q} K_C^{-1} \sqrt{K_Q}$  corresponding to the eigenvalue  $g^2 = \|\sqrt{K_Q} K_C^{-1} \sqrt{K_Q}\|_\infty$ . This guarantees that  $s_C = g^2 s_Q$ , and note that by definition of  $g$ ,  $s_C \leq g^2 s_Q$ . Hence this dataset fully utilized the geometric difference between the quantum and classical space. Finally, we can turn this dataset, which maps input  $\mathbf{x}$  to a real value  $y_Q$ , into a classification task by replacing  $y_Q$  with  $+1$  if  $y_Q > \text{median}(y^{(1)}, \dots, y^{(n)})$  and  $-1$  if  $y_Q \leq \text{median}(y^{(1)}, \dots, y^{(n)})$ . The constructed dataset will yield the largest separation between quantum and classical models from a learning theoretic sense, as the model complexity fully saturates the geometric difference. If there is no quantum advantage in this dataset, there will likely be none.

## 3.4 Code Demonstration

In this section, we explore the practical implementation of a quantum kernel. Before diving into concrete code examples, we discuss an efficient strategy for estimating the quantum kernel in practice.

As explained in Chapter 3.2, one method for estimating the quantum kernel is the SWAP test, which is resource-intensive. Alternatively, we can encode the classical data vector  $\mathbf{x}$  using a unitary operation  $U(\mathbf{x})$  and apply the inverse embedding of  $\mathbf{x}'$  using  $U(\mathbf{x}')^\dagger$  on the same qubits. The quantum kernel  $k_Q(\mathbf{x}, \mathbf{x}')$  is then estimated by measuring the expectation of the projector  $O = (|0\rangle\langle 0|)^{\otimes N}$  on the zero state.

The complete quantum circuit architecture for this process is illustrated

in Figure 3.2. Mathematically, the process is expressed as:

$$\begin{aligned}
 & \langle 0^{\otimes N} | U(\mathbf{x}') U(\mathbf{x})^\dagger O U(\mathbf{x}')^\dagger U(\mathbf{x}) | 0^{\otimes N} \rangle \\
 &= \langle 0^{\otimes N} | U(\mathbf{x}') U(\mathbf{x})^\dagger | 0 \rangle^{\otimes N} \langle 0 |^{\otimes N} U(\mathbf{x}')^\dagger U(\mathbf{x}) | 0^{\otimes N} \rangle \\
 &= \left| \langle 0^{\otimes N} | U(\mathbf{x}')^\dagger U(\mathbf{x}) | 0^{\otimes N} \rangle \right|^2 \\
 &= \left| \langle \phi(\mathbf{x}) | \phi(\mathbf{x}') \rangle \right|^2 \\
 &= k_Q(\mathbf{x}, \mathbf{x}'). \tag{3.68}
 \end{aligned}$$

This approach allows the quantum kernel estimation to use the same number of qubits required for the quantum feature mapping of the classical vector  $\mathbf{x}$ .

Next, we provide an example demonstrating the workflow of applying quantum kernels for classification tasks on the MNIST dataset, with step-by-step code implementation.

### 3.4.1 Classification on MNIST dataset

We train a Support Vector Machine (SVM) classifier associated with a quantum kernel on the MNIST dataset. The pipeline involves the following steps.

Step 1 Load and preprocess the dataset.

Step 2 Define the quantum feature mapping.

Step 3 Construct the quantum kernel.

Step 4 Train and evaluate the SVM classifier.

We begin by importing the required libraries.

```

1 import pennylane as qml
2 from sklearn.datasets import fetch_openml
3 from sklearn.decomposition import PCA
4 from sklearn.model_selection import train_test_split
5 from sklearn.preprocessing import StandardScaler
6 from sklearn.svm import SVC
7 from sklearn.metrics import accuracy_score
8 import numpy as np

```

**Step 1: Dataset preparation.** We focus on the digits 3 and 6 in the MNIST dataset, forming a binary classification problem. Principal Component Analysis (PCA) (Abdi and Williams, 2010) is applied to reduce the

feature dimension of the images, minimizing the number of required qubits for encoding. The compressed features are normalized to align with the periodicity of the quantum feature mapping.

```

1 def load_mnist(n_qubit):
2     # Load MNIST dataset from OpenML
3     mnist = fetch_openml('mnist_784', version=1)
4     X, y = mnist.data, mnist.target
5
6     # Filter out the digits 3 and 6
7     mask = (y == '3') | (y == '6')
8     X_filtered = X[mask]
9     y_filtered = y[mask]
10
11    # Convert labels to binary (0 for digit 3 and 1 for
12      # digit 6)
13    y_filtered = np.where(y_filtered == '3', 0, 1)
14
15    # Apply PCA to reduce feature dimension
16    pca = PCA(n_components=n_qubit)
17    X_reduced = pca.fit_transform(X_filtered)
18
19    # Normalize the input features
20    scaler = StandardScaler().fit(X_reduced)
21    X_scaled = scaler.transform(X_reduced)
22
23    # Split into training and testing sets
24    X_train, X_test, y_train, y_test = train_test_split(
25        X_scaled, y_filtered, test_size=0.2, random_state
26        =42)
27    return X_train, X_test, y_train, y_test
28
29 n_qubit = 8
30 X_train, X_test, y_train, y_test = load_mnist(n_qubit)
```

To better understand the structure of the dataset, we visualize the training data using t-distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton, 2008). The following code generates the visualization:

```

1 def visualize_dataset(X, labels):
2     import matplotlib.pyplot as plt
3     from sklearn.manifold import TSNE
4
5     tsne = TSNE(n_components=2, random_state=42,
```

```
6     perplexity=30)
7     label2name = {
8         0: '3',
9         1: '6'
10    }
11    mnist_tsne = tsne.fit_transform(X)
12    for label in np.unique(labels):
13        indices = labels == label
14        plt.scatter(mnist_tsne[indices, 0], mnist_tsne[
15            indices, 1], cmap='coolwarm', s=20, label=f'16        Number{label2name[label]}')
17
18    # Add labels and legend
19    plt.title("t-SNE Visualization of Two Classes (3 and 6)")
20    plt.xlabel("t-SNE Dimension 1")
21    plt.ylabel("t-SNE Dimension 2")
22    plt.legend()
23
24    plt.tight_layout()
25    plt.show()
26
27 visualize_dataset(X_train, y_train)
```

The resulting t-SNE visualization is shown in Figure 3.6.

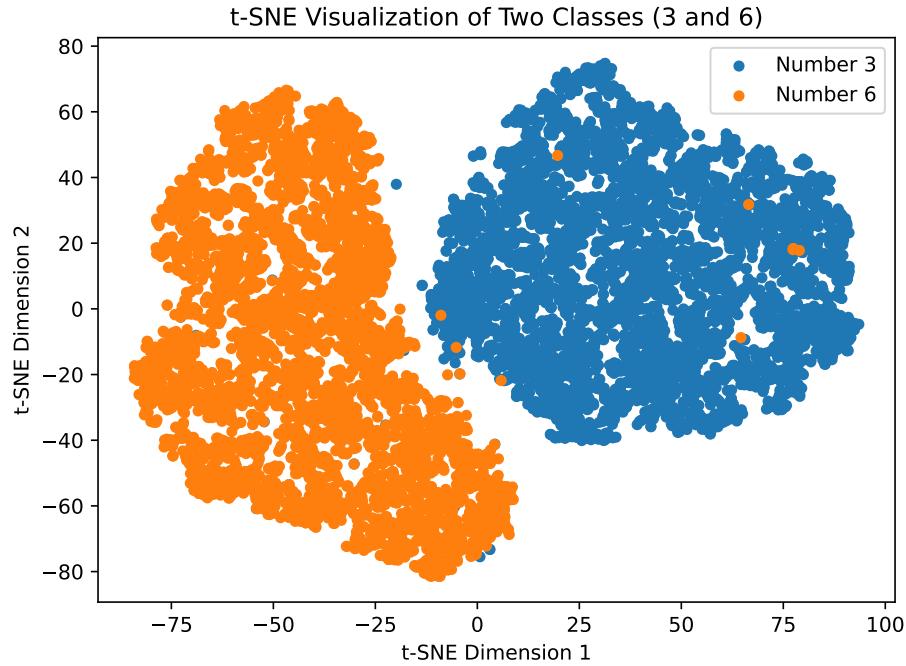


Figure 3.6: T-SNE visualization of MNIST dataset of two classes ‘3’ and ‘6’.

**Steps 2&3: Define quantum feature mapping and building quantum kernel.** We use angle embedding as the quantum feature mapping method. The quantum kernel is implemented as follows.

```

1 dev = qml.device('default.qubit', wires=n_qubit)
2
3 @qml.qnode(dev)
4 def kernel(x1, x2, n_qubit):
5     qml.AngleEmbedding(x1, wires=range(n_qubit))
6     qml.adjoint(qml.AngleEmbedding)(x2, wires=range(
7         n_qubit))
8     return qml.expval(qml.Projector([0]*n_qubit, wires=
9         range(n_qubit)))

```

Using the quantum kernel, we construct the kernel matrix by computing the kernel values for all pairs of samples:

```

1 def kernel_mat(A, B):
2     mat = []

```

```

3   for a in A:
4     row = []
5     for b in B:
6       row.append(kernel(a, b, n_qubit))
7     mat.append(row)
8   return np.array(mat)

```

Next, we visualize the quantum kernel matrix to gain insight into its structure.

```

1 def visualize_kernel(X, y, n_sample):
2   X_vis = []
3   for label in np.unique(y):
4     index = y == label
5     X_vis.append(X[index][:n_sample])
6
7   X_vis = np.concatenate(X_vis, axis=0)
8   n_sample_per_class = len(X_vis) // 2
9
10  sim_mat = kernel_mat(X_vis, X_vis)
11  np.save('code/chapter_4_kernel/sim_mat.npy', sim_mat)
12
13  import matplotlib.pyplot as plt
14  plt.imshow(sim_mat, cmap='viridis', interpolation='nearest')
15
16  # Add color bar to show the scale
17  plt.colorbar(label='Similarity')
18
19  plt.axhline(n_sample_per_class - 0.5, color='red',
20             linewidth=1.5) # Horizontal line
21  plt.axvline(n_sample_per_class - 0.5, color='red',
22             linewidth=1.5) # Vertical line
23
24  xticks = yticks = np.arange(0, len(X_vis))
25  xtick_labels = [f"3-{i+1}" if i < n_sample_per_class
26                  else f"6-{i+1-n_sample_per_class}" for i in range
27                  (len(X_vis))]
28  ytick_labels = xtick_labels
29
30  plt.xticks(xticks, labels=xtick_labels, rotation=90,
31             fontsize=8)
32  plt.yticks(yticks, labels=ytick_labels, fontsize=8)
33
34  # Title and axis labels

```

```

30     plt.title("Quantum Kernel Matrix")
31     plt.xlabel("Sample Index")
32     plt.ylabel("Sample Index")
33
34
35     plt.tight_layout()
36     plt.show()
37
38 visualize_kernel(X_train, y_train, 10)

```

The resulting kernel matrix is shown in Figure 3.7.

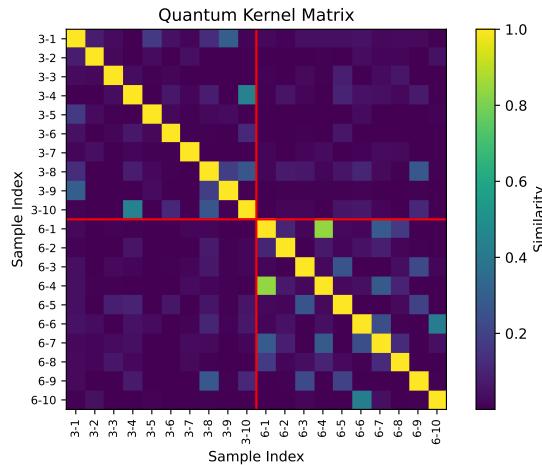


Figure 3.7: Visualization of quantum kernel matrix on 20 samples, equally drawn from two classes ‘3’ and ‘6’ in MNIST dataset.

From the visualization, we observe a clear block structure:

- Most of the elements in the top-left and bottom-right blocks, where samples belong to the same class, show higher similarity values.
- Most of the elements in the top-right and bottom-left blocks, where samples belong to different classes, exhibit lower similarity values.

This indicates that it may be possible to distinguish the two classes by setting a similarity threshold.

**Step 4: Training SVM.** To achieve higher classification accuracy, we build, train, and evaluate the SVM classifier with the quantum kernel matrix:

```

1  svm = SVC(kernel=kernel_mat)
2  svm.fit(X_train, y_train)
3  pred = svm.predict(X_test)
4  print("Accuracy:", accuracy_score(y_test, pred))

```

To further analyze how the performance of the SVM with a quantum kernel depends on the size of the training dataset, we vary the number of training samples from 10 to 100 in increments of 10. For each configuration, we record the corresponding classification accuracy on the test data.

```

1  svm = SVC(kernel='precomputed')
2  n_sample_max = 100
3  X_train_sample = []
4  y_train_sample = []
5  for label in np.unique(y_train):
6      index = y_train == label
7      X_train_sample.append(X_train[index][:n_sample_max])
8      y_train_sample.append(y_train[index][:n_sample_max])
9  X_train_sample = np.concatenate(X_train_sample, axis=0)
10 y_train_sample = np.concatenate(y_train_sample, axis=0)
11 kernel_mat_train = kernel_mat(X_train_sample,
12     X_train_sample)
12 kernel_mat_test = kernel_mat(X_test, X_train_sample)
13
14 accuracy = []
15 n_samples = []
16 for n_sample in range(10, n_sample_max+10, 10):
17     class1_indices = np.arange(n_sample)
18     class2_indices = np.arange(n_sample_max, n_sample_max
19         +n_sample)
20     selected_indices = np.concatenate([class1_indices,
21         class2_indices])
22
23     svm.fit(kernel_mat_train[np.ix_(selected_indices,
24         selected_indices)], np.concatenate([
25             y_train_sample[:n_sample], y_train_sample[
26                 n_sample_max:n_sample_max+n_sample]]))
27     pred = svm.predict(np.concatenate([kernel_mat_test[:, :
28         n_sample], kernel_mat_test[:, n_sample_max:
29             n_sample_max+n_sample]], axis=1))
30     accuracy.append(accuracy_score(y_test, pred))
31     n_samples.append(n_sample)
32
33 plt.plot(n_sample, accuracy, marker='o')

```

```

27 plt.title('Classification Accuracy vs. #Training Samples')
28     )
29 plt.xlabel('#Training Samples')
30 plt.xticks(n_sample, n_sample)
31 plt.ylabel('Accuracy')
32 plt.grid()
33 plt.tight_layout()
34 plt.show()

```

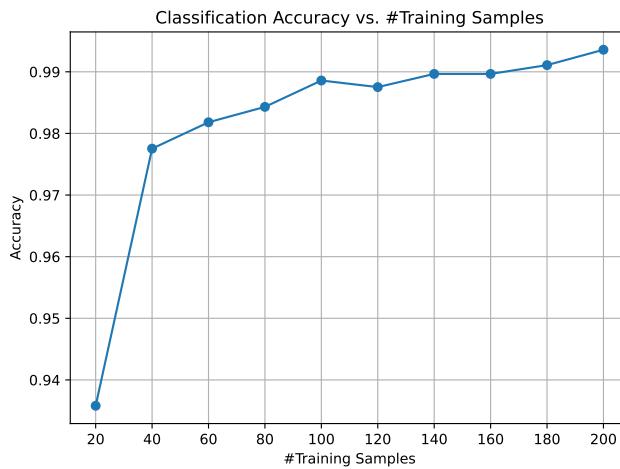


Figure 3.8: The classification accuracy on test data as a function of the number of training samples.

### 3.5 Bibliographic Remarks

The foundational concept of using quantum computers to evaluate kernel functions, namely the concept of quantum kernels, was first explored by Schuld et al. (2017), who highlighted the fundamental differences between quantum kernels and quantum support vector machines. Building on this, Havlíček et al. (2019) and Schuld and Killoran (2019) established a connection between quantum kernels and parameterized quantum circuits (PQCs), demonstrating their practical implementation. These works emphasized the parallels between quantum feature maps and the classical kernel trick. Since then, a large number of studies delved into figuring out the potential of quantum kernels for solving practical real-world problems.

The recent advancements in quantum kernel machines can be roughly

categorized into three key areas: *kernel design*, *theoretical findings*, and *applications*. Specifically, the advances in kernel design focus on addressing challenges such as vanishing similarity and kernel concentration by exploring innovative frameworks. Theoretical studies delve into the limitations and capabilities of quantum kernels, examining factors such as generalization error bounds, noise resilience, and their capacity to demonstrate quantum advantage. Finally, applications of quantum kernels showcase their potential across diverse domains. In the rest of this section, we separately review the existing developments within each of these three branches.

### 3.5.1 Quantum kernel design

A crucial research line in this field focuses on constructing *trainable quantum kernels* to maximize performance for specific datasets and problem domains. In particular, traditional quantum kernels, with fixed data embedding schemes, are limited to specific feature representation spaces and often fail to capture the complex and diverse patterns inherent in real-world data. To address this limitation, [Lloyd et al. \(2020\)](#) explored the construction of task-specific quantum feature maps using measurement theory. Building on this idea, [Hubregtsen et al. \(2022\)](#) introduced a method to optimize quantum feature maps variationally within quantum kernels, linking this approach to the concept of data re-uploading techniques ([Pérez-Salinas et al., 2020](#); [Schuld et al., 2021](#)). Additionally, [Vedaie et al. \(2020\)](#) and [Lei et al. \(2024\)](#) proposed leveraging multiple kernel learning and architecture search to construct quantum kernels, respectively. Last, [Glick et al. \(2024\)](#) proposed the covariant quantum kernels for efficiently solving the problems with group structure.

An orthogonal research direction is addressing the issue of exponential kernel concentration, also known as vanishing similarity, in quantum kernels ([Thanasilp et al., 2022](#)). Specifically, quantum kernels, which are defined as the inner product between quantum feature mappings, often suffer from the phenomenon of vanishing similarity. This issue was first highlighted by [Huang et al. \(2021a\)](#), who found that quantum feature mappings are typically “far” from one another in high-dimensional feature spaces, leading to vanishing similarity and, consequently, poor generalization performance.

To mitigate this issue, [Huang et al. \(2021a\)](#) introduced projected quantum kernels, which store feature vectors in classical memory and evaluate a Gaussian kernel. This approach replaces the reliance on the inner product of quantum states, as seen in traditional quantum embedding kernels. Moreover, [Suzuki et al. \(2022\)](#) proposed the anti-symmetric logarithmic derivative

quantum Fisher kernel, which avoids the exponential kernel concentration problem by encoding geometric information of the input data. Beyond developing new types of quantum kernels, [Shaydulin and Wild \(2022\)](#) and [Canatar et al. \(2022\)](#) explored strategies to mitigate the exponential concentration issue for quantum embedding kernels by scaling input data with carefully chosen hyperparameters. This approach clusters the data-encoded quantum states closer together in feature space, reducing the risk of vanishing similarity at the cost of slightly lowering expressivity.

### 3.5.2 Theoretical studies of quantum kernels

The theoretical studies of quantum kernels aim to rigorously understand their performance potential and limitations under realistic conditions, enabling the design of more effective, robust, and generalizable quantum kernel methods. Prior literature in this context focuses on exploring two aspects of quantum kernels, namely, expressivity and generalization ability.

#### Expressivity of quantum kernels

The expressivity of quantum kernels refers to their capacity to capture complex data relationships and represent intricate patterns in the feature space. A common approach to studying this is through the analysis of the reproducing kernel Hilbert space (RKHS), which provides insights into the underlying feature representations of quantum kernels.

[Schuld \(2021\)](#) rigorously analyzed the RKHS of embedding-based quantum kernels and established the universality approximation theorem, demonstrating that quantum kernels can approximate a wide class of functions. Building on this, [Jerbi et al. \(2023\)](#) extended the analysis by investigating parameterized quantum embedding kernels, introducing a data-reuploading structure and proving a corresponding universality approximation theorem. These results underscore the expressive power of quantum kernels in representing complex data structures.

Despite these advances, the studies on expressive power and universality approximation often overlook the efficiency of constructing quantum kernels. Specifically, if the computational cost of constructing a universal quantum kernel is comparable to that of classical methods, the practical advantages of quantum kernels become questionable.

To narrow this gap, [Gil-Fuster et al. \(2024\)](#) examined the expressive power of efficient quantum kernels that can be implemented on quantum computers within polynomial time. Their work provides a detailed analysis

of the types of kernels that are achievable with a polynomial number of qubits and within polynomial time, offering insights into the feasibility and practical utility of quantum kernels in real-world scenarios.

However, alongside the exploration of expressive power, a significant challenge known as exponential kernel concentration has been identified. [Thanasilp et al. \(2022\)](#) identified four key factors contributing to this issue: high expressivity of data embeddings, global measurements, entanglement, and noise. To address this limitation, substantial research has focused on designing advanced quantum kernels to mitigate exponential kernel concentration, as discussed in Chapter 3.5.1.

### Generalization of quantum kernels

The generalization ability of a learning model—its capacity to perform well on unseen data—is a critical factor in evaluating its effectiveness. In this context, a considerable body of research has investigated the generalization ability of quantum kernels. [Huang et al. \(2021a\)](#) established a data-dependent generalization error bound for quantum kernels and demonstrated that, for certain types of data (such as those generated by quantum circuits), quantum kernels can achieve a generalization advantage over classical learning models.

In addition, [Wang et al. \(2021a\)](#) explored the generalization performance of quantum kernels in the noisy scenario, where practical limitations such as finite measurements and quantum noise are taken into account. Their work rigorously showed that the generalization performance of quantum kernels could be significantly degraded in scenarios involving large training datasets, limited measurement repetitions, or high levels of system noise. To address these challenges, they proposed an effective method based on indefinite kernel learning to help preserve generalization performance under such constraints.

Beyond quantum data, [Liu et al. \(2021b\)](#) examined the generalization error of quantum kernels using artificial classical datasets, such as those based on the discrete logarithm problem. Their results demonstrated that quantum kernels could achieve accurate predictions in polynomial time for such problems, whereas classical learning models require exponential time, highlighting the potential computational advantages of quantum kernels.

Despite these promising results, [Kübler et al. \(2021\)](#) studied the generalization ability of quantum kernels from the perspective of inductive bias. They argued that quantum kernels, lacking inductive bias, often fail to outperform classical models in practical scenarios. This underscores the im-

portance of carefully designing embeddings and aligning kernels to achieve meaningful and practical quantum advantages.

### Provable advantages of quantum kernels

The potential for quantum kernels to demonstrate quantum advantage has been a central focus of research. For instance, Huang et al. (2021a) provided evidence of generalization advantages for quantum kernels on quantum data. Similarly, Liu et al. (2021b) presented a rigorous framework showing that quantum kernels can efficiently solve problems like the discrete logarithm problem, which is believed to be intractable for classical computers under standard cryptographic assumptions. Moreover, Sweke et al. (2021) demonstrated quantum advantage in distribution learning tasks, offering some of the earliest theoretical evidence of quantum advantage in machine learning.

However, many of these tasks are artificial, designed specifically to showcase quantum advantages. This raises the question of how these theoretical benefits can be translated to real-world applications. In this regard, the next significant challenge is to demonstrate that quantum models can consistently outperform classical models in solving practical, real-world problems.

### 3.5.3 Applications of quantum kernels

Motivated by the potential of quantum kernels to recognize complex data patterns, numerous studies have explored their practical applications across diverse fields, including classification, drug discovery, anomaly detection, and financial modeling.

For instance, Beaulieu et al. (2022) investigate the use of quantum kernels for image classification, specifically in identifying real-world manufacturing defects. Similarly, Rodriguez-Grasa et al. (2024) apply quantum kernels to satellite image classification, a task of particular importance in the earth observation industry. In the field of quantum physics, Sancho-Lorente et al. (2022) and Wu et al. (2023) leverage quantum kernels to recognize phases of quantum matter, where quantum kernels outperform classical learning models in solving certain problems. In drug discovery, Batra et al. (2021) explore the potential of quantum kernels to accelerate and improve the identification of promising compounds.

Quantum kernels have also been explored in anomaly detection. Liu and Rebentrost (2018) demonstrate their superior performance over classical methods in detecting anomalies within quantum data. Furthermore, Grossi et al. (2022) employ quantum kernel methods for fraud classification

tasks, showing improvements when benchmarked against classical methods. Miyabe et al. (2023) expand their application to the financial domain by proposing a quantum multiple-kernel learning methodology. This approach broadens the scope of quantum kernels to include credit scoring and directional forecasting of asset price movements, highlighting their potential utility in financial services.

Despite the promise shown in these applications, the realization of quantum advantage in practical tasks remains an ongoing area of research, with current efforts directed toward identifying real-world problems where quantum kernels outperform classical alternatives.



## Chapter 4

# Quantum Neural Networks

Classical neural networks (LeCun et al., 2015) are the foundation of modern artificial intelligence technologies and have achieved widespread success in fields such as computer vision (Voulodimos et al., 2018) and natural language processing (Otter et al., 2020). However, despite these achievements, classical neural networks face significant challenges, including excessively large model sizes and the corresponding high computational costs (Hoffmann et al., 2022), especially in terms of energy consumption (de Vries, 2023). These limitations result from their dependence on classical computational resources, which are becoming increasingly unsustainable as models grow in complexity.

Quantum neural networks (QNNs) (Jeswal and Chakraverty, 2019) offer a promising solution by enhancing neural networks with the computational potential of quantum circuits (Liu et al., 2024a). In QNNs, classical input data is encoded into quantum states, and quantum gates with trainable parameters process these states in ways that classical systems cannot easily replicate. This computational regime leverages quantum mechanics to explore new forms of pattern recognition and problem-solving that go beyond classical methods. Thus, QNNs have the potential to outperform classical neural networks in specific learning tasks (Huang et al., 2022), where the advantages in processing and learning can be explored.

Despite these promising features, there are challenges in realizing the full potential of QNNs, such as quantum noise (Peters et al., 2021) and the requirement for scalable quantum hardware (Acharya et al., 2024). Nonetheless, ongoing advancements in quantum hardware and algorithm design promise QNNs to address the inefficiencies of classical models, especially in areas such as quantum many-body physics (Gardas et al., 2018) and

quantum chemistry (Cao et al., 2019).

In this chapter, to provide a systematic overview, we begin by outlining the structure and function of classical neural networks in Chapter 4.1, before transitioning to fault-tolerant and near-term quantum neural networks in Chapters 4.2 and 4.3, respectively. We also discuss the theoretical foundations of QNNs in Chapter 4.4, focusing on trainability, expressivity, and generalization capabilities. Finally, we provide illustrative code implementations of QNNs using the wine (Dua and Graff, 2017) and MNIST datasets (LeCun et al., 1998) in Chapter 4.5.

## 4.1 Classical Neural Networks

Neural networks (McCulloch and Pitts, 1943; Gardner and Dorling, 1998; Vaswani, 2017) are computer models inspired by the structure of the human brain, designed to process and analyze complex patterns in data. Originally developed from the concept of neurons connected by weighted pathways, neural networks have become one of the most powerful tools in artificial intelligence (LeCun et al., 2015). Each neuron processes its inputs by applying weights and non-linear activations, producing an output that feeds into the next layer of neurons. This structure enables neural networks to learn complex functions during training (Hornik, 1993). For example, given a dataset of images and their labels, a neural network can learn to classify categories, such as distinguishing between cats and dogs, by adjusting its parameters during training. Guided by optimization algorithms such as gradient descent (Amari, 1993), the learning process allows the network to gradually reduce the error between the predicted and actual outputs, allowing it to learn the best parameters for the given task.

After nearly a century of exploration, neural networks have undergone remarkable advancements in both architectures and capabilities. The simplest model, the perceptron (McCulloch and Pitts, 1943), laid the foundation by showing how neural networks could learn to separate linearly classifiable categories. Building on this, deeper and more complex networks—such as multilayer perceptrons (MLPs) (Gardner and Dorling, 1998) and transformers (Vaswani, 2017)—have enabled breakthroughs in tasks such as autonomous driving and content generation.

### 4.1.1 Perceptron

The perceptron model, first introduced by (McCulloch and Pitts, 1943), is widely regarded as a foundational structure in artificial neural networks, in-

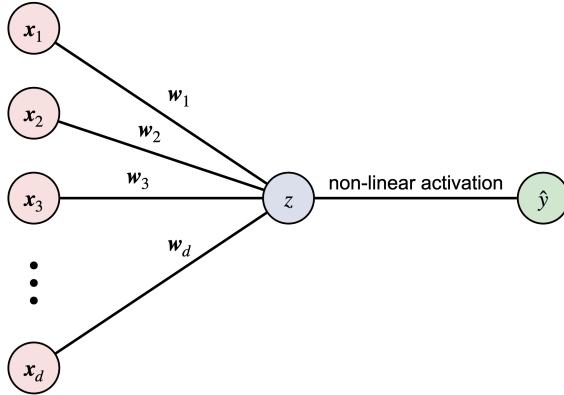


Figure 4.1: **Illustration of the perceptron.** Inputs are processed with a weighted linear combination followed by a non-linear activation to produce the output.

spiring architectures ranging from convolutional neural networks (CNNs) (Le-Cun et al., 1989) and residual neural networks (ResNets) (He et al., 2016) to transformers (Vaswani, 2017). Due to its fundamental role, we next introduce the mechanism of single-layer perceptrons.

A single-layer perceptron comprises three fundamental components: *input neurons*, *a weighted layer*, and *an output neuron* as illustrated in Figure 4.1. Given a  $d$ -dimensional input vector  $\mathbf{x} \in \mathbb{R}^d$ , the input layer consists of  $d$  neurons, each representing the feature  $x_i$  for  $\forall i \in [d]$ . This input is processed through a weighted summation, i.e.,

$$z = \mathbf{w}^\top \mathbf{x}, \quad (4.1)$$

where  $\mathbf{w}^\top$  is the transpose of the weight vector and  $z$  is the output of the weighted layer. A non-linear activation function is then applied to produce the output neuron  $\hat{y}$ . For the standard perceptron model, the sign function is typically used as the activation function:

$$\hat{y} = f(z) = \begin{cases} 1, & \text{if } z \geq 0, \\ -1, & \text{if } z < 0. \end{cases} \quad (4.2)$$

The perceptron learns from input data by iteratively adjusting its trainable parameters  $\mathbf{w}$ . In particular, let  $\mathcal{D} = \{(\mathbf{x}^{(a)}, y^{(a)})\}_{a=1}^n$  be the training dataset, where  $\mathbf{x}^{(a)}$  represents the input features of the  $a$ -th example, and

$y^{(a)} \in \{-1, 1\}$  denotes the corresponding label. When the perceptron outputs a prediction  $\hat{y}^{(s)}$ , the parameters are updated accordingly, i.e.,

$$\mathbf{w} \leftarrow \mathbf{w} + (y^{(s)} - \hat{y}^{(s)}) \mathbf{x}^{(s)}. \quad (4.3)$$

The training process is repeated iteratively until the error reaches a predefined threshold.

Perceptrons can perfectly classify linearly separable data with a finite number of mistakes, as stated in Theorem 4.1.

**Theorem 4.1** (Convergence of perceptrons (Novikoff, 1962)). *Suppose the training data consists of unit vectors separated by a margin of  $\gamma$  with labels  $y^{(i)} \in \{-1, 1\}$ . Then there exists a perceptron training algorithm that achieves zero error with at most  $\mathcal{O}(\frac{1}{\gamma^2})$  mistakes.*

*Proof of Theorem 4.1.* Consider the initial parameter of the perceptron  $\mathbf{w} = \mathbf{0}$ . Since the training dataset is linearly separable by a margin of  $\gamma$ , there exists a unit vector  $\mathbf{w}^*$  such that  $y^{(i)} \mathbf{w}^{*\top} \mathbf{x}^{(i)} \geq \gamma$  for all samples  $i \in [n]$ . Let  $\mathbf{x}^{(s,t)}$  be the sample that is misclassified in the  $t$ -th step, which is then used for adjusting the parameter. Let  $\mathbf{w}(t)$  be the parameter after the  $t$ -th step. Using Eqn. (4.3), it can be shown that

$$\begin{aligned} & \mathbf{w}^{*\top} \mathbf{w}(t) - \mathbf{w}^{*\top} \mathbf{w}(t-1) \\ &= (y^{(s,t)} - \hat{y}^{(s,t)}) \mathbf{w}^{*\top} \mathbf{x}^{(s,t)} \\ &= 2y^{(s,t)} \mathbf{w}^{*\top} \mathbf{x}^{(s,t)} \geq 2\gamma, \end{aligned} \quad (4.4)$$

where Eqn. (4.4) is derived by noticing the sample  $(\mathbf{x}^{(s,t)}, y^{(s,t)})$  is misclassified with  $\hat{y}^{(s,t)} \neq y^{(s,t)}$  and  $y^{(s,t)}, \hat{y}^{(s,t)} \in \{-1, 1\}$ . By considering the initialization  $\mathbf{w}(0) = \mathbf{0}$ , the norm of the parameter after the  $t$ -th step can be bounded by

$$\|\mathbf{w}(t)\| \geq \left| \mathbf{w}^{*\top} \mathbf{w}(t) \right| \quad (4.5)$$

$$= \left| \mathbf{w}^{*\top} \sum_{t'=1}^t (\mathbf{w}(t') - \mathbf{w}(t'-1)) \right| \quad (4.6)$$

$$\geq 2\gamma t, \quad (4.7)$$

where Eqn. (4.5) is follows from the condition  $\|\mathbf{w}^*\| = 1$ . Eqn. (4.7) is

derived by using the result in Eqn. (4.4). On the other hand,

$$\begin{aligned} & \|\mathbf{w}(t)\|^2 - \|\mathbf{w}(t-1)\|^2 \\ &= \left\| \mathbf{w}(t-1) + \left( y^{(s,t)} - \hat{y}^{(s,t)} \right) \mathbf{x}^{(s,t)} \right\|^2 - \|\mathbf{w}(t-1)\|^2 \end{aligned} \quad (4.8)$$

$$= \left\| \mathbf{w}(t-1) + 2y^{(s,t)} \mathbf{x}^{(s,t)} \right\|^2 - \|\mathbf{w}(t-1)\|^2 \quad (4.9)$$

$$= 4\|\mathbf{x}^{(s,t)}\|^2 + 4\mathbf{w}(t-1)^\top y^{(s,t)} \mathbf{x}^{(s,t)} \quad (4.10)$$

$$\leq 4 + 4\mathbf{w}(t-1)^\top y^{(s,t)} \mathbf{x}^{(s,t)} \quad (4.11)$$

where Eqn. (4.8) follows from the weight update rule in Eqn. (4.3). Eqn. (4.9) is derived by noticing that  $y^{(s,t)} \neq \hat{y}^{(s,t)}$  and  $y^{(s,t)}, \hat{y}^{(s,t)} \in \{-1, 1\}$ . Eqn. (4.10) follows from the condition  $\|\mathbf{x}^{(i)}\| = 1$  for all samples. Eqn. (4.11) is derived by noticing that the sample  $(\mathbf{x}^{(s,t)}, y^{(s,t)})$  is misclassified by the perceptron with the parameter  $\mathbf{w}(t-1)$ , *i.e.*

$$y^{(s,t)} \mathbf{w}(t-1)^\top \mathbf{x}^{(s,t)} \leq 0.$$

Thus, after  $t$  steps, the parameter is bounded by

$$\|\mathbf{w}(t)\| \leq 2\sqrt{t}. \quad (4.12)$$

Combining Eqn. (4.7) and Eqn. (4.12), it can be shown that

$$t \leq \frac{1}{\gamma^2}. \quad (4.13)$$

□

Since the parameters are used in an inner product operation, as shown in Eqn. (4.1), the single-layer perceptron can be considered as a basic kernel method employing the identity feature mapping. Consequently, the single-layer perceptron can only classify linearly separable data and is inadequate for handling more complex tasks, such as the XOR problem (Rosenblatt, 1958). This limitation has driven the development of advanced neural networks, such as multilayer perceptrons (MLPs) (Gardner and Dorling, 1998), which can capture non-linear relationships by incorporating non-linear activation functions and multi-layer structures.

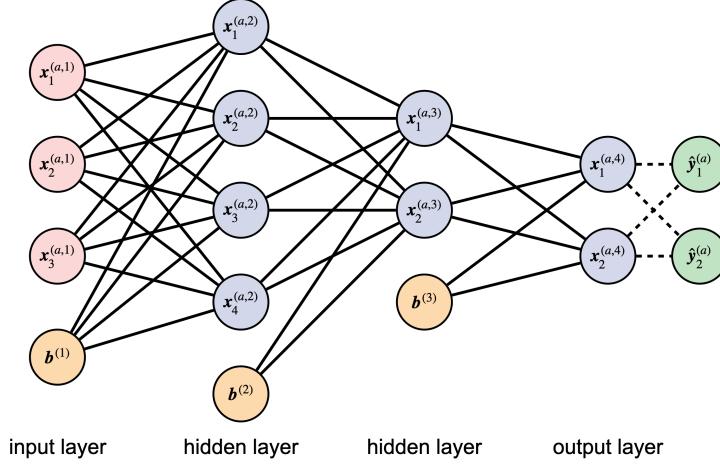


Figure 4.2: **Illustration of a multilayer perceptron with two hidden layers.** Dashed lines denote softmax operations.

### 4.1.2 Multilayer perceptron

The multilayer perceptron (MLP) is a fully connected neural network architecture consisting of three components: the *input layer*, *hidden layers*, and *output layer*, as illustrated in Figure 4.2. Similar to the single-layer perceptron introduced in Chapter 4.1.1, the neurons in the MLP are connected through weighted sums, followed by non-linear activation functions.

The mathematical expression of MLP is as follows. Let  $\mathbf{x}^{(a,1)}$  be the  $a$ -th input data and  $\ell = 1$  denote the input layer. Define  $L$  as the number of total layers. The forward propagation at the  $(\ell + 1)$ -th layer  $\forall \ell \in \{1, 2, \dots, L - 2\}$  yields

$$\begin{aligned}\mathbf{z}^{(a,\ell+1)} &= W^{(\ell)} \mathbf{x}^{(a,\ell)} + \mathbf{b}^{(\ell)}, \\ \mathbf{x}^{(a,\ell+1)} &= \sigma(\mathbf{z}^{(a,\ell+1)}),\end{aligned}$$

where  $\sigma$  represents the non-linear activation function, and  $W^{(\ell)}$  and  $\mathbf{b}^{(\ell)}$  denotes trainable weight and the bias term, respectively. Similar to the notation  $z$  in the perceptron in Chapter 4.1.1,  $\mathbf{z}^{(a,\ell+1)}$  denotes the output of the linear sum in the  $\ell + 1$ -th layer, which is expressed in a more generalized vector form. Therefore, the parameter for the weighted linear sum is represented in matrix form as  $W^{(\ell)}$ . Various methodologies have been proposed for implementing non-linear activations, with some common approaches summarized in Table 4.1.

Table 4.1: Common non-linear activation functions.

Name	Formulation
sigmoid function	$\sigma(x) = 1/(1 + \exp(-x))$
hyperbolic tangent function	$\sigma(x) = \tanh(x)$
rectified linear unit (ReLU) function	$\sigma(x) = \max(0, x)$

After passing through  $L - 2$  hidden layers, the output of MLP given by the equation below serves as the prediction to approximate the target label  $\mathbf{y}^{(a)}$ , i.e.,

$$\hat{\mathbf{y}}^{(a)} = \text{softmax}(\mathbf{x}^{(a,L)}) := \frac{\left( \exp(\mathbf{x}_1^{(a,L)}), \dots, \exp(\mathbf{x}_p^{(a,L)}) \right)^\top}{\sum_{i=1}^p \exp(\mathbf{x}_i^{(a,L)})},$$

with  $p$  here denotes the dimension of  $\mathbf{x}^{(a,L)}$ .

Next, we provide a toy example of binary classification to explain the MLP learning procedure. Let  $\{(\mathbf{x}^{(a)}, \mathbf{y}^{(a)})\}_{a \in \mathcal{D}}$  be the training dataset  $\mathcal{D}$ , where  $\mathbf{x}^{(a)}$  is the feature vector and  $\mathbf{y}^{(a)} \in \{(1, 0)^\top, (0, 1)^\top\}$  is the label for two categories. Consider the MLP with one hidden layer. The prediction can be expressed as follows:

$$\begin{aligned} \hat{\mathbf{y}}^{(a)} &= \text{softmax}(\mathbf{x}^{(a,3)}) = \text{softmax} \circ \sigma(\mathbf{z}^{(a,3)}) \\ &= \text{softmax} \circ \sigma(W^{(2)}\mathbf{x}^{(a,2)} + \mathbf{b}^{(2)}) \\ &= \text{softmax} \circ \sigma(W^{(2)}\sigma(\mathbf{z}^{(a,2)}) + \mathbf{b}^{(2)}) \\ &= \text{softmax} \circ \sigma(W^{(2)}\sigma(W^{(1)}\mathbf{x}^{(a,1)} + \mathbf{b}^{(1)}) + \mathbf{b}^{(2)}), \end{aligned}$$

where  $\circ$  denotes the function composition. We use  $\sigma(x) = 1/(1 + \exp(-x))$  as the non-linear activation function.

MLP learns from the given dataset by minimizing the loss function with respect to the parameters  $\boldsymbol{\theta} = (W^{(1)}, W^{(2)}, \mathbf{b}^{(1)}, \mathbf{b}^{(2)})$ , which is defined as the  $\ell_2$  norm distance between the prediction and the label,

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{D}|} \sum_{a \in \mathcal{D}} \mathcal{L}^{(a)}(\boldsymbol{\theta}) = \frac{1}{2|\mathcal{D}|} \sum_{a \in \mathcal{D}} \left\| \hat{\mathbf{y}}^{(a)}(\boldsymbol{\theta}) - \mathbf{y}^{(a)} \right\|^2. \quad (4.14)$$

We use gradient descent with learning rate  $\eta$  to optimize the parameters:

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}(t)).$$

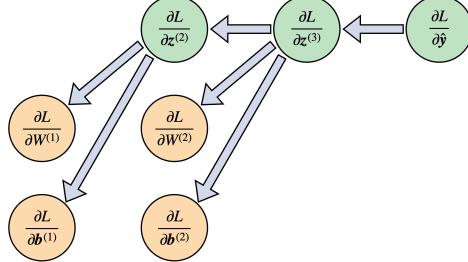


Figure 4.3: **Illustration of backpropagation when calculating the gradient of an MLP with one hidden layer.** The index of sample  $a$  is omitted for simplicity.

As illustrated in Figure 4.3, the gradient is computed using backpropagation (LeCun et al., 1988) as follows. First, the gradient with respect to the output layer is given by

$$\begin{aligned}\frac{\partial \mathcal{L}^{(a)}}{\partial \hat{\mathbf{y}}^{(a)}} &= \hat{\mathbf{y}}^{(a)} - \mathbf{y}^{(a)}, \\ \frac{\partial \mathcal{L}^{(a)}}{\partial \mathbf{x}^{(a,3)}} &= \frac{\partial \hat{\mathbf{y}}^{(a)}}{\partial \mathbf{x}^{(a,3)}} \frac{\partial \mathcal{L}^{(a)}}{\partial \hat{\mathbf{y}}^{(a)}} = \left[ \text{diag}(\hat{\mathbf{y}}^{(a)}) - \hat{\mathbf{y}}^{(a)} \hat{\mathbf{y}}^{(a)\top} \right] \frac{\partial \mathcal{L}^{(a)}}{\partial \hat{\mathbf{y}}^{(a)}}, \\ \frac{\partial \mathcal{L}^{(a)}}{\partial \mathbf{z}^{(a,3)}} &= \frac{\partial \mathbf{x}^{(a,3)}}{\partial \mathbf{z}^{(a,3)}} \frac{\partial \mathcal{L}^{(a)}}{\partial \mathbf{x}^{(a,3)}} = \text{diag} \left[ (\mathbf{1} - \mathbf{z}^{(a,3)}) \odot \mathbf{z}^{(a,3)} \right] \frac{\partial \mathcal{L}^{(a)}}{\partial \mathbf{x}^{(a,3)}},\end{aligned}$$

where  $\mathbf{1}$  denotes the vector  $(1, 1, \dots, 1)^\top$ , and  $\odot$  denotes the element-wise multiplication (Hadamard product). For convenience, we omit the dimension of  $\mathbf{1}$  here, which has the same dimension with  $\mathbf{z}^{(a,3)}$ . Next, the gradient with respect to the hidden layer can be obtained using the chain rule:

$$\begin{aligned}\frac{\partial \mathcal{L}^{(a)}}{\partial \mathbf{x}^{(a,2)}} &= \frac{\partial \mathbf{z}^{(a,3)}}{\partial \mathbf{x}^{(a,2)}} \frac{\partial \mathcal{L}^{(a)}}{\partial \mathbf{z}^{(a,3)}} = W^{(2)} \frac{\partial \mathcal{L}^{(a)}}{\partial \mathbf{z}^{(a,3)}}, \\ \frac{\partial \mathcal{L}^{(a)}}{\partial W^{(2)}} &= \frac{\partial \mathbf{z}^{(a,3)}}{\partial W^{(2)}} \frac{\partial \mathcal{L}^{(a)}}{\partial \mathbf{z}^{(a,3)}} = \frac{\partial \mathcal{L}^{(a)}}{\partial \mathbf{z}^{(a,3)}} \mathbf{x}^{(a,2)\top}, \\ \frac{\partial \mathcal{L}^{(a)}}{\partial \mathbf{b}^{(2)}} &= \frac{\partial \mathbf{z}^{(a,3)}}{\partial \mathbf{b}^{(2)}} \frac{\partial \mathcal{L}^{(a)}}{\partial \mathbf{z}^{(a,3)}} = \frac{\partial \mathcal{L}^{(a)}}{\partial \mathbf{z}^{(a,3)}}, \\ \frac{\partial \mathcal{L}^{(a)}}{\partial \mathbf{z}^{(a,2)}} &= \frac{\partial \mathbf{x}^{(a,2)}}{\partial \mathbf{z}^{(a,2)}} \frac{\partial \mathcal{L}^{(a)}}{\partial \mathbf{x}^{(a,2)}} = \text{diag} \left[ (\mathbf{1} - \mathbf{z}^{(a,2)}) \odot \mathbf{z}^{(a,2)} \right] \frac{\partial \mathcal{L}^{(a)}}{\partial \mathbf{x}^{(a,2)}}.\end{aligned}$$

The gradient with respect to the parameters for the input layer is derived

similarly:

$$\begin{aligned}\frac{\partial \mathcal{L}^{(a)}}{\partial W^{(1)}} &= \frac{\partial \mathbf{z}^{(a,2)}}{\partial W^{(1)}} \frac{\partial \mathcal{L}^{(a)}}{\partial \mathbf{z}^{(a,2)}} = \frac{\partial \mathcal{L}^{(a)}}{\partial \mathbf{z}^{(a,2)}} \mathbf{x}^{(a,1)\top}, \\ \frac{\partial \mathcal{L}^{(a)}}{\partial \mathbf{b}^{(1)}} &= \frac{\partial \mathbf{z}^{(a,2)}}{\partial \mathbf{b}^{(1)}} \frac{\partial \mathcal{L}^{(a)}}{\partial \mathbf{z}^{(a,2)}} = \frac{\partial \mathcal{L}^{(a)}}{\partial \mathbf{z}^{(a,2)}}.\end{aligned}$$

After multiple training epochs, the loss function converges to a value below a predefined threshold, which leads to a small classification error.

Compared to single-layer perceptrons, MLP can model non-linear relationships by employing hidden layers and activation functions. This enables them to learn abstract representations by capturing the complex patterns inherent in the data. Mathematically, the power of MLPs is guaranteed by the universal approximation theorem, as stated in Theorem 4.2, which asserts that a single hidden layer is sufficient to approximate any arbitrary continuous function.

**Fact 4.2** (Universal Approximation Theorem, informal version adapted from Hornik et al. (1989)). *Let  $\mathcal{C}(\mathcal{X}, \mathbb{R}^m)$  denote the set of continuous functions from a subset  $\mathcal{X}$  of a Euclidean space  $\mathbb{R}^n$  to a Euclidean space  $\mathbb{R}^m$ . Denote by  $\sigma$  a function that is not polynomial. Then for every  $n, m \in \mathbb{N}$ , compact set  $\mathcal{K} \subseteq \mathbb{R}^n$ ,  $f \in \mathcal{C}(\mathcal{K}, \mathbb{R}^m)$ , and  $\epsilon > 0$ , there exist  $k \in \mathbb{N}$ ,  $A \in \mathbb{R}^{k \times n}$ ,  $\mathbf{b} \in \mathbb{R}^k$ , and  $C \in \mathbb{R}^{m \times k}$  such that*

$$\sup_{\mathbf{x} \in \mathcal{X}} \|f(\mathbf{x}) - g(\mathbf{x})\| < \epsilon,$$

where  $g(\mathbf{x}) = C\sigma(A\mathbf{x} + \mathbf{b})$ .

#### Remark

MLPs involve a large number of parameters due to their fully connected multilayer architecture. This high parameter count enables MLPs to possess considerable representational power, allowing them to model complex data distributions. However, the excessive capacity to fit the training data often leads to overfitting (Caruana et al., 2000), where the MLP captures noise and irrelevant patterns instead of generalizable features. As a result, MLPs tend to perform poorly on unseen data, especially when the training set is limited or noisy. To mitigate this issue, advanced techniques such as dropout (Srivastava et al., 2014), weight decay (Krogh and Hertz, 1991), and attention mechanisms (Vaswani, 2017) have been proposed to reduce overfit-

ting in MLPs while maintaining sufficient expressivity.

## 4.2 Fault-tolerant Quantum Perceptron

The primary aim of advancing quantum machine learning is to harness the computational advantages of quantum mechanics to enhance performance across various learning tasks. As outlined in Chapter 1.1.2, these advantages manifest in several ways, including reduced runtime, lower query complexity, and improved sample efficiency compared to classical models. A notable example of this is the quantum perceptron model (Kapoor et al., 2016). As a FTQC-based QML algorithm grounded in the Grover search, quantum perceptron offers a quadratic improvement in the query complexity during the training over its classical counterpart. For comprehensiveness, we first introduce the Grover search algorithm, followed by a detailed explanation of the quantum perceptron model.

### 4.2.1 Grover search

Grover search (Grover, 1996) provides runtime speedups for unstructured search problems, which have broad applications in cryptography, quantum machine learning, and constraint satisfaction problems. Unlike classical search methods that require  $\mathcal{O}(d)$  queries for a dataset with  $d$  entries, Grover’s algorithm can identify the target element with high probability using only  $\mathcal{O}(\sqrt{d})$  queries to a quantum oracle. Consequently, quantum algorithms incorporating Grover search have the potential to achieve a quadratic speedup over classical approaches.

In general, a search task can be abstracted as a function  $f(x)$  such that  $f(x) = 1$  if  $x$  belongs to the solution set of the search problem, and  $f(x) = 0$  otherwise. We consider a dataset consisting of  $d = 2^N$  elements, where each element is represented by the quantum state  $|x\rangle$  with  $x = 0, 1, \dots, d - 1$ . In this process, two key quantum oracles are introduced. The first oracle,  $U_0 = 2(|0\rangle\langle 0|)^{\otimes N} - \mathbb{I}_d$ , applies a phase shift of  $e^{i\pi} = -1$  to all quantum states except  $|0\rangle^{\otimes N}$ , which remains unchanged. The second oracle,  $U_f$ , operates in a similar manner: it applies a phase shift of  $-1$  to quantum states that belong to the solution set while leaving all other states unaffected. The procedure for Grover search is described in Algorithm 2.

**Theorem 4.3** (Time complexity of Grover search). *Grover search finds a solution to the unstructured search problem with high probability in time*

**Algorithm 2** Grover search

**Require:** Quantum oracles  $U_f$  and  $U_0$ . The size of the dataset and the solution set, denoted by  $d = 2^N$  and  $M$ , respectively.

**Ensure:** An index corresponds to one of the solution states with high probability.

1: Initialize a register of  $N$  qubits with the state of uniform superposition:

$$|\phi_0\rangle = \frac{1}{\sqrt{d}} \sum_{x=0}^{d-1} |x\rangle = \bigotimes_{n=1}^N \frac{|0\rangle + |1\rangle}{\sqrt{2}} = \left( \bigotimes_{n=1}^N H \right) |0\rangle.$$

2: Let  $m = \lfloor \frac{\pi}{4} \sqrt{\frac{d}{M}} - \frac{1}{2} \rfloor$ . Apply the following operation:

$$|\phi_m\rangle = [H^{\otimes N} U_0 H^{\otimes N} U_f]^m |\phi_0\rangle.$$

3: Measure the state  $|\phi_m\rangle$  to generate an index.

$\mathcal{O}(\sqrt{\frac{d}{M}}(\log d + T_f))$ , where  $d$  is the size of the dataset,  $M$  is the size of the solution set, and  $T_f$  denotes the time complexity of implementing the oracle  $U_f$ .

## Remark

The Grover search achieves a quadratic speed-up in the query complexity of the oracle  $U_f$ . It provides a quantum advantage in the runtime only if the time complexity of the oracle  $U_f$ , denoted as  $T_f$ , is less than  $\sqrt{d}$ .

*Proof of Theorem 4.3.* Let the superposition of the solution state be

$$|\text{target}\rangle = \frac{1}{\sqrt{M}} \sum_{x:f(x)=1} |x\rangle.$$

Similarly, the superposition of the states outside the solution set is given by

$$|\text{other}\rangle = \frac{1}{\sqrt{2^N - M}} \sum_{x|f(x)=0} |x\rangle.$$

Thus, the initial uniform superposition state can be expressed as

$$|\phi_0\rangle = \sqrt{\frac{M}{2^N}} |\text{target}\rangle + \sqrt{\frac{2^N - M}{2^N}} |\text{other}\rangle := \alpha_0 |\text{target}\rangle + \beta_0 |\text{other}\rangle.$$

In principle, the coefficient associated with the target state is expected to increase during the quantum state evolution, such that the solution could be obtained through quantum measurement with high probability. The dynamics of these coefficients can be described as follows:

$$\begin{aligned}\alpha_k &= \langle \text{target} | \phi_k \rangle \\ &= \langle \text{target} | H^{\otimes N} U_0 H^{\otimes N} U_f | \phi_{k-1} \rangle \\ &= \langle \text{target} | (2|\phi_0\rangle\langle\phi_0| - \mathbb{I}) U_f (\alpha_{k-1} |\text{target}\rangle + \beta_{k-1} |\text{other}\rangle) \\ &= \langle \text{target} | (2|\phi_0\rangle\langle\phi_0| - \mathbb{I})(-\alpha_{k-1} |\text{target}\rangle + \beta_{k-1} |\text{other}\rangle) \\ &= (1 - 2\alpha_0^2)\alpha_{k-1} + 2\alpha_0\beta_0\beta_{k-1}, \\ \beta_k &= \langle \text{other} | \phi_k \rangle \\ &= \langle \text{other} | (2|\phi_0\rangle\langle\phi_0| - \mathbb{I})(-\alpha_{k-1} |\text{target}\rangle + \beta_{k-1} |\text{other}\rangle) \\ &= (2\beta_0^2 - 1)\beta_{k-1} - 2\alpha_0\beta_0\alpha_{k-1}.\end{aligned}$$

Let the angle  $\theta = \arccos \sqrt{\frac{2^N - M}{2^N}}$ , then by induction, it can be shown that

$$\alpha_k = \sin[(2k+1)\theta], \quad \beta_k = \cos[(2k+1)\theta].$$

To ensure that the coefficient  $\alpha_m = \mathcal{O}(1)$ , there is a condition  $(2m+1)\theta \approx \pi/2$ . Therefore,  $m = \mathcal{O}(1/\theta) = \mathcal{O}(\sqrt{\frac{2^N}{M}})$  suffices to obtain the solution with high probability.

□

#### 4.2.2 Online quantum perceptron with quadratic speedups

As stated in Theorem 4.1, for a linearly separable dataset with a margin  $\gamma$ , a perceptron model can achieve perfect classification after making  $\mathcal{O}(1/\gamma^2)$  mistakes during training. In classical approaches, identifying a sample that is misclassified by the current model may require up to  $\mathcal{O}(d)$  queries, where  $d$  denotes the size of the training dataset. In contrast, the quantum perceptron model (Kapoor et al., 2016) can identify misclassified samples more efficiently by employing the Grover search algorithm, achieving a quadratic speed-up in the query complexity.

To begin, we introduce the setup of the input data. We consider the classification of a dataset  $\{z^{(i)}\}_{i=1}^d = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^d$ , where the label  $y^{(i)} \in \{-1, 1\}$ . For convenience, we assume that the number of samples is a power of 2, i.e.,  $d = 2^N$ . Each data vector  $\mathbf{x}^{(i)}$  is assumed to be represented by using  $B$  bits. The information of each sample  $z^{(i)}$  is stored in the quantum state  $|z^{(i)}\rangle$  by using  $B + 1$  qubits.

**Example 4.4.** For the sample  $(\mathbf{x}^{(i)}, y^{(i)}) = ([0, 0, 1, 0], 1)$ , the corresponding quantum state is  $|z^{(i)}\rangle = |00101\rangle$ , where the last qubit encodes the label (with “0” for the label “−1”), and the remaining qubits represent the data vector. When  $\mathbf{x}^{(i)}$  is a float vector, a similar bit sequence can be obtained by concatenating the binary representations of the elements in  $\mathbf{x}^{(i)}$ .

Next, we introduce the oracle models. We assume the existence of a quantum oracle  $U$  for encoding training data as the corresponding quantum state, i.e.,

$$U|i\rangle|0\rangle = |i\rangle|z^{(i)}\rangle, \quad U^\dagger|i\rangle|z^{(i)}\rangle = |i\rangle|0\rangle. \quad (4.15)$$

Due to the linearity of unitary,

$$U \sum_{i=0}^{d-1} \frac{1}{\sqrt{d}} |i\rangle|0\rangle = \sum_{i=0}^{d-1} \frac{1}{\sqrt{d}} |i\rangle|z^{(i)}\rangle. \quad (4.16)$$

In addition to the input oracle  $U$  described in Eqn. (4.15), the quantum perceptron model employs another oracle to distinguish between correctly classified and misclassified quantum states. Specifically, the oracle  $F'_w$  satisfies

$$F'_w|z^{(i)}\rangle = (-1)^{f(\mathbf{w}, z^{(i)})}|z^{(i)}\rangle, \quad (4.17)$$

where  $f : (\mathbf{w}, z^{(i)}) \rightarrow \{0, 1\}$ . The function outputs 1 if the current perceptron model with weight  $\mathbf{w}$  misclassifies the training sample  $z^{(j)}$ ; otherwise, it outputs 0. Furthermore, we define

$$F_w = U^\dagger(\mathbb{I} \otimes F'_w)U, \quad (4.18)$$

which is used as the oracle  $U_f$  in the Grover search. The online quantum perceptron procedure is given in Algorithm 3. The query complexity of the online quantum perceptron is provided in Theorem 4.5.

**Theorem 4.5** (Online quantum perceptron (Kapoor et al., 2016)). *Consider a training dataset that consists of unit vectors  $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(d)}\}$  and labels  $\{y^{(1)}, \dots, y^{(d)}\}$  with a margin  $\gamma$ . Denote by  $n_{\text{quant}}$  the number of queries to  $F_w$  needed to learn the weight  $\mathbf{w}$ , such that the training dataset is perfectly classified with probability at least  $1 - \epsilon$ , then*

$$n_{\text{quant}} \in \mathcal{O}\left(\frac{\sqrt{d}}{\gamma^2} \log \frac{1}{\gamma^2 \epsilon}\right).$$

---

**Algorithm 3** Online quantum perceptron

---

**Require:** Linearly separable dataset  $\{z^{(i)}\}_{i=1}^d = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^d$ , where  $d = 2^N$ . Margin threshold  $\gamma$ . Constants  $\epsilon \in (0, 1)$  and  $c \in (1, 2)$ .

**Ensure:** Weight  $\mathbf{w}$  for a perceptron that correctly classifies the dataset with a margin  $\gamma$  with probability at least  $1 - \epsilon$ .

- 1: Initialize the weight  $\mathbf{w} = \mathbf{0}$ .
- 2: **for**  $h = 1, \dots, \lceil \frac{1}{\gamma^2} \rceil$  **do**
- 3:     **for**  $k = 1, \dots, \lceil \log_{3/4} \gamma^2 \epsilon \rceil$  **do**
- 4:         **for**  $j = 1, \dots, \lceil \log_c \frac{1}{\sin(2 \sin^{-1}(1/\sqrt{d}))} \rceil$  **do**
- 5:             Draw  $m$  uniformly from  $\{0, \dots, \lceil c^j \rceil - 1\}$ .
- 6:             Prepare the quantum state

$$|\phi_0\rangle = \frac{1}{\sqrt{d}} \sum_{i=0}^{d-1} |i\rangle.$$

- 7:     Generate the state

$$|\phi_1\rangle = \{(2|\phi_0\rangle\langle\phi_0| - \mathbb{I}_d) \otimes \mathbb{I}_d\}^m |\phi_0\rangle |0\rangle^{\otimes N}.$$

- 8:     Measure the first register of the state  $|\phi_1\rangle$  to obtain an outcome  $q$ .
  - 9:     **if**  $f(\mathbf{w}, z^{(q)}) = 1$  **then**
  - 10:         Update  $\mathbf{w} \leftarrow \mathbf{w} + y^{(q)} \mathbf{x}^{(q)}$ .
  - 11:     **end if**
  - 12:     **end for**
  - 13:     **end for**
  - 14: **end for**
  - 15: Output  $\mathbf{w}$ .
-

For the classical case where the training vectors are uniformly sampled from the training dataset, the number of queries to  $f_w$  is bounded by

$$\Omega(d) \ni n_{\text{class}} \in \mathcal{O}\left(\frac{d}{\gamma^2} \log \frac{1}{\epsilon\gamma^2}\right).$$

*Proof of Theorem 4.5.* The main idea of the quantum perceptron model in Algorithm 3 is to replace the procedure of finding the misclassified sample in classical perceptrons with the Grover search. Due to convergence result for perceptrons in Theorem 4.1,  $h = 1, \dots, \lceil \frac{1}{\gamma^2} \rceil$  iterations of Steps (3-13) suffice to update the weight  $w$  towards the case of perfect classification. Therefore, Theorem 4.5 is the direct consequence of the following lemmas and Theorem 4.1. The query complexity of classical perceptrons has the lower bound  $\Omega(d)$ , since the model needs to go through the entire dataset in the worst case.  $\square$

**Lemma 4.6.** *Given only uniform sampling access to the training dataset, there exists a classical perceptron that either finds a misclassified sample to update the weight  $w$  or concludes that no such example exists with probability  $1 - \epsilon\gamma^2$ , using  $\mathcal{O}(d \log(1/\epsilon\gamma^2))$  queries to  $f_w$ .*

**Lemma 4.7.** *The procedure of Steps 3-13 in Algorithm 3 either finds a misclassified sample to update the weight  $w$  or concludes that no such example exists with probability  $1 - \epsilon\gamma^2$ , using  $\mathcal{O}(\sqrt{d} \log(1/\epsilon\gamma^2))$  queries to  $F_w$ .*

*Proof of Lemma 4.6.* First, let  $m_c = d[\log(1/\epsilon\gamma^2)]$  be the number of samples drawn from the dataset uniformly in each iteration of training. Suppose these samples are classified correctly, then the probability that the entire dataset is classified correctly is

$$\Pr(\text{Correct classification}) \geq 1 - \left(1 - \frac{1}{d}\right)^{m_c} \geq 1 - \exp\left(-\frac{m_c}{d}\right) \geq 1 - \epsilon\gamma^2.$$

$\square$

*Proof of Lemma 4.7.* For convenience, denote  $\theta_a := \arccos \sqrt{\frac{d-d_0}{d}}$ , where  $d_0$  the number of misclassified samples in the dataset according to the current model. Let  $d_1 := \lceil \log_c \frac{1}{\sin(2\sin^{-1}(1/\sqrt{d}))} \rceil$ . Here, an exponential expansion strategy is used in Steps 4-12 to handle the scenario of unknown  $d_0$ . Namely, quantum operations in the Grover search are repeated for  $m$  times, where  $m$  is drawn from an exponentially expanded set  $0, \dots, \lceil c^j \rceil - 1$  uniformly for

a predefined  $c \in (1, 2)$  and  $j = 1, \dots, d_1$ . It can be shown that this strategy can find a misclassified sample before the convergence of Algorithm 3 with an average probability at least  $1/4$ :

$$\begin{aligned} \Pr(f(\mathbf{w}, z^{(q)}) = 1) &= \sum_{j=1}^{d_1} \frac{1}{\lceil c^j \rceil} \sum_{m=0}^{\lceil c^j \rceil - 1} \sin^2((2m+1)\theta_a) \\ &\geq \frac{1}{\lceil c^{d_1} \rceil} \sum_{m=0}^{\lceil c^{d_1} \rceil - 1} \sin^2((2m+1)\theta_a) \\ &= \frac{1}{2} \left[ 1 - \frac{\sin(4\lceil c^{d_1} \rceil \theta_a)}{2\lceil c^{d_1} \rceil \sin(2\theta_a)} \right] \\ &\geq \frac{1}{4}. \end{aligned}$$

The procedure of Steps 4-12 is repeated for  $k = 1, \dots, \lceil \log_{3/4} \gamma^2 \epsilon \rceil$  iterations to accumulate the success probability. The probability of finding a misclassified sample in Steps 3-13 before the convergence of Algorithm 3 is at least

$$1 - \left(1 - \frac{1}{4}\right)^{\lceil \log_{3/4} \epsilon \gamma^2 \rceil} \geq 1 - \epsilon \gamma^2. \quad (4.19)$$

Finally, the query complexity  $Q$  of Steps 3-13 in Algorithm 3 can be upper bounded as follows:

$$\begin{aligned} Q &\leq \sum_{k=1}^{\lceil \log_{3/4} \gamma^2 \epsilon \rceil} \sum_{j=1}^{d_1} c^j \\ &\leq \left(1 + \log_{3/4} \gamma^2 \epsilon\right) \frac{c}{1-c} \left[1 - c^{d_1}\right] \\ &\leq \left(1 + \log_{3/4} \gamma^2 \epsilon\right) \frac{c^2}{c-1} \left[\frac{1}{\sin(2 \sin^{-1}(1/\sqrt{d}))} - 1\right] \\ &= \mathcal{O}(\sqrt{d} \log \frac{1}{\epsilon \gamma^2}). \end{aligned}$$

□

### 4.3 Near-term Quantum Neural Networks

Following recent experimental breakthroughs in superconducting quantum hardware architectures (Arute et al., 2019; Acharya et al., 2024; AbuGhanem,

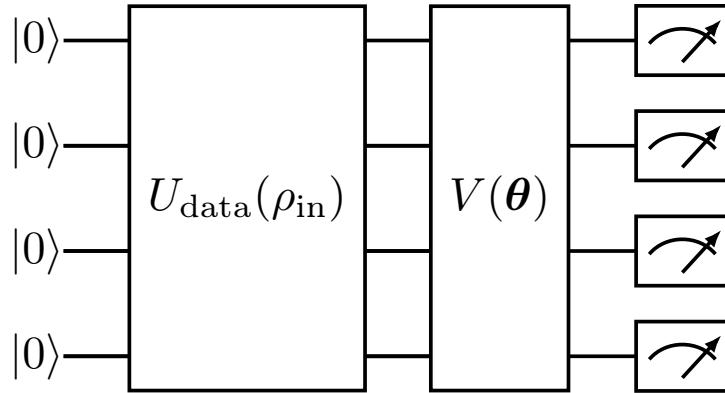
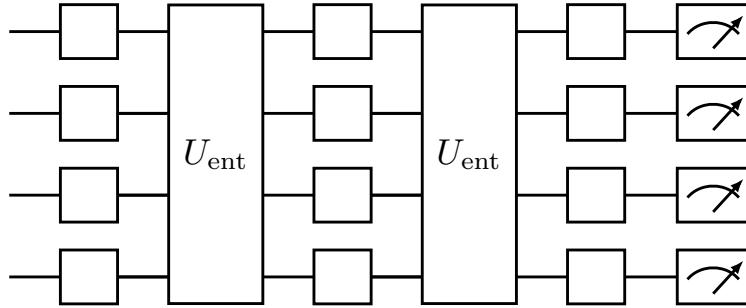


Figure 4.4: **Illustration of a QNN.** The input state  $\rho_{\text{in}}$  is prepared using the operation  $U_{\text{data}}$ , followed by a variational quantum circuit (VQC)  $V(\theta)$  and the measurement operation.

2024; Gao et al., 2024), researchers have devoted considerable effort to developing and implementing quantum machine learning algorithms optimized for current and near-term quantum devices (Wang and Liu, 2024). Compared to fault-tolerant quantum computers, these devices face three primary limitations: quantum noise, limited coherence time, and circuit connectivity constraints. Regarding quantum noise, state-of-the-art devices have single-qubit gate error rates of  $10^{-4} \sim 10^{-3}$  and two-qubit gate error rates of approximately  $10^{-3} \sim 10^{-2}$  (AbuGhanem, 2024; Gao et al., 2024). The coherence time is around  $10^2 \mu\text{s}$  (Acharya et al., 2024; AbuGhanem, 2024; Gao et al., 2024), primarily limited by decoherence in noisy quantum channels. Regarding circuit connectivity, most superconducting quantum processors employ architectures that exhibit two-dimensional connectivity patterns and their variants (Acharya et al., 2024; AbuGhanem, 2024; Gao et al., 2024). Gate operations between non-adjacent qubits must be executed through intermediate relay operations, leading to additional error accumulation. To address these inherent limitations, the quantum neural network (QNN) framework has been proposed. Specifically, these QNNs are designed to perform meaningful computations on near-term quantum devices.

### 4.3.1 General framework

In this section, we introduce the basic architecture of QNNs. As illustrated in Figure 4.4, a basic QNN consists of three components: the input, the



**Figure 4.5: Illustration of a hardware-efficient circuit with two entanglement layers.**

model circuit, and the measurement.

**Input.** The QNN uses quantum states  $\rho_{\text{in}}$  as input data. As shown in Table 4.2, QNNs can process both classical and quantum data. Specifically, the input states  $\rho_{\text{in}}$  may be introduced from physical processes such as quantum Hamiltonian evolutions or be constructed to encode classical vectors using encoding protocols introduced in Chapter 2.3.1, such as angle encoding and amplitude encoding.

**Model circuit.** QNNs employ variational quantum circuits (VQCs), a.k.a, ansatzes, to extract and learn features from input data. A typical VQC, denoted as  $V(\theta)$ , adopts a layered architecture that consists of both parameterized and fixed quantum gates, with the former being trainable. For problem-agnostic implementations, an effective parameterization strategy is to use the parameters  $\theta$  as the phases of single-qubit rotation gates RX, RY, RZ, while quantum entanglement is introduced through fixed two-qubit gates, such as CX and CZ. Standard circuit architectures include the hardware-efficient circuit (HEC) (Kandala et al., 2017a), shown in Figure 4.5, and the quantum convolutional neural network (QCNN) (Cong et al., 2019), shown in Figure 4.6. For problem-specific applications, such as finding the ground states of molecular Hamiltonians, specialized circuits like the unitary coupled cluster *ansatz* (Peruzzo et al., 2014) are employed.

**Example 4.8.** Hardware-efficient circuits incorporate several widely adopted ansatzes. Single-qubit rotations  $\{RX, RY, RZ\}$  are used to construct parameterized single-qubit unitaries. The entangled unitary layer

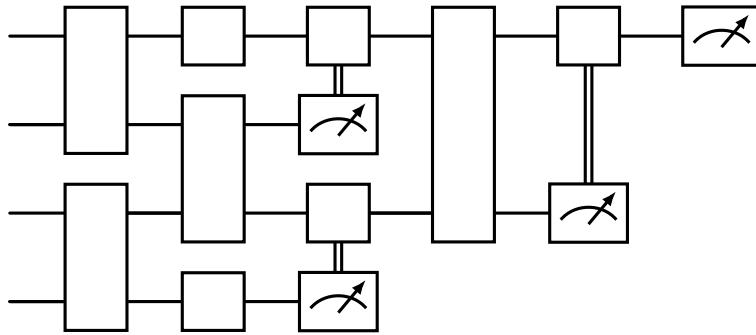


Figure 4.6: Illustration of a quantum convolutional neural network.

Table 4.2: Examples of classical and quantum data employed in QNNs, where  $H$  denotes the system Hamiltonian,  $k_B$  is the Boltzmann constant, and  $|\phi_0\rangle$  is a predefined initial state.

	Example	Input state formulation
Classical data	Angle encoding	$\bigotimes_{n=1}^N [RY(\mathbf{x}_n) 0\rangle]$
	Amplitude encoding	$\sum_{i=0}^{d-1} \mathbf{x}_i / \ \mathbf{x}\ _2  i\rangle$
Quantum data	Gibbs state	$\frac{\exp(-H/k_B T)}{\text{Tr}[\exp(-H/k_B T)]}$
	Hamiltonian evolution	$\exp(-iHt) \phi_0\rangle$

can be implemented using two-qubit gates such as

$$U_{\text{ent}} = \bigotimes_{n=1}^{\lceil \frac{N}{2} \rceil} U(2n - 1 + k\%2, 2n + k\%2) \quad \text{for the } k\text{-th layer},$$

$$U_{\text{ent}} = \bigotimes_{n=1}^{\lceil \frac{N}{2} \rceil} U(2n - 1, 2n) \bigotimes_{n=1}^{\lceil \frac{N-1}{2} \rceil} U(2n, 2n + 1),$$

where  $U \in \{CX, CZ\}$ .

**Measurement.** After implementing the model circuit, the quantum state is measured using specific observables, denoted as  $O$ , to extract classical information. The choice of observables depends on the experimental objectives. In the case of a variational quantum eigensolver, where the goal is to find the ground state and energy of a given Hamiltonian, the observable is chosen to be the target Hamiltonian itself. In quantum machine learning applications involving classical data, the measurement outcomes are used to approximate label information, which typically lacks direct physical significance. As a result, the observable can, in principle, be any Hermitian operator. However, for practical experimental considerations, a linear combination of Pauli-Z operators is commonly used as the observable:

$$O = \sum_{j=1}^N \mathbf{c}_j \mathbb{I}^{\otimes(j-1)} \otimes Z_j \otimes \mathbb{I}^{\otimes(N-j)}, \quad (4.20)$$

where  $\mathbf{c} \in \mathbb{R}^N$  is a weight vector. The measurement outcome of QNN can be expressed as a function of  $\boldsymbol{\theta}$ , i.e.,

$$f(\boldsymbol{\theta}; \rho_{\text{in}}, V, O) = \text{Tr}[OV(\boldsymbol{\theta})\rho_{\text{in}}V(\boldsymbol{\theta})^\dagger]. \quad (4.21)$$

**Training of QNNs.** As a QML framework, the optimization of QNNs amounts to updating parameters  $\boldsymbol{\theta}$  using gradient-based methods. Due to the linearity of quantum mechanics and the unitary evolution constraint, for certain cases, the acquisition of gradients can be elegantly performed using the parameter-shift rule.

**Theorem 4.9** (Parameter-shift rule (Crooks, 2019)). *Suppose the gate  $G_j(\boldsymbol{\theta}_j)$  in a VQC  $V(\boldsymbol{\theta})$  has a unitary Hamiltonian  $H_j$ , then the corresponding gradient could be obtained as*

$$\frac{\partial f}{\partial \boldsymbol{\theta}_j}(\boldsymbol{\theta}) = \frac{1}{2} \left[ f\left(\boldsymbol{\theta} + \frac{\pi}{2}\mathbf{e}^{(j)}\right) - f\left(\boldsymbol{\theta} - \frac{\pi}{2}\mathbf{e}^{(j)}\right) \right],$$

where the function  $f$  follows the Eqn. (4.21) and the one-hot vector  $\mathbf{e}^{(j)}$  has the same dimension with  $\boldsymbol{\theta}$  with the  $j$ -th element being 1.

*Proof of Theorem 4.9.* For convenience, we denote the detailed structure of VQC as

$$V(\boldsymbol{\theta}) = \prod_{i=L}^1 G_i(\boldsymbol{\theta}_i) W_i,$$

where  $L$  is the number of parameters in VQC,  $G_i$  is the parameterized gate, and  $W_i$  is the fixed gate. By assumption, the gate takes the form as

$$G_j(\boldsymbol{\theta}_j) = \exp(-iH_j\boldsymbol{\theta}_j/2),$$

where the Hamiltonian  $H_j$  is a unitary. For convenience, unnecessary parameterized and fixed gates can be merged into the state  $\rho_{\text{in}}$  and the observable  $O$ , i.e.,

$$\begin{aligned} \rho'_{\text{in}} &= W_j \left( \prod_{i=j-1}^1 G_i(\boldsymbol{\theta}_i) W_i \right) \rho_{\text{in}} \left( \prod_{i=1}^{j-1} W_i^\dagger G_i(\boldsymbol{\theta}_i)^\dagger \right) W_j^\dagger, \\ O' &= \left( \prod_{i=j}^L W_i^\dagger G_i(\boldsymbol{\theta}_i)^\dagger \right) O \left( \prod_{i=L}^j G_i(\boldsymbol{\theta}_i) W_i \right). \end{aligned}$$

It can be shown that

$$\begin{aligned} f(\boldsymbol{\theta}) &= \text{Tr} [OV(\boldsymbol{\theta})\rho_{\text{in}}V(\boldsymbol{\theta})^\dagger] \\ &= \text{Tr} [O'G_j(\boldsymbol{\theta}_j)\rho'_{\text{in}}G_j(\boldsymbol{\theta}_j)^\dagger] \\ &= \text{Tr} [O' \exp(-iH_j\boldsymbol{\theta}_j/2)\rho'_{\text{in}} \exp(iH_j\boldsymbol{\theta}_j/2)] \\ &= \cos^2 \frac{\boldsymbol{\theta}_j}{2} \text{Tr} [O'\rho'_{\text{in}}] + \frac{i}{2} \sin \boldsymbol{\theta}_j [[H_j, O']\rho'_{\text{in}}] + \sin^2 \frac{\boldsymbol{\theta}_j}{2} \text{Tr} [H_j O' H_j \rho'_{\text{in}}], \end{aligned} \tag{4.22}$$

where  $[A, B] := AB - BA$  denotes the commutator.

After some calculations from Eqn. (4.22), it can be shown that

$$\begin{aligned} f\left(\boldsymbol{\theta} + \frac{\pi}{2}\mathbf{e}^{(j)}\right) &= \frac{1 - \sin \boldsymbol{\theta}_j}{2} \text{Tr}[O' \rho'_{\text{in}}] + \frac{i}{2} \cos \boldsymbol{\theta}_j [[H_j, O'] \rho'_{\text{in}}] \\ &\quad + \frac{1 + \sin \boldsymbol{\theta}_j}{2} \text{Tr}[H_j O' H_j \rho'_{\text{in}}] \\ f\left(\boldsymbol{\theta} - \frac{\pi}{2}\mathbf{e}^{(j)}\right) &= \frac{1 + \sin \boldsymbol{\theta}_j}{2} \text{Tr}[O' \rho'_{\text{in}}] - \frac{i}{2} \cos \boldsymbol{\theta}_j [[H_j, O'] \rho'_{\text{in}}] \\ &\quad + \frac{1 - \sin \boldsymbol{\theta}_j}{2} \text{Tr}[H_j O' H_j \rho'_{\text{in}}] \\ \frac{\partial f}{\partial \theta_j}(\boldsymbol{\theta}) &= -\frac{1}{2} \sin \boldsymbol{\theta}_j \text{Tr}[O' \rho'_{\text{in}}] + \frac{i}{2} \cos \boldsymbol{\theta}_j [[H_j, O'] \rho'_{\text{in}}] \\ &\quad + \frac{1}{2} \sin \boldsymbol{\theta}_j \text{Tr}[H_j O' H_j \rho'_{\text{in}}]. \end{aligned}$$

Comparing the above equations, Theorem 4.9 is proved.  $\square$

### 4.3.2 Discriminative learning with QNNs

In this section, we present an example in which a QNN is employed for discriminative learning. Specifically, we focus on binary classification, where the label  $y^{(i)} = \pm 1$  corresponds to the input state  $\rho^{(i)}$ . In the case of classical data, the state  $\rho^{(i)} = |\psi(\mathbf{x}^{(i)})\rangle\langle\psi(\mathbf{x}^{(i)})|$  can be generated from the classical vector  $\mathbf{x}^{(i)}$  using a read-in approach

$$|\psi(\mathbf{x}^{(i)})\rangle = U_\phi(\mathbf{x}^{(i)})|0\rangle, \quad (4.23)$$

where a simple feature map can be constructed via angle encoding, as introduced in Chapter 2.3.1,

$$U_\phi(\mathbf{x}^{(i)}) = \bigotimes_{n=1}^N \text{RY}(x_n^{(i)}) = \bigotimes_{n=1}^N \exp(-iY x_n^{(i)}/2). \quad (4.24)$$

Denote by  $O$  and  $V(\boldsymbol{\theta})$  the quantum observable and the VQC, respectively. The prediction function of the QNN is given by

$$\hat{y}^{(i)}(\boldsymbol{\theta}) = \text{Tr}[OV(\boldsymbol{\theta})\rho^{(i)}V(\boldsymbol{\theta})^\dagger]. \quad (4.25)$$

In the binary classification task, the QNN learns by training the parameter  $\boldsymbol{\theta}$  to minimize the distance between the label  $y^{(i)}$  and the prediction

$\hat{y}^{(i)}(\boldsymbol{\theta})$ . Specifically, the mean square error (MSE) is used as the loss function:

$$\boldsymbol{\theta}^* = \operatorname{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}), \text{ where } \mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^n \ell(\boldsymbol{\theta}, \mathbf{x}^{(i)}, y^{(i)}) = \frac{1}{2} \sum_{i=1}^n \left( \hat{y}^{(i)}(\boldsymbol{\theta}) - y^{(i)} \right)^2. \quad (4.26)$$

The gradient of the loss in Eqn. (4.26) can be calculated via the chain rule, i.e.,

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^n \left( \hat{y}^{(i)}(\boldsymbol{\theta}) - y^{(i)} \right) \nabla_{\boldsymbol{\theta}} \hat{y}^{(i)}(\boldsymbol{\theta}), \quad (4.27)$$

where the gradient of the prediction  $\hat{y}^{(i)}$  can be obtained by using the parameter-shift rule in Theorem 4.9. Consequently, a variety of gradient-based optimization algorithms, such as stochastic gradient descent (Amari, 1993), Adagrad (Duchi et al., 2011), and Adam (Kingma, 2014), can be employed to train QNNs.

#### Remark

The QNN binary classification framework can be naturally extended to multi-label classification using the **one-vs-all** strategy. Specifically, we train  $k$  QNN binary classifiers for  $k$  classes, with each classifier distinguishing a specific class from the others.

#### Remark

The QNN classification framework presented in this section can be extended to quantum regression learning by incorporating continuous labels.

### 4.3.3 Generative learning with QNNs

In this section, we introduce a quantum generative model implemented by QNNs, namely quantum generative adversarial network (QGAN) (Lloyd and Weedbrook, 2018). Similar to its classical counterparts, QGAN learns to generate samples by employing a discriminator and a generator, which are engaged in a two-player minimax game. Specifically, both the discriminator  $D$  and the generator  $G$  can be implemented using QNNs. By leveraging the expressive power of QNNs, QGAN has the potential to exhibit quantum advantages in certain tasks (Bravyi et al., 2018; Zhu et al., 2022).

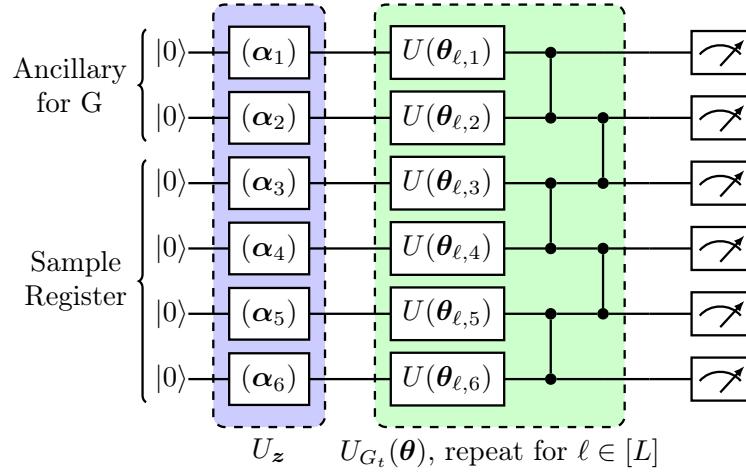


Figure 4.7: **The quantum generator used in the quantum patch GAN**, where each  $U(\theta_{\ell,n}) \in \mathcal{U}(2)$  is a trainable single-qubit unitary.

To illustrate the training and sampling processes of QGAN, we present two examples based on the quantum patch and batch GANs proposed by Huang et al. (2021b). Let  $N$  denote the number of qubits and  $M$  the number of training samples. The patch and batch strategies are designed for the cases where  $N < \lceil \log M \rceil$  and  $N > \lceil \log M \rceil$ , respectively. Specifically, the patch strategy enables the generation of high-dimensional images with limited quantum resources, while the batch strategy facilitates parallel training when sufficient quantum resources are available.

### Quantum patch GAN

We begin by introducing the quantum patch GAN, which consists of a quantum generator, as illustrated in Figure 4.7, a classical discriminator, and a classical optimizer. Both the learning and sampling processes of an image are performed in patches, involving  $T$  sub-generators. For the  $t$ -th sub-generator, the model takes a latent state  $\mathbf{z}$  as input and generates a sample  $G_t(\mathbf{z})$ . Specifically, the latent state is prepared from the initial state  $|0\rangle^{\otimes N}$  using a single-qubit rotation layer, where the parameters  $\{\alpha_n\}_{n=1}^N$  are sampled from the uniform distribution over  $[0, 2\pi)$ . The latent state is then processed through an  $N$ -qubit hardware-efficient circuit  $U_{G_t}(\boldsymbol{\theta})$ , which leads to the state

$$|\psi_t(\mathbf{z})\rangle = U_{G_t}(\boldsymbol{\theta})|\mathbf{z}\rangle. \quad (4.28)$$

To perform non-linear operations, partial measurements are conducted, and a subsystem  $\mathcal{A}$  (ancillary qubits) is traced out from the state  $|\psi_t(\mathbf{z})\rangle$ . The resulting mixed state is

$$\rho_t(\mathbf{z}) = \frac{\text{Tr}_{\mathcal{A}}[\Pi \otimes \mathbb{I} |\psi_t(\mathbf{z})\rangle\langle\psi_t(\mathbf{z})|]}{\text{Tr}[\Pi \otimes \mathbb{I} |\psi_t(\mathbf{z})\rangle\langle\psi_t(\mathbf{z})|]}, \quad (4.29)$$

where  $\Pi$  is the projective operator acting on the subsystem  $\mathcal{A}$ . Subsequently, the mixed state  $\rho_t(\mathbf{z})$  is measured in the computational basis to obtain the sample  $G_t(\mathbf{z})$ . Specifically, let  $\Pr(J = j) := \text{Tr}[|j\rangle\langle j|\rho_t(\mathbf{z})]$ , where the probabilities of the outcomes can be estimated by the measurement. The sample  $G_t(\mathbf{z})$  is then defined as

$$G_t(\mathbf{z}) = [\Pr(J = 0), \dots, \Pr(J = j), \dots, \Pr(J = 2^{N-N_{\mathcal{A}}} - 1)], \quad (4.30)$$

where  $N_{\mathcal{A}}$  is the number of qubits in  $\mathcal{A}$ . Finally, the complete image is reconstructed by aggregating these samples from all sub-generators as follows:

$$G(\mathbf{z}) = [G_1(\mathbf{z}), \dots, G_T(\mathbf{z})]. \quad (4.31)$$

In principle, the discriminator  $D$  in a quantum patch GAN can be any classical neural network that takes the training data  $\mathbf{x}$  or the generated sample  $G(\mathbf{z})$  as input, with the output

$$D(\mathbf{x}), D(G(\mathbf{z})) \in [0, 1]. \quad (4.32)$$

Let  $\boldsymbol{\gamma}$  and  $\boldsymbol{\theta}$  denote the parameters of the discriminator  $D$  and the generator  $G$ , respectively. The optimization problem for the quantum patch GAN can be formulated as:

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\gamma}} \mathcal{L}(D_{\boldsymbol{\gamma}}(G_{\boldsymbol{\theta}}(\mathbf{z})), D_{\boldsymbol{\gamma}}(\mathbf{x})) := \mathbb{E}_{\mathbf{x}}[\log D_{\boldsymbol{\gamma}}(\mathbf{x})] + \mathbb{E}_{\mathbf{z}}[\log(1 - D_{\boldsymbol{\gamma}}(G_{\boldsymbol{\theta}}(\mathbf{z})))]. \quad (4.33)$$

Similar to quantum discriminative learning, the quantum patch GAN can be trained using gradient-based optimization algorithms.

### Quantum batch GAN

As illustrated in Figure 4.8, the quantum batch GAN differs from the quantum patch GAN by employing a quantum discriminator. In a quantum batch GAN, all qubits are divided into two registers: the index register, consisting of  $N_I$  qubits, and the feature register, consisting of  $N_F$  qubits.

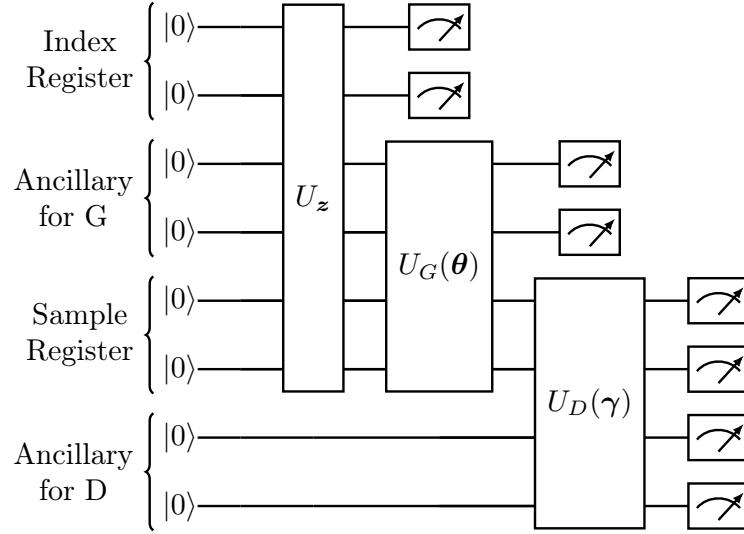


Figure 4.8: **The main structure of the quantum batch GAN.** The figure illustrates the process of generation and training using fake data. The oracle  $U_z$  for encoding latent vectors, the quantum generator  $U_G(\theta)$ , and the quantum discriminator  $U_D(\gamma)$  are applied sequentially to the initial state  $|0\rangle^{\otimes N}$ . Both  $U_G$  and  $U_D$  share the same hardware-efficient structure as shown in Figure 4.7. In the case of real data, the operations  $U_z$  and  $U_G(\theta)$  are replaced by the oracle  $U_x$ .

The qubits in the feature register are further partitioned into three parts:  $N_D$  qubits for generating quantum samples,  $N_{A_G}$  qubits for implementing non-linear operations in the generator  $G_\theta$ , and  $N_{A_D}$  qubits for implementing non-linear operations in the discriminator  $D_\gamma$ . For a batch with size  $|B_k| = 2^{N_I}$ , two oracles are used to encode the information of latent vectors and training samples:

$$|0\rangle_I \otimes |0\rangle_F \xrightarrow{U_z} \frac{1}{2^{N_I}} \sum_i |i\rangle_I \otimes |\mathbf{z}^{(i)}\rangle_F, \quad (4.34)$$

$$|0\rangle_I \otimes |0\rangle_F \xrightarrow{U_x} \frac{1}{2^{N_I}} \sum_i |i\rangle_I \otimes |\mathbf{x}^{(i)}\rangle_F. \quad (4.35)$$

## Remark

For data with  $M$  features, state preparation for amplitude encoding in  $U_{\mathbf{x}}$  requires  $\tilde{\mathcal{O}}(2^{N_I} M)$  multi-controlled quantum gates, which is infeasible for current quantum devices. This challenge can be addressed by employing pre-trained shallow circuit approximations of the given oracle (Benedetti et al., 2019a).

After the encoding stage, a PQC  $U_G(\boldsymbol{\theta})$  and the corresponding partial measurement are employed as the quantum generator. Thus, the generated state corresponding to  $|B_k|$  fake samples is obtained as follows:

$$\begin{aligned} & \frac{1}{2^{N_I}} \sum_i |i\rangle_I \otimes |\mathbf{z}^{(i)}\rangle_F \\ \xrightarrow{U_G(\boldsymbol{\theta})} & \frac{1}{2^{N_I}} \sum_i |i\rangle_I \otimes \left( U_G(\boldsymbol{\theta}) \otimes \mathbb{I}_{2^{N_{A_D}}} |\mathbf{z}^{(i)}\rangle_F \right) := |\psi(\mathbf{z})\rangle \\ \xrightarrow{\Pi_{A_G}} & \frac{\mathbb{I}_{2^{N_I}} \otimes \Pi_{A_G} \otimes \mathbb{I}_{2^{N_D+N_{A_D}}} |\psi(\mathbf{z})\rangle}{\text{Tr}[\mathbb{I}_{2^{N_I}} \otimes \Pi_{A_G} \otimes \mathbb{I}_{2^{N_D+N_{A_D}}} |\psi(\mathbf{z})\rangle\langle\psi(\mathbf{z})|]} := |G_{\boldsymbol{\theta}}(\mathbf{z})\rangle, \end{aligned}$$

where the partial measurement  $\Pi_{A_G} = (|0\rangle\langle 0|)^{\otimes N_{A_G}}$  serves as the non-linear operation. In the sampling stage, the reconstructed image is generated similarly to the quantum patch GAN. Specifically, the  $i$ -th image  $G_{\boldsymbol{\theta}}(\mathbf{z}^{(i)})$  in the batch is

$$G_{\boldsymbol{\theta}}(\mathbf{z}^{(i)}) = [\Pr(J=0|I=i), \dots, \Pr(J=2^{N_D}-1|I=i)], \quad (4.36)$$

where

$$\Pr(J=j|I=i) = \text{Tr}[|i\rangle_I |j\rangle_F \langle i|_I \langle j|_F |G(\mathbf{z})\rangle \langle G(\mathbf{z})|]. \quad (4.37)$$

Finally, we introduce the training stage. A quantum discriminator is applied to either the fake generated state  $|G_{\boldsymbol{\theta}}(\mathbf{z})\rangle$  or the real data state  $|\mathbf{x}\rangle$ . Similar to the quantum generator, the quantum discriminator  $D_{\gamma}$  consists of a PQC  $U_D(\gamma)$ , followed by the corresponding partial measurement. In

the case of the real state, the state evolution proceeds as follows:

$$\begin{aligned} & \frac{1}{2^{N_I}} \sum_i |i\rangle_I \otimes |\boldsymbol{x}^{(i)}\rangle_F \\ \xrightarrow{U_D(\gamma)} & \frac{1}{2^{N_I}} \sum_i |i\rangle_I \otimes \left( \mathbb{I}_{2^{N_{A_G}}} \otimes U_D(\gamma) |\boldsymbol{x}^{(i)}\rangle_F \right) := |\psi(\boldsymbol{x})\rangle \\ \xrightarrow{\Pi_{A_D}} & \frac{\mathbb{I}_{2^{N-N_{A_D}}} \otimes \Pi_{A_G} |\psi(\boldsymbol{x})\rangle}{\text{Tr} \left[ \mathbb{I}_{2^{N-N_{A_D}}} \otimes \Pi_{A_G} |\psi(\boldsymbol{x})\rangle \langle \psi(\boldsymbol{x})| \right]} := |D_\gamma(\boldsymbol{x})\rangle, \end{aligned}$$

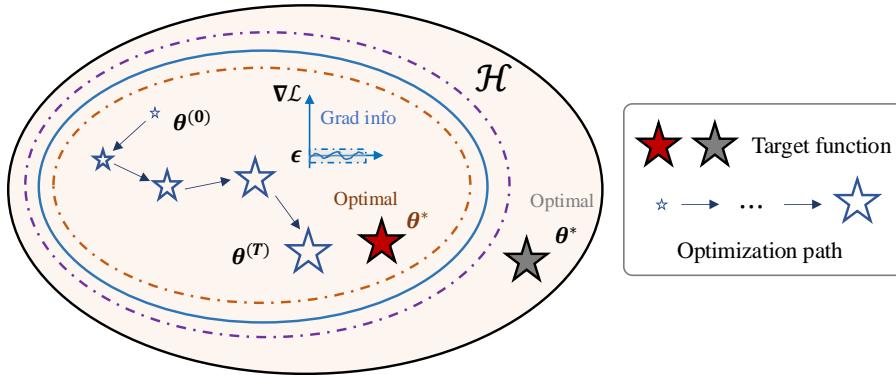
where the partial measurement is  $\Pi_{A_G} = (|0\rangle\langle 0|)^{\otimes N_{A_D}}$ . The classical description  $D_\gamma(\boldsymbol{x})$  is generated similarly to Eqn. (4.36). The generated state  $G_\theta|\boldsymbol{z}\rangle$  undergoes the same procedure to obtain the description  $D_\gamma(G_\theta(\boldsymbol{z}))$ . These classical vectors are then used in the loss function in Eqn. (4.33) to train parameters  $\boldsymbol{\theta}$  and  $\gamma$ .

## 4.4 Theoretical Foundations of Quantum Neural Networks

The primary goal of QNNs is to make accurate predictions on unseen data. Achieving this goal depends on three key factors: expressivity, generalization ability, and trainability, as illustrated in Figure 4.9. A thorough analysis of these factors is crucial for understanding the potential advantages and limitations of QNNs compared to classical counterparts. Instead of providing an exhaustive review of all theoretical results, this section focuses on emphasizing key conceptual insights of QNNs.

As explained in Chapter 3.3, expressivity refers to a model's ability to represent a wide range of functions, determining the smallest achievable training error. In Chapter 4.4.1, we will characterize the expressivity of QNNs using the *covering number*, an advanced tool from statistical learning theory. This analysis will reveal the relationship between the expressivity of QNNs and their structural factors, such as the size of the quantum system and the number of exploited quantum gates. Understanding this connection helps clarify how QNNs' expressivity scales with their architecture.

Generalization ability evaluates the discrepancy between a model's performance on the training data and on unseen test data. In Chapter 4.4.1, we will further explore the relationship between the generalization ability and expressivity of QNNs by deriving a generalization error bound in terms of the covering number. This bound provides insights into how the expressivity of QNNs—specifically their structural factors—may impact their ability to



**Figure 4.9: Overview of the expressivity, generalization ability, and trainability of QNNs.** The expressivity of the employed QNNs determines its hypothesis space  $\mathcal{H}$  (solid blue ellipse). When  $\mathcal{H}$  has a moderate size and encompasses the target concept (solid red star), QNNs can achieve good performance. Conversely, if  $\mathcal{H}$  fails to cover the target concept (solid gray star) due to limited expressivity, the performance of QNNs deteriorates. During QNN optimization, a significant challenge arises from the vanishing gradient problem, commonly referred to as the barren plateau. This issue prohibits a good estimation near the target parameters  $\theta^*$ .

generalize and offers a framework to assess their potential advantages over classical ML models.

While expressivity and generalization ability are crucial for both quantum kernels and QNNs, trainability emerges as an additional consideration for QNNs due to the introduction of trainable parameters in quantum circuits. This leads to fundamentally different optimization challenges, where many existing results from classical ML models no longer apply. Specifically, trainability refers to a model's ability to efficiently converge to a good solution during training, directly influencing the computational cost of training. In Chapter 4.4.2, we will introduce a well-known challenge in training QNNs, referred to as the barren plateau problem, where gradients vanish exponentially as the system size increases, making optimization intractable. Additionally, we will discuss various strategies to address this issue, offering practical insights into enhancing the trainability of QNNs.

#### 4.4.1 Expressivity and generalization of quantum neural networks

The expressivity and generalization are deeply interconnected within the framework of statistical learning theory for understanding the prediction ability of any learning model. To better understand these two terms in the context of quantum neural networks, we first review the framework of empirical risk minimization (ERM), which is a popular framework for analyzing these abilities in statistical learning theory.

Consider the training dataset  $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n \in \mathcal{X} \times \mathcal{Y}$  sampled independently from an unknown distribution  $\mathcal{P}$ , a learning algorithm  $\mathcal{A}$  aims to use the dataset  $\mathcal{D}$  to infer a hypothesis  $h_{\theta^*} : \mathcal{X} \rightarrow \mathcal{Y}$  from the hypothesis space  $\mathcal{H}$  that could accurately predict all labels of  $\mathbf{x} \in \mathcal{X}$  following the distribution  $\mathcal{P}$ . This amounts to identifying an optimal hypothesis in  $\mathcal{H}$  minimizing the expected risk

$$R(h) = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{P}} \ell(h_{\theta^*}(\mathbf{x}), y), \quad (4.38)$$

where  $\ell(\cdot, \cdot)$  refers to the per-sample loss predefined by the learner. Unfortunately, the inaccessible distribution  $\mathcal{P}$  forbids us to assess the expected risk directly. In practice,  $\mathcal{A}$  alternatively learns an empirical hypothesis  $h_{\hat{\theta}} \in \mathcal{H}$ , as the global minimizer of the (regularized) loss function

$$\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}) = \frac{1}{n} \sum_{i=1}^n \ell(h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}), y^{(i)}) + \mathcal{R}(\boldsymbol{\theta}), \quad (4.39)$$

where  $\mathcal{R}(\boldsymbol{\theta})$  refers to an optional regularizer, as will be detailed in the following. Moreover, the first term on the right-hand side refers to the empirical risk

$$R_{\text{ERM}}(h_{\hat{\theta}}) = \frac{1}{n} \sum_{i=1}^n \ell(h_{\hat{\theta}}(\mathbf{x}^{(i)}), y^{(i)}), \quad (4.40)$$

which is also known as the training error. To address the intractability of  $R(h_{\hat{\theta}})$ , one can decompose it into two measurable terms,

$$R(h_{\hat{\theta}}) = R_{\text{ERM}}(h_{\hat{\theta}}) + R_{\text{Gene}}(h_{\hat{\theta}}), \quad (4.41)$$

where  $R_{\text{Gene}}(h_{\hat{\theta}}) = R(h_{\hat{\theta}}) - R_{\text{ERM}}(h_{\hat{\theta}})$  refers to the generalization error. In this regard, achieving a small prediction error requires the learning model to achieve both a small training error and a small generalization error.

### An overview

Before moving to analyze the training error (ERM) and generalization error of QNNs rigorously, we first delve into better understanding the meaning of expressivity and generalization ability of the learning models with the ERM framework and try to give an intuition about the necessities and benefits of exploring such theoretical aspects of QNNs as a special learning model.

In particular, the expressivity could be directly understood as the size of the hypothesis space  $\mathcal{H} = \{h_{\theta} : \theta \in \Theta\}$  related to the learning model. Intuitively, the achievable smallest empirical risk is determined by the expressivity of learning models. Specifically, a learning model with low expressivity may not fit the training data with complex patterns, e.g., the hypothesis space of linear model  $\mathcal{H} = \{h_{\theta} = \theta \cdot x\}$  cannot fit the nonlinear data  $\{\mathbf{x}^{(i)}, (\mathbf{x}^{(i)})^2\}$  perfectly.

In general, the cardinality of the hypothesis space is infinity, as the parameters  $\theta$  are continuous. This makes it hard to compare the expressivity of different learning models. An alternative measure is model complexity, which measures the richness of the hypothesis space through the structural factors of the specific learning models, such as the number of parameters, depth, or architectural design. Remarkably, model complexity is measurable and bounded. In this tutorial, we will employ the covering number to measure the model complexity of QNNs later.

The generalization capability of learning models is directly measured by the generalization error  $R_{\text{Gene}}$  in Eqn. (4.41). A good generalization ability means that the learning model predicts well on unseen data as well as on the training data. In this regard, a small generalization error with a small training error implicates a small prediction error, as a small generalization error guarantees that the prediction performance is as well as the training performance.

In statistical learning theory, it has been well-established that a bias-variance trade-off governs the interplay between model complexity and generalization performance for any learning model, highlighting the delicate balance required for a model to generalize well to unseen data. The relationship is often depicted by a U-shaped curve, as shown in Figure 4.10. This curve suggests that there exists an optimal level of model complexity for improving the generalization ability of any learning model. When under the point related to optimal expressivity, increasing model complexity improves performance on training data and enhances generalization. However, beyond a certain point, higher complexity leads to overfitting, resulting in poor generalization on test data. For QNNs, identifying this optimal level

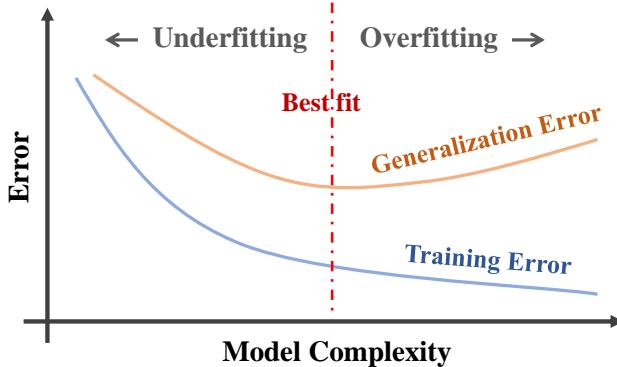


Figure 4.10: Influence of model complexity on generalization error.

of complexity is crucial for achieving the best balance between training performance and generalization.

### Expressivity of QNNs

In this chapter, we analyze the generalization error of QNNs through a specific measure of model complexity: the covering number. By leveraging this measure, we aim to understand better and characterize the generalization performance of QNNs.

To elucidate the specific definition of the covering number, we first review the general structures of QNNs. Define  $\rho \in \mathbb{C}^{2^N \times 2^N}$  as the  $N$ -qubit input quantum states,  $O \in \mathbb{C}^{2^N \times 2^N}$  as the quantum observable,  $U(\boldsymbol{\theta}) = \prod_{l=1}^{N_g} u_l(\boldsymbol{\theta}) \in \mathcal{U}(2^N)$  as the applied ansatz, where  $\boldsymbol{\theta} \in \Theta$  are the trainable parameters living in the parameter space  $\Theta$ ,  $u_l(\boldsymbol{\theta}) \in \mathcal{U}(2^k)$  refers to the  $l$ -th quantum gate operated with at most  $k$ -qubits with  $k \leq N$ , and  $\mathcal{U}(2^N)$  stands for the unitary group in dimension  $2^N$ . In general,  $U(\boldsymbol{\theta})$  is formed by  $N_{gt}$  trainable gates and  $N_g - N_{gt}$  fixed gates, e.g.,  $\Theta \subset [0, 2\pi)^{N_{gt}}$ . Under the above definitions, the explicit form of the output of QNN under the ideal scenarios is

$$h(\boldsymbol{\theta}, O, \rho) := \text{Tr}\left(U(\boldsymbol{\theta})^\dagger O U(\boldsymbol{\theta}) \rho\right). \quad (4.42)$$

Given the training data set  $\mathcal{D} = \{(\rho^{(i)}, y^{(i)})\}_{i=1}^n$  and loss function  $\mathcal{L}(\boldsymbol{\theta}, \mathcal{D})$  defined in Eqn. (4.39), QNN is optimized to find a good approximation  $h^*(\boldsymbol{\theta}, O, \rho) = \arg \min_{h(\boldsymbol{\theta}, O, \rho) \in \mathcal{H}} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D})$  that can well approximate the target concept, where  $\mathcal{H}$  refers to the hypothesis space of QNNs with

$$\mathcal{H} = \left\{ \text{Tr}\left(U(\boldsymbol{\theta})^\dagger O U(\boldsymbol{\theta}) \rho\right) \mid \boldsymbol{\theta} \in \Theta \right\}. \quad (4.43)$$

An intuition about how the hypothesis space  $\mathcal{H}$  affects the performance of QNNs is depicted in Figure 4.9. When  $\mathcal{H}$  has a modest size and covers the target concepts, the estimated hypothesis could well approximate the target concept. By contrast, when the complexity of  $\mathcal{H}$  is too low, there exists a large gap between the estimated hypothesis and the target concept. An effective measure to evaluate the complexity of  $\mathcal{H}$  is covering number, an advanced tool broadly used in statistical learning theory, to bound the complexity of  $\mathcal{H}$  and measure the expressivity of QNNs.

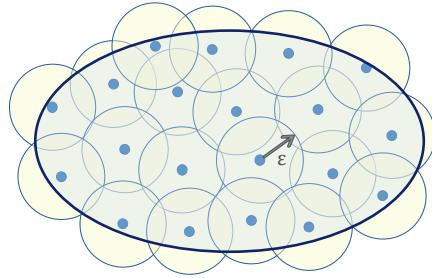


Figure 4.11: **The geometric intuition of covering number.** Covering number concerns the minimum number of spherical balls with radius  $\epsilon$  that occupy the whole space.

**Definition 4.10** (Covering number). *The covering number  $\mathcal{N}(\mathcal{U}, \epsilon, \|\cdot\|)$  denotes the least cardinality of any subset  $V \subset \mathcal{U}$  that covers  $\mathcal{U}$  at scale  $\epsilon$  with a norm  $\|\cdot\|$ , i.e.,  $\sup_{A \in \mathcal{U}} \min_{B \in V} \|A - B\| \leq \epsilon$ . Here we use this notion to measure the expressivity of QNNs.*

The geometric interpretation of the covering number is depicted in Figure 4.11, which refers to the minimum number of spherical balls with radius  $\epsilon$  that are required to completely cover a given space with possible overlaps. This notion has been employed to study other crucial topics in quantum physics such as Hamiltonian simulation and entangled states. Note that  $\epsilon$  is a predefined hyper-parameter, i.e., a small constant with  $\epsilon \in (0, 1)$ , and is independent of any factor. This convention has been broadly adopted in the regime of machine learning to evaluate the model capacity of various learning models.

Following the convention of Du et al. (2022c), we now give a step-by-step analysis of the model complexity of the hypothesis space  $\mathcal{H}$  of QNNs defined in Eqn. (4.43). We will show that the covering number of QNNs is controlled by their structural factors, including the number of parameterized

gates  $N_{gt}$ , the number of qubits  $k$  the gates acting on, and the type of the quantum observable  $O$ . In the end, we first look at a simpler hypothesis space consisting of the operator group

$$\mathcal{H}_{\text{circ}} := \left\{ U(\boldsymbol{\theta})^\dagger O U(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta \right\}, \quad (4.44)$$

where we remove the factor of input states  $\rho$  compared to the hypothesis space  $\mathcal{H}$  related to QNNs. Actually, the covering number of  $\mathcal{H}$  under the metric  $d$  could be connected to the covering number of  $\mathcal{H}_{\text{circ}}$  under the related metric  $d_{\text{circ}}$  through employing their Lipschitz properties, which is encapsulated in the following Fact.

**Fact 4.11.** *Let  $(\mathcal{H}_1, d_1)$  and  $(\mathcal{H}_2, d_2)$  be two metric spaces satisfying  $f : \mathcal{H}_1 \rightarrow \mathcal{H}_2$  be bi-Lipschitz such that*

$$c_l d_1(\mathbf{x}, \mathbf{z}) \leq d_2(f(\mathbf{x}), f(\mathbf{z})) \leq c_r d_1(\mathbf{x}, \mathbf{z}), \quad \forall \mathbf{x}, \mathbf{z} \in \mathcal{H}_1. \quad (4.45)$$

*Then their covering number obey*

$$\mathcal{N}(\mathcal{H}_1, 2\epsilon/c_l, d_1) \leq \mathcal{N}(\mathcal{H}_2, \epsilon, d_2) \leq \mathcal{N}(\mathcal{H}_1, \epsilon/c_r, d_1), \quad (4.46)$$

*where the left inequality requires  $\epsilon \leq c_l c_u / 2$  with  $c_u$  being the upper bound of the distance between any two points in  $\mathcal{H}_1$ , namely,  $d_1(\mathbf{x}, \mathbf{z}) \leq c_u$  for  $\mathbf{x}, \mathbf{z} \in \mathcal{H}_1$ .*

Fact 4.11 indicates that we can derive the covering number of the metric space  $(\mathcal{H}, d)$  by analyzing the covering number of the metric space  $(\mathcal{H}_{\text{circ}}, d_{\text{circ}})$  and the Lipschitz constants of the mapping between  $\mathcal{H}$  and  $\mathcal{H}_{\text{circ}}$ . Intuitively, a quantum circuit consisting of a large number of multi-qubit parameterized gates leads to a complicated QNN with a large model complexity. These intuitions are formalized into Theorem 4.13. Specifically, the result of the covering number of the metric space  $(\mathcal{H}_{\text{circ}}, d_{\text{circ}})$  is encapsulated in Lemma 4.12.

**Lemma 4.12.** *Suppose that the employed  $N$ -qubit quantum circuit containing in total  $N_g$  gates with  $N_g > N$ , each gate  $u_i(\boldsymbol{\theta})$  acting on most  $k$  qubits, and  $N_{gt} \leq N_g$  gates in  $U(\boldsymbol{\theta})$  are trainable. The  $\epsilon$ -covering number for the operator group  $\mathcal{H}_{\text{circ}}$  in Eqn. (4.44) with respect to the operator-norm distance obeys*

$$\mathcal{N}(\mathcal{H}_{\text{circ}}, \epsilon, \|\cdot\|) \leq \left( \frac{7N_{gt}\|O\|}{\epsilon} \right)^{2^{2kN_{gt}}}, \quad (4.47)$$

*where  $\|O\|$  denotes the operator norm of  $O$ .*

*Proof of Lemma 4.12.* To measure the covering number the operator group of  $\mathcal{H}_{\text{circ}} = \{U(\boldsymbol{\theta})^\dagger O U(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta\}$ , one could first consider a fixed  $\epsilon$ -covering  $\mathcal{S}$  for the set  $\mathcal{N}(U(2^k), \epsilon, \|\cdot\|)$  of all possible gates and define the set

$$\tilde{\mathcal{S}} := \left\{ \prod_{i \in \{N_{gt}\}} u_i(\boldsymbol{\theta}_i) \prod_{j \in \{N_g - N_{gt}\}} u_j \mid u_i(\boldsymbol{\theta}_i) \in \mathcal{S} \right\}, \quad (4.48)$$

where  $u_i(\boldsymbol{\theta}_i)$  and  $u_j$  specify the trainable and fixed quantum gates in the employed quantum circuit, respectively. Note that for any circuit  $U(\boldsymbol{\theta}) = \prod_{i=1}^{N_g} u_i(\boldsymbol{\theta}_i)$ , one can always find a  $U_\epsilon(\boldsymbol{\theta}) \in \tilde{\mathcal{S}}$  where each  $u_i(\boldsymbol{\theta}_i)$  of trainable gates is replaced with the nearest element in the covering set  $\mathcal{S}$ , and the discrepancy  $\|U(\boldsymbol{\theta})^\dagger O U(\boldsymbol{\theta}) - U_\epsilon(\boldsymbol{\theta})^\dagger O U_\epsilon(\boldsymbol{\theta})\|$  satisfies

$$\begin{aligned} & \|U(\boldsymbol{\theta})^\dagger O U(\boldsymbol{\theta}) - U_\epsilon(\boldsymbol{\theta})^\dagger O U_\epsilon(\boldsymbol{\theta})\| \\ & \leq \|U - U_\epsilon\| \|O\| \\ & \leq N_{gt} \|O\| \epsilon, \end{aligned} \quad (4.49)$$

where the first inequality uses the triangle inequality, and the second inequality follows from  $\|U - U_\epsilon\| \leq N_{gt} \epsilon$ .

Therefore, by Definition 4.10,  $\tilde{\mathcal{S}}$  forms an  $N_{gt} \|O\| \epsilon$ -covering set for  $\mathcal{H}_{\text{circ}}$ . An upper bound for the group  $\mathcal{S}$ , as established by Barthel and Lu (2018, Lemma 1), gives  $|\mathcal{S}| \leq \left(\frac{7}{\epsilon}\right)^{2^{2k}}$ . Since there are  $|\mathcal{S}|^{N_{gt}}$  combinations for the gates in  $\tilde{\mathcal{S}}$ , it follows that  $|\tilde{\mathcal{S}}| \leq \left(\frac{7}{\epsilon}\right)^{2^{2k} N_{gt}}$  and the covering number for  $\mathcal{H}_{\text{circ}}$  satisfies

$$\mathcal{N}(\mathcal{H}_{\text{circ}}, N_{gt} \|O\| \epsilon, \|\cdot\|) \leq \left(\frac{7}{\epsilon}\right)^{2^{2k} N_{gt}}. \quad (4.50)$$

An equivalent representation of the above inequality is

$$\mathcal{N}(\mathcal{H}_{\text{circ}}, \epsilon, \|\cdot\|) \leq \left(\frac{7N_{gt} \|O\|}{\epsilon}\right)^{2^{2k} N_{gt}}. \quad (4.51)$$

□

With the established covering number of operator group  $\mathcal{H}_{\text{circ}}$ , one could directly analyze the covering number of the hypothesis space  $\mathcal{H}$  related to QNNs, which is encapsulated in the following theorem.

**Theorem 4.13.** *For  $0 < \epsilon < 1/10$ , the covering number of the hypothesis space  $\mathcal{H}$  in Eqn. (4.43) yields*

$$\mathcal{N}(\mathcal{H}, \epsilon, |\cdot|) \leq \left(\frac{7N_{gt} \|O\|}{\epsilon}\right)^{2^{2k} N_{gt}}, \quad (4.52)$$

where  $\|O\|$  denotes the operator norm of  $O$ .

*Proof of Theorem 4.13.* The intuition of the proof is as follows. Recall the definition of the hypothesis space  $\mathcal{H}$  in Eqn. (4.43) and Lemma 4.11. When  $\mathcal{H}_1$  refers to the hypothesis space  $\mathcal{H}$  and  $\mathcal{H}_2$  refers to the unitary group  $\mathcal{U}(2^N)$ , the upper bound of the covering number of  $\mathcal{H}$ , i.e.,  $\mathcal{N}(\mathcal{H}_1, d_1, \epsilon)$ , can be derived by first quantifying  $c_r$  Eqn. (4.45), and then interacting with  $\mathcal{N}(\mathcal{H}_{\text{circ}}, \epsilon, \|\cdot\|)$  in Lemma 4.12. Based on the above observations, the following addresses the upper bound of the covering number  $\mathcal{N}(\mathcal{H}, \epsilon, |\cdot|)$ .

The Lipschitz constant  $c_r$  in Eqn. (4.45) is derived as a prerequisite for establishing the upper bound of  $\mathcal{N}(\mathcal{H}, \epsilon, |\cdot|)$ . Define  $U \in \mathcal{U}(2^N)$  as the employed quantum circuit composed of  $N_g$  gates, i.e.,  $U = \prod_{i=1}^{N_g} u_i$ . Let  $U_\epsilon$  be the quantum circuit where each of the  $N_g$  gates is replaced by the nearest element in the covering set. The relation between the distance  $d_2(\text{Tr}(U_\epsilon^\dagger O U_\epsilon \rho), \text{Tr}(U^\dagger O U \rho))$  and the distance  $d_1(U_\epsilon, U)$  yields

$$\begin{aligned} & d_2(\text{Tr}(U_\epsilon^\dagger O U_\epsilon \rho), \text{Tr}(U^\dagger O U \rho)) \\ &= |\text{Tr}(U_\epsilon^\dagger O U_\epsilon \rho) - \text{Tr}(U^\dagger O U \rho)| \\ &= |\text{Tr}((U_\epsilon^\dagger O U_\epsilon - U^\dagger O U)\rho)| \\ &\leq \|U_\epsilon^\dagger O U_\epsilon - U^\dagger O U\| \text{Tr}(\rho) \\ &= d_1(U_\epsilon^\dagger O U_\epsilon, U^\dagger O U), \end{aligned} \quad (4.53)$$

where the first equality comes from the explicit form of the hypothesis, the first inequality uses the Cauchy-Schwartz inequality, and the last inequality employs  $\text{Tr}(\rho) = 1$  and

$$\|U_\epsilon^\dagger O U_\epsilon - U^\dagger O U\| = d_1(U_\epsilon^\dagger O U_\epsilon, U^\dagger O U). \quad (4.54)$$

The above equation indicates  $c_r = 1$ . Combining the above result with Lemma 4.11 (i.e., Eqn. (4.45)) and Lemma 4.12, we obtain

$$\mathcal{N}(\mathcal{H}, \epsilon, |\cdot|) \leq \mathcal{N}(\mathcal{H}_{\text{circ}}, \epsilon, \|\cdot\|) \leq \left( \frac{7N_{gt}\|O\|}{\epsilon} \right)^{2^{2k}N_{gt}}. \quad (4.55)$$

This relation ensures

$$\mathcal{N}(\mathcal{H}, \epsilon, |\cdot|) \leq \left( \frac{7N_{gt}\|O\|}{\epsilon} \right)^{2^{2k}N_{gt}}. \quad (4.56)$$

□

Theorem 4.13 indicates that the most decisive factor, which controls the complexity of  $\mathcal{H}$ , is the employed quantum gates in  $U(\boldsymbol{\theta})$ . This claim is ensured by the fact that the term  $2^{2^k N_{gt}}$  exponentially scales the complexity  $\mathcal{N}(\mathcal{H}, \epsilon, |\cdot|)$ . Meanwhile, the qubits count  $N$  and the operator norm  $\|O\|$  polynomially scale the complexity of  $\mathcal{N}(\mathcal{H}, \epsilon, |\cdot|)$ . These observations suggest a succinct and direct way to compare the expressivity of QNNs with different quantum circuits. Moreover, the dependence of the expressivity of QNNs on the type of quantum gates (denoted by the term  $k$ ) demonstrated that the expressivity of QNNs depends on the structure information of ansatz such as the location of different quantum gates and the types of the employed quantum gates. The expressivity measured by the covering number could provide practical guidance for designing the circuit structure of QNNs.

### Generalization error of QNNs

As the relation between generalization error and covering number is well-established in statistical learning theory, we can directly obtain the generalization error bound with the above bounds of covering number following the same conventions.

**Theorem 4.14.** *Assume that the loss function  $\ell$  defined in Eqn. (4.38) is  $L_1$ -Lipschitz and upper bounded by a constant  $C$ , the QNN-based learning algorithm outputs a hypothesis  $h_{\hat{\boldsymbol{\theta}}}$  from the training dataset  $\mathcal{S}$  of size  $n$ . Following the notations of risk  $R_{\text{Gene}}(h_{\hat{\boldsymbol{\theta}}}) = R(h_{\hat{\boldsymbol{\theta}}}) - R_{\text{ERM}}(h_{\hat{\boldsymbol{\theta}}})$  defined in Eqn. (4.41), for  $0 < \epsilon < 1/10$ , with probability at least  $1 - \delta$  with  $\delta \in (0, 1)$ , we have*

$$R_{\text{Gene}}(h_{\hat{\boldsymbol{\theta}}}) \leq \mathcal{O}\left(\frac{8L + c + 24L\sqrt{N_{gt}} \cdot 2^k}{\sqrt{n}}\right). \quad (4.57)$$

*Proof sketch of Theorem 4.14.* Recall that the bound of generalization error in terms of Rademacher complexity has been established by Kakade et al. (2008) as follows

$$R_{\text{Gene}}(h_{\hat{\boldsymbol{\theta}}}) \leq 2L_1 \mathcal{R}(\mathcal{H}_{QNN}) + 3C \sqrt{\frac{\ln(2/\delta)}{2n}}, \quad (4.58)$$

where  $\mathcal{R}(\mathcal{H}_{QNN})$  represents the empirical Rademacher complexity of the hypothesis space of QNNs. Furthermore, the relationship between Rademacher complexity and covering number can be derived using the Dudley entropy integral bound Dudley (1967), which is given by

$$\mathcal{R}(\mathcal{H}) \leq \inf_{\alpha > 0} \left( 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^1 \sqrt{\ln \mathcal{N}(\mathcal{H}|_{\mathcal{S}}, \epsilon, \|\cdot\|_2)} d\epsilon \right), \quad (4.59)$$

where  $\mathcal{H}_{|\mathcal{S}}$  denotes the set of vectors formed by the hypothesis with  $n$  examples in the dataset  $\mathcal{S}$ . In this regard, the generalization error bound in Eqn. (4.57) could be obtained by combining the Eqn. (4.58) and Eqn. (4.59) with direct but tedious calculations, which is omitted here. For details of the calculations, please refer to the proof of Theorem 2 in Du et al. (2022c).  $\square$

The assumption used in this analysis is quite mild, as the loss functions in QNNs are generally Lipschitz continuous and can be bounded above by a constant  $C$ . This property has been broadly employed to understand the capability of QNNs. The results obtained have three key implications. First, the generalization bound exhibits an exponential dependence on the term  $k$  and a sublinear dependence on the number of trainable quantum gates  $N_{gt}$ . This observation reflects the quantum version of Occam’s razor (Haussler and Warmuth, 1987), where the parsimony of the output hypothesis implies greater predictive power. Second, increasing the number of training examples  $n$  improves the generalization bound. This suggests that incorporating more training data is essential for optimizing complex quantum circuits. Lastly, the sublinear dependence on  $N_{gt}$  may limit the ability to accurately assess the generalization performance of overparameterized QNNs (Larocca et al., 2023). Together, these implications provide valuable insights for designing more powerful QNNs.

#### 4.4.2 Trainability of quantum neural networks

The parameters in QNNs are often trained using gradient-based optimizers. As such, the magnitude of the gradient plays a crucial role in the trainability of QNNs. Specifically, large gradients are desirable, as they allow the loss function to decrease rapidly and consistently. However, this favorable property does not hold across a wide range of problem settings. In contrast, training QNNs usually encounters the barren plateau (BP) problem (McClean et al., 2018), i.e., *the variance of the gradient, on average, decreases exponentially as the number of qubits increases*. In this section, we first introduce an example demonstrating how quantum circuits that form unitary 2-designs (Dankert et al., 2009) lead to BP, and then discuss several techniques to avoid or mitigate this issue.

We begin by introducing some basic notations. For convenience, let  $L$  denote the number of parameters in the QNN  $V(\boldsymbol{\theta})$ . Consider the loss function defined as the measurement outcome of an  $N$ -qubit quantum state  $\rho$  after applying the QNN operation, i.e.,

$$f(\boldsymbol{\theta}) = \text{Tr} \left[ O V(\boldsymbol{\theta}) \rho V(\boldsymbol{\theta})^\dagger \right], \quad (4.60)$$

Then, the mathematical formulation of the BP phenomenon is given by

$$\mathbb{E}_{\mathcal{P}} \left[ \frac{\partial f}{\partial \theta_k} \right] = 0, \quad \text{Var}_{\mathcal{P}} \left[ \frac{\partial f}{\partial \theta_k} \right] = \exp(-\alpha N) \cdot \beta, \quad (4.61)$$

where  $\mathcal{P}$  represents the probability distribution of the quantum circuit, and  $\alpha, \beta > 0$  are constants. In the case where the circuit  $V(\boldsymbol{\theta})$  has a random structure with a polynomial number of single-qubit rotations and CNOT or CZ gates in  $N$ , a uniform distribution over the parameter space can approximate a 2-design for the unitary  $V(\boldsymbol{\theta})$  (Harrow and Low, 2009; Haferkamp, 2022). Moreover, a unitary sampled from an exact 2-design exhibits the following statistical properties.

**Fact 4.15** (Cerezo et al. (2021b)). *Let  $\{W_y\}_{y \in Y} \subset \mathcal{U}(d)$  form a unitary 2-design, and let  $A, B, C, D : \mathcal{H}_w \rightarrow \mathcal{H}_w$  be arbitrary linear operator. Then*

$$\frac{1}{|Y|} \sum_{y \in Y} \text{Tr}[W_y A W_y^\dagger B] = \frac{\text{Tr}[A] \text{Tr}[B]}{d}, \quad (4.62)$$

$$\begin{aligned} & \frac{1}{|Y|} \sum_{y \in Y} \text{Tr}[W_y A W_y^\dagger B] \text{Tr}[W_y C W_y^\dagger D] \\ &= \frac{1}{d^2 - 1} (\text{Tr}[AC] \text{Tr}[BD] + \text{Tr}[A] \text{Tr}[B] \text{Tr}[C] \text{Tr}[D]) \\ &\quad - \frac{1}{d(d^2 - 1)} (\text{Tr}[A] \text{Tr}[C] \text{Tr}[BD] + \text{Tr}[AC] \text{Tr}[B] \text{Tr}[D]), \end{aligned} \quad (4.63)$$

$$\begin{aligned} & \frac{1}{|Y|} \sum_{y \in Y} \text{Tr}[W_y A W_y^\dagger B W_y C W_y^\dagger D] \\ &= \frac{1}{d^2 - 1} (\text{Tr}[A] \text{Tr}[C] \text{Tr}[BD] + \text{Tr}[AC] \text{Tr}[B] \text{Tr}[D]) \\ &\quad - \frac{1}{d(d^2 - 1)} (\text{Tr}[AC] \text{Tr}[BD] + \text{Tr}[A] \text{Tr}[B] \text{Tr}[C] \text{Tr}[D]). \end{aligned} \quad (4.64)$$

Fact 4.15 can be derived from Facts C.4 and C.5 in Appendix C, which provides a more detailed discussion of unitary designs, potentially of independent interest. By applying Fact 4.15, it can be shown that QNNs with quantum circuits forming 2-designs exhibit barren plateau loss landscapes.

**Theorem 4.16** (Adapted from McClean et al. (2018)). *Consider the loss function given in Eqn. (4.60), where the QNN  $V(\boldsymbol{\theta}) = \prod_{j=1}^L V_j(\boldsymbol{\theta}_j) W_j$  with fixed gate  $W_j$  and variational gate  $V_j(\boldsymbol{\theta}_j) = \exp(-i\boldsymbol{\theta}_j H_j/2)$ . Suppose all hermitian matrices  $\{H_j\}$  are traceless. For a integer  $k \in [1, L]$ , denote*

$U_- = \prod_{j=1}^{k-1} V_j(\boldsymbol{\theta}_j) W_j$  and  $U_+ = \prod_{j=k+1}^L V_j(\boldsymbol{\theta}_j) W_j$ . Then, if both  $U_-$  and  $U_+$  form 2-designs, there is

$$\mathbb{E}\left[\frac{\partial f}{\partial \boldsymbol{\theta}_k}\right] = 0, \quad \text{Var}\left[\frac{\partial f}{\partial \boldsymbol{\theta}_k}\right] \approx \frac{1}{2^{3N+1}} \text{Tr}[O^2] \text{Tr}[\rho^2] \text{Tr}[H_j^2]. \quad (4.65)$$

*Proof of Theorem 4.16.* By using notations  $U_-$  and  $U_+$ , the function  $f(\boldsymbol{\theta})$  in Eqn. (4.60) can be expressed as:

$$\begin{aligned} f &= \text{Tr}\left[OV\rho V^\dagger\right] \\ &= \text{Tr}\left[OU_-V_kW_kU_+\rho U_+^\dagger W_k^\dagger V_k^\dagger U_-^\dagger\right] \\ &= \text{Tr}\left[U_-^\dagger OU_-V_kW_kU_+\rho U_+^\dagger W_k^\dagger V_k^\dagger\right] \\ &= \text{Tr}\left[O' \exp(-i\boldsymbol{\theta}_k H_k/2) \rho' \exp(i\boldsymbol{\theta}_k H_k/2)\right], \end{aligned}$$

where  $O' := U_-^\dagger O U_-$  and  $\rho' := W_k U_+ \rho U_+^\dagger W_k^\dagger$ . Thus, the gradient could be calculated as

$$\frac{\partial f}{\partial \boldsymbol{\theta}_k} = \frac{i}{2} \text{Tr}\left[O'\left[V_k \rho' V_k^\dagger, H_k\right]\right].$$

The expectation of the gradient is zero since

$$\begin{aligned} \mathbb{E}_{U_+, U_-} \frac{\partial f}{\partial \boldsymbol{\theta}_k} &= \mathbb{E} \frac{i}{2} \text{Tr}\left[O'\left[V_k \rho' V_k^\dagger, H_k\right]\right] \\ &= \mathbb{E}_{U_+, U_-} \frac{i}{2} \text{Tr}\left[U_-^\dagger O U_- \left[V_k \rho' V_k^\dagger, H_k\right]\right] \\ &= \mathbb{E}_{U_+} \frac{i}{2^{N+1}} \text{Tr}[O] \text{Tr}\left[\left[V_k \rho' V_k^\dagger, H_k\right]\right] \quad (4.66) \\ &= 0, \quad (4.67) \end{aligned}$$

where Eqn. (4.66) follows from Eqn. (4.62), and Eqn. (4.67) is derived by noticing  $\text{Tr}[[A, B]] = \text{Tr}[AB - BA] = 0$ . Therefore, the variance of the

gradient equals to the expectation of its square, i.e.,

$$\begin{aligned} \text{Var}_{U_+, U_-} \left[ \frac{\partial f}{\partial \theta_k} \right] &= \mathbb{E}_{U_+, U_-} \left[ \frac{\partial f}{\partial \theta_k} \right]^2 \\ &= -\frac{1}{4} \mathbb{E}_{U_+, U_-} \text{Tr} \left[ U_-^\dagger O U_- \left[ V_k \rho' V_k^\dagger, H_k \right] \right]^2 \\ &= -\frac{1}{4 \times (2^{2N} - 1)} \mathbb{E}_{U_+} \text{Tr}[O^2] \text{Tr} \left[ \left[ V_k \rho' V_k^\dagger, H_k \right]^2 \right] \\ &\quad + \frac{1}{2^{N+2}(2^{2N} - 1)} \mathbb{E}_{U_+} \text{Tr}[O]^2 \text{Tr} \left[ \left[ V_k \rho' V_k^\dagger, H_k \right]^2 \right] \quad (4.68) \\ &= -\frac{1}{4 \times (2^{2N} - 1)} \text{Tr}[O^2] \mathbb{E}_{U_+} \text{Tr} \left[ \left[ V_k \rho' V_k^\dagger, H_k \right]^2 \right], \quad (4.69) \end{aligned}$$

where Eqn. (4.68) follows from Eqn. (4.63), and Eqn. (4.69) follows from  $\text{Tr}[O] = 0$ . Further, it can be shown that

$$\begin{aligned} &\mathbb{E}_{U_+} \text{Tr} \left[ \left[ V_k \rho' V_k^\dagger, H_k \right]^2 \right] \\ &= 2 \mathbb{E}_{U_+} \text{Tr} \left[ \left( V_k \rho' V_k^\dagger H_k \right)^2 \right] - 2 \mathbb{E}_{U_+} \text{Tr} \left[ \left( V_k \rho' V_k^\dagger \right)^2 (H_k)^2 \right] \\ &= 2 \mathbb{E}_{U_+} \text{Tr} \left[ \left( V_k W_k U_+ \rho U_+^\dagger W_k^\dagger V_k^\dagger H_k \right)^2 \right] \\ &\quad - 2 \mathbb{E}_{U_+} \text{Tr} \left[ \left( V_k W_k U_+ \rho U_+^\dagger W_k^\dagger V_k^\dagger \right)^2 (H_k)^2 \right] \\ &= \frac{2}{2^{2N} - 1} \{ \text{Tr}[\rho]^2 \text{Tr}[H_k^2] + \text{Tr}[\rho^2] \text{Tr}[H_k]^2 \} \\ &\quad - \frac{2}{2^N(2^{2N} - 1)} \{ \text{Tr}[\rho^2] \text{Tr}[H_k^2] + \text{Tr}[\rho]^2 \text{Tr}[H_k]^2 \} \\ &\quad - \frac{2}{2^N} \text{Tr}[H_k^2] \text{Tr}[\rho^2] \quad (4.70) \end{aligned}$$

$$\approx -\frac{2^{N+1}}{2^{2N} - 1} \text{Tr}[H_k^2] \text{Tr}[\rho^2], \quad (4.71)$$

where Eqn. (4.70) follows from Eqn. (4.62) and Eqn. (4.64). Eqn. (4.71) is derived by ignoring minor terms and using  $\text{Tr}[H_k] = 0$ . Combining

Eqn. (4.69) and Eqn. (4.71), it can be shown that

$$\begin{aligned}\text{Var}_{U_+, U_-} \left[ \frac{\partial f}{\partial \theta_k} \right] &\approx \frac{2^N}{2 \times (2^{2N} - 1)^2} \text{Tr}[O^2] \text{Tr}[H_k^2] \text{Tr}[\rho^2] \\ &\approx \frac{1}{2^{3N+1}} \text{Tr}[O^2] \text{Tr}[\rho^2] \text{Tr}[H_k^2].\end{aligned}$$

Thus, Theorem 4.16 is proved. □

### Remark

The influence of barren plateau can be categorized into three folds.

1. **Training efficiency.** Exponentially small gradients imply that the training of QNNs with gradient-based optimizers may require exponential numbers of iterations to converge.
2. **Optimization effectiveness.** The calculation of the gradient via the parameter-shift rule would introduce unbiased statistical noise due to finite shot numbers. A small gradient could be vulnerable to these measurement noises, which may induce additional barriers in optimizations.
3. **Quantum advantage.** QNNs are expected to exhibit quantum advantages by employing intermediate to large numbers of qubits. However, the barren plateau phenomenon may induce exponential training steps, which can offset potential quantum advantages.

Since the barren plateau could seriously affect the trainability of scaled QNNs and raise concerns about the utility of QNNs for achieving quantum advantages, researchers have been focused on developing techniques to address this problem. Existing efforts including specific architecture design (Cerezo et al., 2021b; Pesah et al., 2021; Zhang et al., 2021b), parameter initialization schemes (Grant et al., 2019; Zhang et al., 2022b), and advanced training protocols (Skolik et al., 2021; Haug and Kim, 2021). Here we briefly introduce two related results with theoretical guarantees.

**Fact 4.17** (Shallow hardware-efficient circuits are BP-free, informal version adapted from Cerezo et al. (2021b)). *Suppose the observable has a local form in the Pauli basis decomposition. For QNNs employing N-qubit shallow hardware-efficient circuits with logarithmic depths, the variance of the*

gradient has the lower bound

$$\text{Var}\left[\frac{\partial f}{\partial \theta_k}\right] \geq \Omega\left(\frac{1}{\text{poly}(N)}\right). \quad (4.72)$$

**Fact 4.18** (Gaussian initializations help to escape the BP region, informal version adapted from [Zhang et al. \(2022b\)](#)). Suppose the observable is the tensor product of Pauli matrices  $\sigma_i = \sigma_{i_1} \otimes \cdots \otimes \sigma_{i_N}$ , where the number of non-identity matrices in  $\{\sigma_{i_1}, \dots, \sigma_{i_N}\}$  is  $S$ . For QNNs employing  $N$ -qubit shallow hardware-efficient circuits with the depth  $L$ , the gradient norm has the lower bound

$$\mathbb{E}_{\theta} \|\nabla_{\theta} f\|^2 \geq \frac{L}{S^S(L+2)^{S+1}} \text{Tr}[\sigma_j \rho_{\text{in}}]^2, \quad (4.73)$$

where  $S$  is the number of non-zero elements in  $\mathbf{i}$ , and the index  $\mathbf{j} = (j_1, j_2, \dots, j_N)$  such that  $j_m = 0, \forall i_m = 0$  and  $j_m = 3, \forall i_m \neq 0$ . The expectation is taken with the Gaussian distribution  $\mathcal{N}\left(0, \frac{1}{4S(L+2)}\right)$  for the parameters  $\theta$ .

## 4.5 Code Demonstration

This section provides hands-on demonstrations of QNNs for both discriminative and generative tasks, where we will illustrate practical implementations of quantum classifiers and quantum patch GAN. Each subsection corresponds to a specific application, offering a step-by-step explanation and code walkthrough.

### 4.5.1 Quantum classifier

We now demonstrate how to utilize QNN to solve discriminative tasks, specifically a binary classification problem based on the Wine dataset. Below, we outline the major steps in the pipeline:

- Step 1 Load and preprocess the dataset.
- Step 2 Implement a quantum read-in protocol to encode classical data into quantum states.
- Step 3 Construct a parameterized quantum circuit model to process the input quantum states.
- Step 4 Train and test the QNN to evaluate its performance.

We first import all the necessary libraries:

```

1 import sklearn
2 import sklearn.datasets
3 import pennylane as qml
4 from pennylane import numpy as np
5 from pennylane.optimize import AdamOptimizer
6 import matplotlib.pyplot as plt

```

**Step 1: Dataset preparation.** We prepare the Wine dataset for the classification task. For simplicity, we focus on the first two classes of the Wine dataset. The dataset consists of 13 attributes per sample, each with a distinct range. We apply normalization to rescale these attributes to the interval  $[0, \pi]$ . Furthermore, the labels are remapped from  $\{0, 1\}$  to  $\{-1, 1\}$  to align with the output range of the quantum circuit model. The dataset is split into training and test sets to fairly evaluate the classifier.

```

1 def load_wine(split_ratio = 0.5):
2     feat, label = sklearn.datasets.load_wine(return_X_y=True)
3
4     # normalization
5     feat = np.pi * (feat - np.min(feat, axis=0, keepdims=True)) / np.ptp(feat, axis=0, keepdims=True)
6
7     index_c0 = label == 0
8     index_c1 = label == 1
9
10    label = label * 2 - 1
11
12    n_c0 = sum(index_c0)
13    n_c1 = sum(index_c1)
14
15    X_train = np.concatenate((feat[index_c0][:int(split_ratio*n_c0)], feat[index_c1][:int(split_ratio*n_c1)]), axis=0)
16    y_train = np.concatenate((label[index_c0][:int(split_ratio*n_c0)], label[index_c1][:int(split_ratio*n_c1)]), axis=0)
17    X_test = np.concatenate((feat[index_c0][int(split_ratio*n_c0):], feat[index_c1][int(split_ratio*n_c1):]), axis=0)
18    y_test = np.concatenate((label[index_c0][int(split_ratio*n_c0):], label[index_c1][int(split_ratio*n_c1):]), axis=0)

```

```
19
20     return X_train, y_train, X_test, y_test
21 X_train, y_train, X_test, y_test = load_wine()
```

To better understand the dataset, we apply t-SNE to visualize its distribution. As shown in Figure 4.12, each data point is projected into a 2D space for visualization, with distinct colors representing different classes.

```
1 def visualize_dataset(X, y):
2     from sklearn.manifold import TSNE
3
4     tsne = TSNE(n_components=2, random_state=42,
5                  perplexity=30)
6
7     wine_tsne = tsne.fit_transform(X)
8     for label in np.unique(y):
9         indices = y == label
10        plt.scatter(wine_tsne[indices, 0], wine_tsne[
11            indices, 1], edgecolor='black', cmap='coolwarm', s=20, label=f'Class {label}')
12
13    # Add labels and legend
14    plt.title("t-SNE Visualization of Wine dataset (two classes)")
15    plt.xlabel("t-SNE Dimension 1")
16    plt.ylabel("t-SNE Dimension 2")
17    plt.legend()
18
19    plt.tight_layout()
20    plt.show()
21 visualize_dataset(X_train, y_train)
```

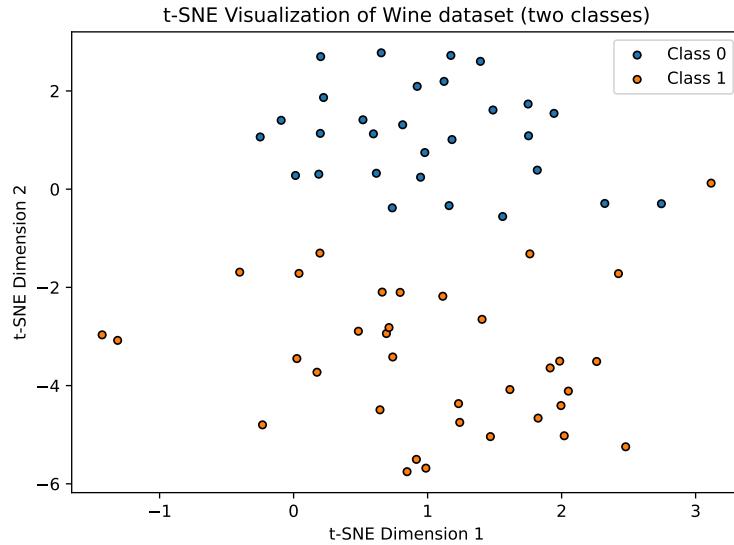


Figure 4.12: **T-SNE visualization of Wine dataset of the first two classes.**

**Step 2: Data encoding.** To encode the 13 attributes of the Wine dataset into a quantum system, we use angle encoding introduced in Section 2.3.1, followed by a layer of CNOT gates acting on neighboring qubits to introduce entanglement.

```

1 def data_encoding(x):
2     n_qubit = len(x)
3    qml.AngleEmbedding(features =x , wires = range(
4         n_qubit) , rotation ="X")
5     for i in range(n_qubit):
6         if i+1 < n_qubit:
7             qml.CNOT(wires=[i, i+1])

```

**Step 3: Building quantum classifier.** With the data encoding in place, we construct a quantum binary classifier. The circuit model is composed of multiple layers, where each layer includes parameterized single-qubit rotation gates with trainable angles, followed by a block of non-parametric CNOT gates to introduce entanglement among qubits. To read-out the category information of each input sample from the prepared quantum state, we compute the expectation value of the Pauli-Z operator on the first qubit.

```

1 def classifier(param , x=None):
2     data_encoding(x)

```

```
3     n_layer, n_qubit = param.shape[0], param.shape[1]
4     for i in range(n_layer):
5         for j in range(n_qubit):
6             qml.Rot(param[i, j, 0], param[i, j, 1], param
7                     [i, j, 2], wires=j)
8         for j in range(n_qubit):
9             if j+1 < n_qubit:
10                 qml.CNOT(wires=[j, j+1])
11     return qml.expval(qml.PauliZ(0))
12
13 n_qubit = X_train.shape[1]
14 dev = qml.device('default.qubit', wires=n_qubit)
circuit = qml.QNode(classifier, dev)
```

We visualize the whole quantum circuit of 2 layers by drawing the diagram, as shown in Figure 4.13.

```
1 fig, ax = qml.draw_mpl(circuit)(np.pi * np.random.randn
2                                 (2, n_qubit, 3), X_train[0])
3 fig.show()
```

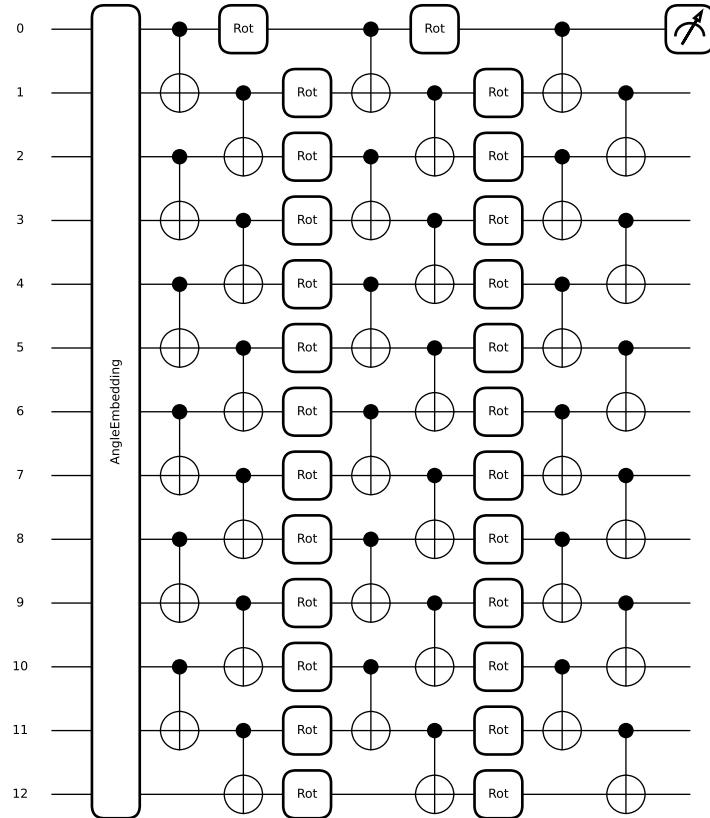


Figure 4.13: The circuit diagram of the quantum classifier.

**Step 4: Training and evaluation of quantum classifier.** With the data and circuit model ready, we now move to the optimization of the quantum classifier. The mean squared error (MSE) is used as the loss function. The goal is to minimize the difference between the predicted and actual labels over the training set.

```

1 def mse_loss(predict, label):
2     return np.mean((predict - label)**2)
3
4 def cost(param, circuit, X, y):
5     exp = []
6     for i in range(len(X)):
7         pred = circuit(param, x=X[i])
8         exp.append(pred)

```

```
9     return mse_loss(np.array(exp), y)
```

To evaluate the performance of the quantum classifier, we use classification accuracy as the metric. Specifically, if the sign of the read-out result matches the corresponding label, the prediction is considered correct; otherwise, it is deemed incorrect. The accuracy is then calculated as the proportion of correctly classified samples out of the total.

```
1 def accuracy(predicts, labels):
2     assert len(predicts) == len(labels)
3     return np.sum((np.sign(predicts)*labels+1)/2)/len(
4         predicts)
```

The Adam optimizer is utilized to minimize the loss function. To ensure efficient computation, the dataset is divided into smaller batches for each training iteration. At the end of each epoch, both the training and test losses, along with the classification accuracy, are recorded to track the model's performance.

```
1 lr = 0.01
2 opt = AdamOptimizer(lr)
3 batch_size = 4
4
5 n_epoch = 50
6 cost_train, cost_test, acc_train, acc_test = [], [], [], []
7 for i in range(n_epoch):
8     index = np.random.permutation(len(X_train))
9     feat_train, label_train = X_train[index], y_train[
10        index]
11    for j in range(0, len(X_train), batch_size):
12        feat_train_batch = feat_train[j*batch_size:(j+1)*
13            batch_size]
14        label_train_batch = label_train[j*batch_size:(j
15            +1)*batch_size]
16        param = opt.step(lambda v: cost(v, circuit,
17            feat_train_batch, label_train_batch), param)
18
19        # compute cost
20        cost_train.append(cost(param, circuit, X_train,
21            y_train))
22        cost_test.append(cost(param, circuit, X_test, y_test)
23            )
24
25        # compute accuracy
```

```

20     pred_train = []
21     for j in range(len(X_train)):
22         pred_train.append(circuit(param, x=X_train[j]))
23     acc_train.append(accuracy(np.array(pred_train),
24                               y_train))
25
26     pred_test = []
27     for j in range(len(X_test)):
28         pred_test.append(circuit(param, x=X_test[j]))
29     acc_test.append(accuracy(np.array(pred_test), y_test))
30

```

After training the QNN, the training and test cost, as well as the accuracy over epochs, can be visualized using the following code.

```

1 plt.figure(figsize=(12, 6))
2
3 plt.subplot(1, 2, 1)
4 epochs = np.arange(n_epoch) + 1
5 plt.plot(epochs, cost_train, label='Training Cost',
6           marker='o')
7 plt.plot(epochs, cost_test, label='Test Cost', marker='o')
8 plt.title('Cost Over Epochs')
9 plt.xlabel('Epochs')
10 plt.ylabel('Cost')
11 plt.legend()
12 plt.grid()
13
14 # Plot training and test accuracy
15 plt.subplot(1, 2, 2)
16 plt.plot(epochs, acc_train, label='Training Accuracy',
17           marker='o')
18 plt.plot(epochs, acc_test, label='Test Accuracy', marker='o')
19 plt.title('Accuracy Over Epochs')
20 plt.xlabel('Epochs')
21 plt.ylabel('Accuracy')
22 plt.legend()
23 plt.grid()
24 plt.tight_layout()
25 plt.show()

```

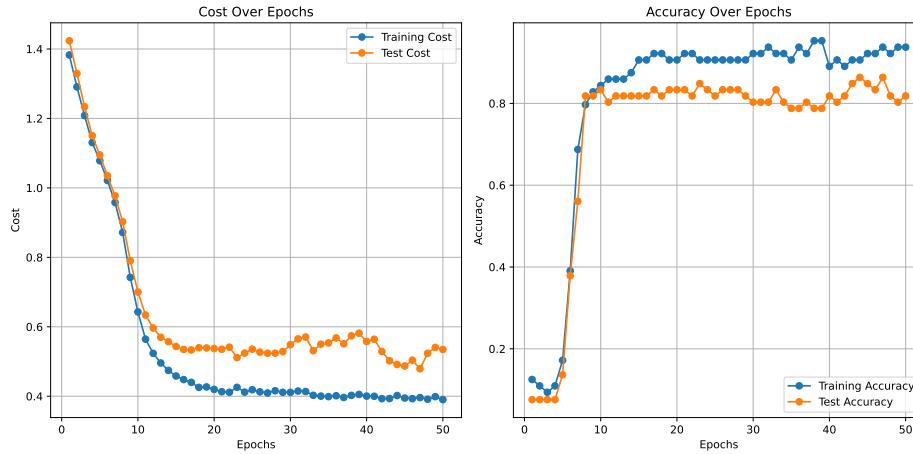


Figure 4.14: The training curve of the quantum classifier.

As demonstrated in Figure 4.14, this QNN achieves a test accuracy exceeding 0.8. The performance of the QNN could potentially be further enhanced by employing more advanced read-in protocols, as discussed in Chapter 2.3.1, which could enable more efficient and expressive representations of the input data. Additionally, optimizing the circuit design, such as adjusting the arrangement of layers or introducing more complex parameterized gates to increase the model’s capacity, as highlighted in Chapter 4.6, could further improve the QNN’s ability to capture intricate patterns in the dataset.

### 4.5.2 Quantum patch GAN

We next demonstrate how to implement a quantum patch GAN introduced in Chapter 4.3.3 for the generation of hand-written digits of five. The whole pipeline includes the following steps:

- Step 1 Load and pre-process the dataset.
- Step 2 Build the classical discriminator.
- Step 3 Build the quantum generator.
- Step 4 Train the quantum patch GAN.
- Step 5 Visualize the generated images.

We begin by importing the required libraries:

```

1 import torch
2 import torch.nn as nn
3 import torch.optim as optim
4 from torch.utils.data import Dataset, DataLoader
5
6 import numpy as np
7 import matplotlib.pyplot as plt
8 import pennylane as qml
9 import math

```

**Step 1: Dataset preparation.** We will use the Optical Recognition of Handwritten Digits dataset (`optdigits`), where each data point represents an  $8 \times 8$  grayscale image. The following code defines a custom dataset class to load and process the data.

```

1 class OptdigitsData(Dataset):
2     def __init__(self, data_path, label):
3         """
4             Dataset class for Optical Recognition of
5                 Handwritten Digits.
6         """
7
8         self.data = []
9         with open(data_path, 'r') as f:
10            for line in f.readlines():
11                if int(line.strip().split(',')[-1]) == label:
12                    # Normalize image pixel values from
13                    # [0,16) to [0, 1)
14                    image = [int(pixel)/16 for pixel in
15                            line.strip().split(',')[:-1]]
16                    image = np.array(image, dtype=np.
17                                    float32).reshape(8, 8)
18                    self.data.append(image)
19        self.label = label
20
21    def __len__(self):
22        return len(self.data)
23
24    def __getitem__(self, index):
25        return torch.from_numpy(self.data[index]), self.
26            label

```

After defining the dataset, we can visualize a few examples to better understand the structure of the dataset.

```

1 def visualize_dataset(data_path):
2     """
3         Visualizes the dataset by displaying examples for
4             each digit label.
5     """
6     plt.figure(figsize=(10, 5))
7     for i in range(10):
8         plt.subplot(1, 10, i + 1)
9         data = OptdigitsData(data_path, label=i)
10        plt.imshow(data[0][0], cmap='gray')
11        plt.title(f"Label: {i}")
12        plt.axis('off')
13    plt.tight_layout()
14    plt.show()
15
visualize_dataset('code/chapter5_qnn/data/optdigits.tra')
```

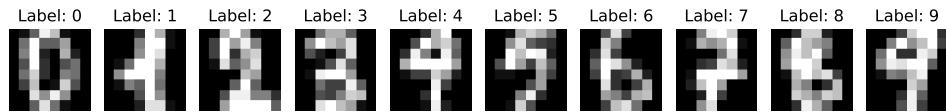


Figure 4.15: Samples in the dataset optdigits.

**Step 2: Building the classical discriminator.** The discriminator is a classical neural network responsible for distinguishing real images from fake ones. It consists of fully connected layers with ReLU activations. The final output is passed through a Sigmoid function, which scales the output to the range (0, 1), serving as a probabilistic indicator of whether the input image is real or fake.

```

1 class ClassicalDiscriminator(nn.Module):
2     """
3         A classical discriminator for classifying real and
4             fake images.
5     """
6     def __init__(self, input_shape):
7         super().__init__()
8         self.model = nn.Sequential(
9             nn.Flatten(),
10            nn.Linear(int(np.prod(input_shape)), 256),
11            nn.ReLU(),
```

```

11         nn.Dropout(),
12         nn.Linear(256, 128),
13         nn.ReLU(),
14         nn.Dropout(),
15         nn.Linear(128, 1),
16         nn.Sigmoid()
17     )
18
19     def forward(self, img):
20         return self.model(img)

```

**Step 3: Defining the quantum patch generator.** The generator in the quantum patch GAN consists of parameterized quantum circuits (PQC). These circuits are responsible for generating patches of the target image. Specifically, the PQC follows the layout in Figure 4.7, which applies layers of single-qubit rotation gates and entangling gates to the latent state.

```

1 def PQC(params):
2     n_layer, n_qubit = params.shape[0], params.shape[1]
3     for i in range(n_layer):
4         for j in range(n_qubit):
5             qml.Rot(params[i, j, 0], params[i, j, 1],
6                     params[i, j, 2], wires=j)
7             # Control Z gates
8             for j in range(n_qubit - 1):
9                 qml.CZ(wires=[j, j + 1])

```

Then, we implement the quantum generator for each patch of an image, i.e., sub-generators. The sub-generator transforms the latent variable  $z$  into a latent quantum state  $|z\rangle$ , applies a PQC, performs partial measurements on the ancillary system  $\mathcal{A}$ , and outputs the probabilities of each computational basis state in the remaining system, which correspond to the generated pixel values.

```

1 def QuantumGenerator(params, z=None, n_qubit_a=1):
2     n_qubit = params.shape[1]
3
4     # angle encoding of latent state z
5     for i in range(n_qubit):
6         qml.RY(z[i], wires=i)
7
8     PQC(params)
9
10    # partial measurement on the ancillary qubits
11    qml.measure(wires=n_qubit-1)

```

```
12     return qml.probs(wires=range(n_qubit-n_qubit_a))
```

Using the sub-generators, the quantum patch generator combines the output patches from multiple sub-generators to construct the complete image.

```
1  class PatchQuantumGenerator(nn.Module):
2      """
3          Combines patches generated by quantum circuits into
4          full images.
5      """
6      def __init__(self, qnode_generator, n_generator,
7          n_qubit, n_qubit_a, n_layer):
8          super().__init__()
9
10         self.params_generator = nn.ParameterList([
11             nn.Parameter(torch.rand((n_layer, n_qubit, 3),
12                 requires_grad=True) for _ in range(
13                 n_generator))
14         ])
15         self.qnode_generator = qnode_generator
16         self.n_qubit_a = n_qubit_a
17
18     def forward(self, zs):
19         images = []
20         for z in zs:
21             patches = []
22             for params in self.params_generator:
23                 patch = self.qnode_generator(params, z=z,
24                     n_qubit_a=self.n_qubit_a).float()
25
26                 # post-processing: min-max scaling
27                 patch = (patch - patch.min()) / (patch.
28                     max() - patch.min() + 1e-8)
29
30                 patches.append(patch.unsqueeze(0))
31             patches = torch.cat(patches, dim=0)
32             images.append(patches.flatten().unsqueeze(0))
33
34     return torch.cat(images, dim=0)
```

**Step 4: Training the quantum patch GAN.** With the dataset and models ready, we initialize the quantum generator, classical discriminator, and their optimizers.

```
1  # Hyperparameters
2  torch.manual_seed(0)
```

```

3  image_width = 8
4  image_height = 8
5  n_generator = 4
6  n_qubit_d = int(np.log2((image_width * image_height) // 
    n_generator))
7  n_qubit_a = 1
8  n_qubit = n_qubit_d + n_qubit_a
9  n_layer = 6
10
11 # Quantum device
12 dev = qml.device("lightning.qubit", wires=n_qubit)
13 qnode_generator = qml.QNode(QuantumGenerator, dev)
14
15 # Initialize generator and discriminator
16 discriminator = ClassicalDiscriminator([image_height,
    image_width])
17 discriminator.train()
18 generator = PatchQuantumGenerator(qnode_generator,
    n_generator, n_qubit, n_qubit_a, n_layer)
19 generator.train()
20
21 # Optimizers
22 lr_generator = 0.3
23 lr_discriminator = 1e-2
24 opt_discriminator = optim.SGD(discriminator.parameters(),
    lr=lr_discriminator)
25 opt_generator = optim.SGD(generator.parameters(), lr=
    lr_generator)
26
27 # Construct dataset and dataloader
28 batch_size = 4
29 dataset = OptdigitsData('code/chapter5_qnn/data/optdigits
    .tra', label=5)
30 dataloader = DataLoader(dataset, batch_size=batch_size,
    shuffle=True, drop_last=True)
31
32 # Loss function
33 loss_fn = nn.BCELoss()
34 labels_real = torch.ones(batch_size, dtype=torch.float)
35 labels_fake = torch.zeros(batch_size, dtype=torch.float)
36
37 # Testing setup
38 n_test = 10
39 z_test = torch.rand(n_test, n_qubit) * math.pi

```

The GAN training process involves alternating updates for the discriminator and the generator. The discriminator is trained to distinguish between real and fake images, while the generator learns to create images that can successfully deceive the discriminator. To track the generator's progress during training, the generated images are saved at the end of each epoch.

```

1 n_epoch = 10
2 record = []
3 for i in range(n_epoch):
4     for data, _ in dataloader:
5
6         zs = torch.rand(batch_size, n_qubit) * math.pi
7         image_fake = generator(zs)
8
9         # Training the discriminator
10        discriminator.zero_grad()
11        pred_fake = discriminator(image_fake.detach())
12        pred_real = discriminator(data)
13
14        loss_discriminator = loss_fn(pred_fake.squeeze(),
15                                      labels_fake) + loss_fn(pred_real.squeeze(),
16                                      labels_real)
17        loss_discriminator.backward()
18        opt_discriminator.step()
19
20        # Training the generator
21        generator.zero_grad()
22        pred_fake = discriminator(image_fake)
23        loss_generator = loss_fn(pred_fake.squeeze(),
24                                  labels_real)
25        loss_generator.backward()
26        opt_generator.step()
27
28        print(f'The {i}-th epoch: discriminator loss={loss_discriminator:0.3f}, generator loss={loss_generator:0.3f}')
29
30    # test
31    generator.eval()
32    image_generated = generator(z_test).view(n_test,
33                                             image_height, image_width).detach()
34
35    record[str(i)] = {
36        'loss_discriminator': loss_discriminator.item(),
37        'loss_generator': loss_generator.item(),
38    }

```

```

34         'image_generated': image_generated.numpy().tolist
35             ())
36     }
generator.train()

```

**Step 5: Visualizing the generated images.** After training, we visualize the images generated by the quantum generator to evaluate its performance. These visualizations allow us to see how well the model has learned to produce realistic image patches.

```

1 n_epochs_to_visualize = len(record) // 2
2 n_images_per_epoch = 10
3
4 fig, axes = plt.subplots(n_epochs_to_visualize,
5                         n_images_per_epoch, figsize=(n_images_per_epoch,
6                         n_epochs_to_visualize))
7
8 # Iterate through the recorded epochs and visualize
9 # generated images
10 for epoch_idx, (epoch, data) in enumerate(record.items()):
11     :
12     if epoch_idx % 2 == 1:
13         continue
14     images = np.array(data['image_generated'])
15
16     for img_idx in range(n_images_per_epoch):
17         ax = axes[epoch_idx // 2, img_idx]
18         ax.imshow(images[img_idx], cmap='gray')
19         ax.axis('off')
20
21         # Add epoch information to the title of each row
22         if img_idx == 0:
23             ax.set_title(f"Epoch {epoch}", fontsize=10)
24
25 plt.tight_layout()
26 plt.show()

```

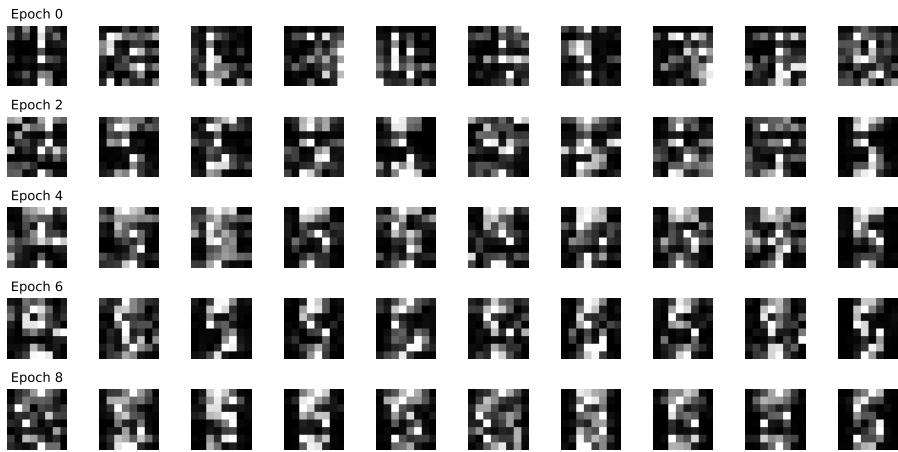


Figure 4.16: The images generated by quantum patch GAN during training.

## 4.6 Bibliographic Remarks

Quantum neural networks (QNNs) have emerged as a prominent paradigm in quantum machine learning, demonstrating potential in both discriminative and generative learning tasks. For discriminative tasks, QNNs utilize high-dimensional Hilbert spaces to efficiently capture and represent complex intrinsic relationships. In generative tasks, QNNs leverage variational quantum circuits to generate complex probability distributions that may exceed the capabilities of classical models. While these approaches share common learning strategies, they each introduce unique challenges and opportunities in terms of model design, theoretical exploration, and practical implementation. In the remainder of this section, we briefly review recent advances in QNNs. Interested readers can refer to [Ablayev et al. \(2019\)](#); [Li and Deng \(2022\)](#); [Massoli et al. \(2022\)](#); [Tian et al. \(2023\)](#); [Wang and Liu \(2024\)](#) for a more comprehensive review.

### 4.6.1 Discriminative learning with QNN

QNNs for discriminative tasks have emerged as one of the most active research areas in quantum machine learning, demonstrating potential advantages in feature representation and processing efficiency. The quantum learning approach leverages the high-dimensional Hilbert space and quantum parallelism to potentially handle complex classification boundaries more

effectively than classical neural networks. Research has shown particular promise in handling datasets with inherent quantum properties and problems where quantum entanglement can be meaningfully exploited (Huang et al., 2022).

### Model designs

In the realm of quantum discriminative models, researchers have developed various quantum neural architectures. In general, variational quantum classifiers (Havlíček et al., 2019; Mitarai et al., 2018) could employ parameterized quantum circuits for classification tasks. Subsequently, quantum convolutional neural networks (Cong et al., 2019) are designed for processing structured data. Hybrid quantum-classical architectures (Arthur et al., 2022) are proposed to combine quantum layers with classical neural networks. Other notable works include the development of quantum versions of popular classical architectures like recurrent neural networks (Bausch, 2020) and attention mechanisms (Shi et al., 2024). Finally, Pérez-Salinas et al. (2020); Fan and Situ (2022) have explored quantum re-uploading strategies for encoding classical data, achieving QML models with more expressive feature maps.

In addition to manually designed architectures, various lightening strategies have been explored to enhance the efficiency of quantum neural networks. For example, quantum architecture search methods have been developed by (Du et al., 2022a; Zhang et al., 2022c; Linghu et al., 2024) to automatically discover optimal quantum circuit designs with reduced gate complexity. Sim et al. (2021); Wang et al. (2022b) introduced quantum pruning techniques that systematically identify and remove redundant quantum gates while preserving the performance. In the realm of knowledge distillation, researchers have demonstrated how to transfer knowledge from the teacher model given as quantum (Alam et al., 2023) or classical (Li et al., 2024) neural networks to more compact quantum circuit architectures that are more robust against quantum noises. These optimization approaches have collectively contributed to improving the practical performance of QNNs on real quantum devices, particularly in the NISQ era.

### Theoretical foundations

To gain a deeper understanding of the potential advantages and limitations of QNNs, a crucial research topic is analyzing their learnability. More concisely, the learnability is determined by the interplay of three key aspects:

expressivity, trainability, and generalization, as preliminarily introduced in Chapter 4.4 with essential theoretical results. Beyond these foundational insights, an extensive body of research has conducted more comprehensive and detailed investigations into these three aspects, which will be reviewed individually in the following.

**Expressivity.** The expressivity of QNNs refers to their ability to represent complex functions or quantum states efficiently. Universal approximation theorems (UAT) incorporating data re-uploading strategies have been established by Pérez-Salinas et al. (2020) firstly with subsequent works (Schuld et al., 2021; Yu et al., 2022a) in various problem settings. Beyond the UAT, Sim et al. (2019), Nakaji and Yamamoto (2021), and Holmes et al. (2022) analyze the expressivity of QNNs by investigating how well the parameterized quantum circuits used in QNNs can approximate the Haar distribution, a critical measure of expressive capacity in quantum systems. Moreover, Yu et al. (2022b) analyze the non-asymptotic error bounds of variational quantum circuits for approximating multivariate polynomials and smooth functions.

**Trainability.** The trainability of QNNs corresponds to two aspects during the optimization of QNNs: the *gradient magnitude* and the *convergence rate*.

For the first research line, McClean et al. (2018) first found the phenomenon of vanishing gradients, dubbed as the barren plateau, where the gradient magnitude scales exponentially small with the quantum system size. Since then, a series of studies explored the cause of barren plateau, including global measurements (Cerezo et al., 2021b), highly entanglement (Ortiz Marroero et al., 2021), and quantum system noise (Wang et al., 2021b).

Efforts to address the barren plateau problem, a major challenge in training deep quantum circuits, have yielded strategies such as proper circuit and parameter initialization techniques (Grant et al., 2019; Zhang et al., 2022b), cost function design (Cerezo et al., 2021b), and proper circuits (Pesah et al., 2021). Quantum-specific regularization techniques have also been developed to mitigate these effects (Larocca et al., 2022).

Another research line on the trainability of QNNs focuses on the convergence of QNNs, which is not introduced in this Chapter. In particular, Kiani et al. (2020) and Wiersema et al. (2020) experimentally found that overparameterized QNNs embrace a benign landscape and permit fast convergence towards near optimal local minima. Afterward, initial attempts have been made to theoretically explain the superiority of over-parameterized QNNs. Specifically, Larocca et al. (2023) and Anschuetz (2021) utilized tools from dynamical Lie algebra and random matrix theory, respectively, to quantify

the critical points in the optimization landscape of overparameterized QNNs. Moreover, You et al. (2022) extended the classical convergence results of Xu et al. (2018) to the quantum domain, proving that overparameterized QNNs achieve an exponential convergence rate. Additionally, Liu et al. (2023) and Wang et al. (2023a) introduced the concept of the quantum neural tangent kernel (QNTK) to further demonstrate the exponential convergence rate of overparameterized QNNs. Besides the overparameterization theory, (Du et al., 2021b, 2022b; Qi et al., 2023; Qian et al., 2024) investigated the required conditions to ensure the convergence of QNNs towards local minima.

**Generalization.** Research has also focused on understanding the sample complexity and generalization error bounds of quantum machine learning algorithms by using different statistical learning tools. In particular, Abbas et al. (2021) compared the generalization power of QNNs and classical learning models based on an information geometry metric. Caro et al. (2022) and Du et al. (2022c) established generalization error bounds using covering numbers, revealing the impact of circuit structural factors—such as the number of gates and types of observables—on generalization ability. Similarly, Bu et al. (2022) analyzed the generalization ability of QNNs from the perspective of quantum resource theory, emphasizing the role of quantum resources such as entanglement and magic in influencing generalization.

Furthermore, frameworks for demonstrating quantum advantage in specific learning scenarios have been proposed (Huang et al., 2021c, 2022), providing insights into the conditions under which quantum models outperform their classical counterparts. Zhang et al. (2024a) investigate the training-dependent generalization abilities of QNNs, while Du et al. (2023) study problem-dependent generalization, highlighting key factors that enable QNNs to achieve strong generalization performance.

Beyond analyses focused on specific datasets and problems, the generalization ability of QNNs has also been examined through the lens of the No-Free-Lunch theorem. Poland et al. (2020) explore the average performance of QNNs across all possible datasets and problems, providing a broader perspective on their generalization potential. Extending this work, Sharma et al. (2022) and Wang et al. (2024a) adapt the No-Free-Lunch theorem to scenarios involving entangled data, demonstrating the potential benefits of entanglement in certain settings. Additionally, Wang et al. (2024b) establish a No-Free-Lunch framework for various learning protocols, considering different quantum resources used in these protocols.

**Potential advantages.** A critical challenge in quantum learning theory is identifying tasks where QNNs demonstrate provable computational ad-

vantages over classical approaches when solving classical problems. While recent studies have explored unconditional quantum advantage, most focus on synthetic tasks. For instance, polynomial advantages over commonly used classical models have been demonstrated through quantum contextuality in sequential learning (Anschuetz et al., 2023; Anschuetz and Gao, 2024). Additionally, quantum entanglement has been shown to reduce communication requirements in non-local machine-learning tasks (Zhao and Deng, 2024). Despite these advancements, a significant issue remains: most QNNs, without careful design, can be efficiently simulated or approximated by classical surrogate models. The review (Cerezo et al., 2023) presented strong evidence that commonly used models with provable absence of barren plateaus are also classically simulable. Concrete algorithms in this research line include Pauli path simulators (Bermejo et al., 2024; Angrisani et al., 2024; Lerch et al., 2024), tensor-network simulators (Shin et al., 2024), and learning protocols (Landman et al., 2022; Schreiber et al., 2023; Du et al., 2024).

## Applications

Practical applications of quantum neural networks for discriminative learning have spanned multiple domains. In computer vision, researchers have demonstrated quantum approaches to image classification (Henderson et al., 2020) and pattern recognition (ALRikabi et al., 2022). In quantum chemistry, QNNs have been applied to proton affinity predictions (Jin and Merz Jr, 2024) and catalyst developments (Roh et al., 2024). Financial applications include market trend classification (Li et al., 2023a) and fraud detection (Innan et al., 2024). Medical applications encompass drug discovery (Batra et al., 2021) and disease diagnosis (Enad et al., 2023).

### 4.6.2 Generative learning with QNNs

QNNs for generative tasks represent a promising avenue in quantum machine learning, offering new methodologies for generating complex data distributions. By leveraging variational quantum circuits, these models aim to generate potentially more complex probability distributions compared to classical counterparts, particularly in domains where high-dimensional data or quantum properties are prominent. We refer the readers to Tian et al. (2023) for a survey of QNNs in generative learning.

### Model designs

Researchers have developed various QNN architectures tailored for generative tasks. Quantum circuit Born machines (QCBMs) (Benedetti et al., 2019a) are one of the pioneering models, utilizing parameterized quantum circuits to generate discrete probability distributions. Quantum generative adversarial networks (Lloyd and Weedbrook, 2018) extend the adversarial framework to the quantum domain, where quantum generators and discriminators compete to learn complex distributions. Quantum Boltzmann machines (Amin et al., 2018) are another notable model, employing quantum devices to prepare Boltzmann distributions for estimating target discrete distributions. Additionally, quantum autoencoders (Romero et al., 2017) have been proposed for tasks like quantum state compression and reconstruction, offering potential advantages in quantum information processing. Recently, quantum diffusion models (Zhang et al., 2024b; Kölle et al., 2024) have been proposed for generating quantum state ensembles or classical images. These models showcase the versatility of QNNs in addressing diverse generative tasks.

### Theoretical foundations

The theoretical understanding of generative quantum neural networks has advanced in several directions. Similar to QNNs for discriminative tasks, quantum generative models like QCBMs face the barren plateau issue with additional mechanisms from the Kullbach-Leibler (KL) divergence loss function (Rudolph et al., 2024). In parallel, QCBMs are more efficient in the data-limited regime than the other classical generative models (Hibat-Allah et al., 2024). Besides, Gao et al. (2018) proved the existence of quantum generative model that is more capable of representing probability distributions compared with classical generative models, which has exponential speedup in learning and inference. Similarly, Gao et al. (2022) proved the separation in expressive power between a class of widely used generative models, known as Bayesian networks, and its minimal quantum-inspired extension. Du et al. (2022d) analyzed the generalization bounds of QCBMs and QGANs under the maximum mean discrepancy loss.

### Applications

Generative QNNs have shown potential in various practical applications. In finance, they have been used to model complex financial data distributions and generate synthetic financial datasets, demonstrating better performance

than classical models in certain scenarios ([Alcazar et al., 2020](#); [Zhu et al., 2022](#)). In the domain of quantum physics, quantum generative models have been applied for quantum state tomography and quantum simulation, aiding in the understanding of quantum systems ([Benedetti et al., 2019b](#)). In image generation, QGANs have been employed to produce high-quality images, showcasing their capability in handling complex visual data ([Huang et al., 2021b](#)). Furthermore, quantum generative models have been explored for drug discovery, where they can potentially accelerate the process by efficiently exploring large chemical spaces ([Li et al., 2021](#)). These applications highlight the broad potential of QNNs in generative tasks across different domains.



## Chapter 5

# Quantum Transformer

Transformers, introduced by [Vaswani \(2017\)](#), have become one of the most important and widely adopted deep learning architectures in modern AI. Transformers were first developed to improve previous architectures for natural language processing on the ability to handle long-range dependencies and capture intricate relationships in data. Unlike previous sequential models, such as recurrent neural networks, which process information in a step-by-step manner, transformers use a mechanism called *self-attention* to capture correlations among all elements in a sequence simultaneously. This parallel processing capability significantly reduces training time and improves learning performance.

Despite its many advantages, the transformer architecture has several drawbacks, particularly the required computational resources. As discussed in previous chapters, quantum computing provides unique advantages over classical computing in certain applications by leveraging quantum phenomena such as superposition, entanglement, and interference. These capabilities have inspired researchers to explore whether integrating quantum computing with Transformers could lead to superior performance compared to their classical counterparts in specific tasks.

To figure out this question, in this chapter, we start by introducing the mechanism of Transformers in Chapter [5.1](#). We then illustrate how to construct a quantum Transformer on a fault-tolerant quantum computer in Chapter [5.2](#). We also analyze the runtime of quantum transformers combined with numerical observations in Chapter [5.3](#), demonstrating a quadratic speedup over the classical counterpart.

## 5.1 Classical Transformer

The transformer architecture is designed to predict the next *token* (formally present in Chapter 5.1.1) in a sequence by leveraging sophisticated neural network components. Its modular design—including residual connections, layer normalization, and feed-forward networks (FFNs) as introduced in Chapter 4.1—makes it highly scalable and customizable. This versatility has enabled its successful application in large-scale foundation models across diverse domains, including natural language processing, computer vision, reinforcement learning, robotics, and beyond.

The full architecture of Transformer is illustrated in Figure 5.1. Note that while the original paper by Vaswani (2017) introduced both encoder and decoder components, contemporary large language models primarily adopt *decoder-only architectures*, which have demonstrated superior practical performance. Therefore, in the remainder of this section, we focus on detailing the implementation of each building block and discussing the optimization of decoder-only Transformer architectures. To deepen the understanding, a toy example of a classical Transformer with the code demonstration is provided in Chapter 5.4.

### 5.1.1 Tokenization and embedding

To handle sequential data, such as natural language, Transformers employ *tokenization* to convert it into discrete units. This preprocessing step makes the data compatible with computational models and optimizes it for parallel processing, particularly on GPUs. More concisely, Tokenization breaks a sentence into smaller pieces called *tokens*, which could be words, subwords, or even characters, depending on the tokenization strategy. For example, the sentence “Transformers are amazing!” might become tokens like (“Transform”, “ers”, “are”, “amazing”, “!”) if subwords are used. Modern tokenization methods (Sennrich et al., 2016; Kudo and Richardson, 2018; Mielke et al., 2021) enable sophisticated mapping of complex inputs into token spaces.

For Transformer, tokens are mapped to high-dimensional real vector representations via *embedding* (Vaswani, 2017), as highlighted by the solid box “Embeddings/Projections” in Figure 5.1. Let  $d_{\text{token}}$  denote the dictionary’s token count and  $d_{\text{model}}$  represent the embedding vector dimension. We define the set containing all token embedding vectors in the dictionary as

$$\mathcal{W} := \{\mathcal{W}_j \in \mathbb{R}^{d_{\text{model}}} : \mathcal{W}_j \text{ is the embedding of token } j \in [d_{\text{token}}]\}.$$

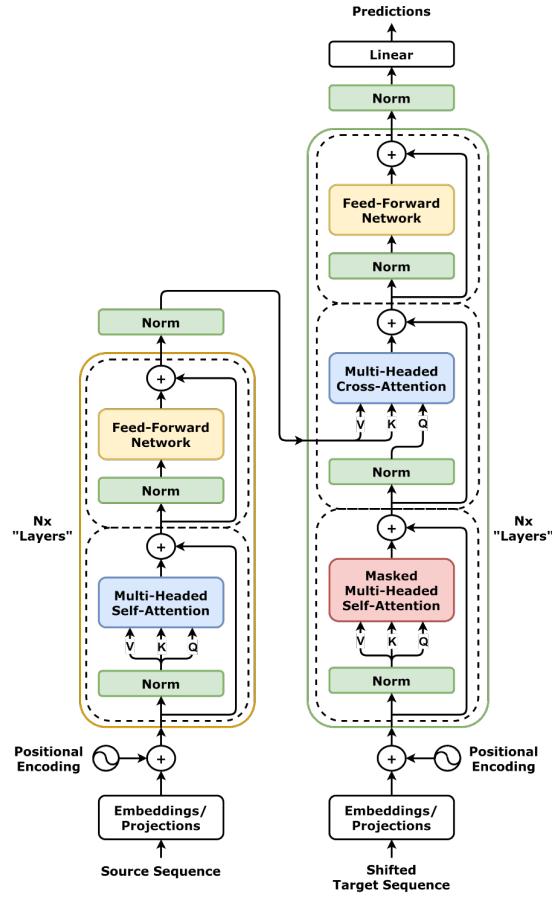


Figure 5.1: **A standard Transformer architecture, showing an encoder on the left, and a decoder on the right.** Image by Daniel Voigt Godoy under CC BY. The encoder processes the source sequence through multiple layers of multi-headed self-attention and feed-forward networks, augmented with residual connections, layer normalization, and positional encodings. The decoder incorporates masked multi-headed self-attention to process the target sequence and multi-headed cross-attention to integrate information from the encoder. The output is passed through a feed-forward network and a final linear layer to generate predictions.

An  $\ell$ -length sentence is represented as a sequence of vectors  $\{S_j\}_{j=1}^\ell$ , where  $S_j \in \mathcal{W}$ . Mathematically, this sequence can be interpreted as a real matrix  $S \in \mathbb{R}^{\ell \times d_{\text{model}}}$  whose  $j$ -th row  $S_j$  representing the  $j$ -th token.

### 5.1.2 Self-attention

*Self-attention* is a core building block of the transformer architecture, which captures intrinsic correlations among tokens. By allowing each token in a sequence to attend to every other token, Transformer generates attention matrices via the inner-product operations, encoding complex inter-token relationships into a transformative vector representation, colloquially termed “scaled dot-product attention”. The generated attention matrices highlight how relevant each part of the input is to every other part. This allows Transformers to handle contextual dependencies across a variety of data structures.

The self-attention mechanism, as highlighted by the blue or red box in Figure 5.1, involves three parameterized weight matrices:  $W_q, W_k \in \mathbb{R}^{d_{\text{model}} \times d_k}$  and  $W_v \in \mathbb{R}^{d_{\text{model}} \times d_v}$ .

#### Remark

Following conventions (Vaswani, 2017), we use the notation  $d$  to specify  $d_{\text{model}}$ ,  $d_k$ , and  $d_v$  in the rest of this chapter, i.e.,  $d := d_{\text{model}} = d_k = d_v$ . This is a widely used setting in practice.

Given a sequence  $S \in \mathbb{R}^{\ell \times d}$ , we define the three new matrices after interacting it with three parameterized weight matrices  $W_q, W_k, W_v$ , i.e.,

- Query matrix:  $Q := SW_q$ ;
- Key matrix:  $K := SW_k$ ;
- Value matrix:  $V := SW_v$ .

The attention block computes the matrix  $G^{\text{soft}} \in \mathbb{R}^{\ell \times d}$  via

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^\top / \alpha_0)V =: G^{\text{soft}}, \quad (5.1)$$

where  $\alpha_0 > 0$  is a scaling factor, and  $\text{softmax}(\cdot)$  is a row-wise nonlinear transformation such that  $\text{softmax}(\mathbf{z})_j := e^{z_j} / (\sum_{k \in [\ell]} e^{z_k})$  for  $\mathbf{z} \in \mathbb{R}^\ell$  and  $j \in [\ell]$ . The row-wise softmax application ensures the controlled attention distribution. The scaling factor  $\alpha_0 = \sqrt{d}$  empirically prevents excessive value amplification, particularly when input matrix rows have zero mean and unit standard deviation.

For decoder-only architectures, *masked* self-attention is employed, strategically hiding tokens subsequent to the current query token, i.e.,

$$M_{jk} = \begin{cases} 0 & k \leq j, \\ -\infty & k > j. \end{cases} \quad (5.2)$$

Conceptually, the mask  $M$  is applied to the scaled dot product  $QK^\top/\alpha_0$  in Eq. (5.1) before the softmax operation. Specifically, the matrix in the softmax operation is modified as  $QK^\top/\alpha_0 + M$ .

Another crucial technique in Transformer is the multi-head attention, which further extends the self-attention mechanism by computing and concatenating multiple attention matrices, enabling parallel representation learning across different subspaces. In practice, the embedding dimensions are often much larger (e.g.,  $d = 512$  or  $d = 768$ ), with multiple attention heads working simultaneously, each capturing distinct relationships between words in the sequence. This mechanism plays a crucial role in modern AI, as it allows words to dynamically interact with one another within the context of the sequence.

While multi-head attention is a pivotal advancement in Transformer architectures, we will not delve into its details here, as *the quantum Transformer implementations presented below primarily support single-head attention*. Nonetheless, these techniques remain essential in classical AI and are promising directions for future developments in quantum Transformers.

### 5.1.3 Residual connection

Residual connections (the arrows bypassing the main components, such as the attention and feed-forward layers in Figure 5.1), paired with layer normalization (green box in the figure), provide crucial architectural flexibility and robustness. By enabling direct information flow between layers, they mitigate challenges in training deep neural networks (He et al., 2015; Ba et al., 2016).

For the  $j$ -th token in an  $\ell$ -length sentence, the residual connection generates  $G_j^{\text{soft}} + S_j \in \mathbb{R}^d$  for  $\forall j \in [\ell]$ , which is subsequently normalized to standardize the vector representation. Let

$$\bar{s}_j := \frac{1}{d} \left( \sum_{k=1}^d (G_{jk}^{\text{soft}} + S_{jk}), \dots, \sum_{k=1}^d (G_{jk}^{\text{soft}} + S_{jk}) \right) \in \mathbb{R}^d,$$

where  $\varsigma := \sqrt{\frac{1}{d} \sum_{k=1}^d ((G_j^{\text{soft}} + S_j - \bar{s}_j)_k)^2}$ . The complete residual connection with the *layer normalization*  $\text{LN}(\cdot, \cdot)$  can be expressed as

$$\text{LN}_{\gamma, \beta}(G_j^{\text{soft}}, S_j) = \gamma \frac{G_j^{\text{soft}} + S_j - \bar{s}_j}{\varsigma} + \beta, \quad (5.3)$$

where  $\gamma$  and  $\beta$  denote the scale and bias parameters, respectively.

#### 5.1.4 Feed-forward network

Recall the definitions of fully-connected neural networks (FFN) in Chapter 4.1. Transformers employ a two-layer fully connected transformation (yellow box in Figure 5.1) to proceed with the output of residual connection, i.e.,

$$\text{FFN}(\text{LN}(z_j, S_j)) = \sigma(\text{LN}(G_j^{\text{soft}}, S_j)M_1 + b_1)M_2 + b_2, \quad (5.4)$$

where  $M_1 \in \mathbb{R}^{d \times d'}$ ,  $M_2 \in \mathbb{R}^{d' \times d}$  are linear transformation matrices, and  $b_1, b_2$  are vectors. In most practical cases,  $d' = 4d$ . Here  $\sigma(x)$  is an activation function, such as  $\tanh(x)$  and  $\text{ReLU}(x) = \max(0, x)$ . Another activation function that has been widely used in Transformers is the Gaussian Error Linear Units function (GELU), i.e.,

$$\text{GELU}(x) = x \cdot \frac{1}{2} \left( 1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right) \right).$$

##### Single-head and single-block Transformer

By integrating all ingredients introduced in Chapters 5.1.1, we reach out the explicit form of single-head and single-block Transformer, i.e.,

$$\text{Transformer}(S, j) := \text{LN}(\text{FFN}(\text{LN}(\text{Attention}(S, j))). \quad (5.5)$$

For the final output, i.e., to predict the next token, one can implement the softmax function to make the vector into a distribution  $\Pr(\cdot | S_1, \dots, S_{j-1})$ , where the dimension is the size of the token dictionary  $d_{\text{token}}$ , and sample from this distribution.

In modern architectures, multiple computational blocks are applied iteratively. Similar to multi-head attention, we will not explore this in detail here, as the quantum Transformer implementations introduced below primarily support single-head attention.

### 5.1.5 Optimization and inference

Upon the architecture of Transformer, its *optimization* involves training the model to achieve high performance on a given task. When applied to language processing tasks, a feasible loss function of the Transformer is the cross entropy between the predicted probability distribution and the correct distribution. Mathematically, given a  $\ell$ -length sequence  $\{S_1, \dots, S_\ell\}$  as input, the loss function can be written as

$$\mathcal{L} = -\frac{1}{\ell} \sum_{j=1}^{\ell} \Pr(S_j | S_1, \dots, S_{j-1}), \quad (5.6)$$

where  $\Pr(S_j | S_1, \dots, S_{j-1})$  is the predicted probability of the correct  $S_j$  coming out of the softmax layer, based on the previous tokens.

This process of training Transformers can be achieved by using Adam optimizer ([Kingma and Ba, 2015](#)). Learning rate schedules, such as warm-up followed by decay, can provide stable and effective convergence.

After optimization, we could use the trained Transformer for *inference*, which refers to making predictions on new data. Given a new initial sequence  $S' = \{S'_1, \dots, S'_{j-1}\}$ , we feed it to the trained Transformer and obtain the distribution  $\Pr(S'_j | S'_1, \dots, S'_{j-1})$ . Then we can use a decoding strategy (e.g., greedy decoding or sampling) to select the token based on the highest probability or a sampling approach.

Efficient inference is crucial for deploying models in real-world applications. Speed optimization techniques, such as quantization, reduce the precision of weights and activations to accelerate inference with minimal accuracy loss ([Jacob et al., 2018](#)). Pruning removes redundant weights or attention heads to reduce the model size and computational cost ([Han et al., 2015](#)). Batching and parallelism are also critical, with batched inference allowing the processing of multiple inputs simultaneously and GPU or TPU acceleration enabling parallel computations. Efficient attention at inference, such as caching key and value tensors, reduces redundant computations in autoregressive tasks like text generation ([Shoeybi et al., 2019](#)).

The inference cost is up to 10 times the training cost as large language models (LLMs) are trained once and applied millions of times ([McDonald et al., 2022; Desislavov et al., 2023](#)). For this reason, in the next section, we explore how to harness quantum computing to address this issue, which is crucial from both scientific and societal perspectives.

## 5.2 Fault-tolerant Quantum Transformer

In this section, we move on to show an end-to-end transformer architecture implementable on a quantum device, which includes all the key building blocks introduced in Chapter 5.1, and a discussion of the potential runtime speedups of this quantum model. In particular, here we focus on the *inference process* in which a classical Transformer has already been trained and is queried to predict the single next token.

Recall that in Chapter 5.1, we suppose that the three parameterized matrices in the self-attention mechanism have the same size, i.e.,  $W_q, W_k, W_v \in \mathbb{R}^{d \times d}$ . Besides, the input sequence  $S$  and the matrix returned by the attention block  $G^{\text{soft}}$  has the size  $\ell \times d$ . Here, we further suppose the length of the sentence and the dimension of hidden features exponentially scale with 2, i.e.,  $\ell = 2^N$  and  $\log d \in \mathbb{N}^+$ . This setting aligns with the scaling of quantum computing, making it easier to understand. For other cases, padding techniques can be applied to expand the matrix and vector dimensions to conform to this requirement.

Since the runtime speedups depend heavily on the capabilities of the available input oracles, it is essential to specify the input oracles used in the quantum Transformer before detailing the algorithms. For the classical Transformers, the memory access to the inputs such as the sentence and the query, key, and value matrices is assumed. In the quantum version, we assume the access of several matrices via *block encoding techniques* introduced in Chapter 2.4.

**Assumption 5.1** (Input oracles). *Following the explicit form of the single-head and single-layer Transformer in Eqn. (5.5), there are five parameterized weight matrices, i.e.,  $W_q, W_k, W_v \in \mathbb{R}^{d \times d}$  in the attention block, as well as  $M_1 \in \mathbb{R}^{d' \times d}$  and  $M_2 \in \mathbb{R}^{d' \times d'}$  in FFN. Note that here,  $M_1$  and  $M_2$  are actually the transpose of parameterized matrices in the classical transformer. Quantum Transformer assumes access to these five parameterized weight matrices, as well as the input sequence  $S \in \mathbb{R}^{\ell \times d}$  via block encoding.*

Mathematically, given any  $A \in \{W_q, W_k, W_v, M_1, M_2, S\}$  corresponding to an  $N$ -qubit operator,  $\alpha, \varepsilon \geq 0$  and  $a \in \mathbb{N}$ , there exists a  $(a + N)$ -qubit unitary  $U_A$  referring to the  $(\alpha, a, \varepsilon)$ -block-encoding of  $A$  with

$$\|A - \alpha(\langle 0 |^{\otimes a} \otimes \mathbb{I}_{2^N})U_A(|0\rangle^{\otimes a} \otimes \mathbb{I}_{2^N})\| \leq \varepsilon, \quad (5.7)$$

where  $\|\cdot\|$  represents the spectral norm.

Under this assumption, we know that the quantum Transformer can access  $U_S$ ,  $U_{W_q}$ ,  $U_{W_k}$ , and  $U_{W_v}$  corresponding to the  $(\alpha_s, a_s)$ -encoding of  $S$  and

$(\alpha_w, a_w)$ -encodings of  $W_q$ ,  $W_k$  and  $W_v$ , respectively. Moreover, the quantum Transformer can access  $(\alpha_m, a_m)$ -encodings  $U_{M_1}$  and  $U_{M_2}$  corresponding two weight matrices  $M_1 \in \mathbb{R}^{d' \times d}$  and  $M_2 \in \mathbb{R}^{d \times d'}$  in FFN.

Remark

For simplicity and clarity, in the following, we consider the *perfect* block encoding of input matrices *without errors*, i.e.,  $\varepsilon = 0$ . As such, we will not explicitly write the error term of the block encoding, and use  $(\alpha, a)$  instead of  $(\alpha, a, 0)$ . The output of the quantum transformer is a quantum state corresponding to the probability distribution of the next token. The complete single-layer structure is described in Figure 5.2.

Under the above assumptions about access to the read-in protocols, the following theorem indicates how to implement a single-head and single-block transformer architecture in Eqn. (5.5) on the quantum computer.

**Theorem 5.2** (Quantum Transformer, informal). *For a single-head and single-block Transformer depicted in Figure 5.2, suppose its embedding dimension is  $d$  and its input sequence  $S$  has the length  $\ell = 2^N$ . Under Assumption 5.1 about the input oracles, for the index  $j \in [\ell]$ , one can construct an  $\epsilon$ -accurate quantum circuit for the quantum state proportional to*

$$\sum_{k=1}^d \text{Transformer}(S, j)_k |k\rangle, \quad (5.8)$$

by using  $\tilde{\mathcal{O}}(dN^2\alpha_s\alpha_w \log^2(1/\epsilon))$  times of the input block encodings.

We show this theorem by explicitly designing the quantum circuit for each computation block of the transformer architecture in a coherent way, i.e., without intermediate measurement. In addition, a subsequent transformation of the amplitude-encoded state, followed by measurement in the computational basis, yields the index of the next predicted token based on the probabilities modeled by the Transformer architecture.

**Roadmap.** In the remainder of this section, we detail the implementation of quantum Transformers, proceeding from the bottom to the top as illustrated in Figure 5.2. Specifically, we first demonstrate how to quantize the attention block  $\text{Attention}(S, j)$  in Chapter 5.2.1. Next, we present the quantization of residual connections and layer normalization operations (i.e., the

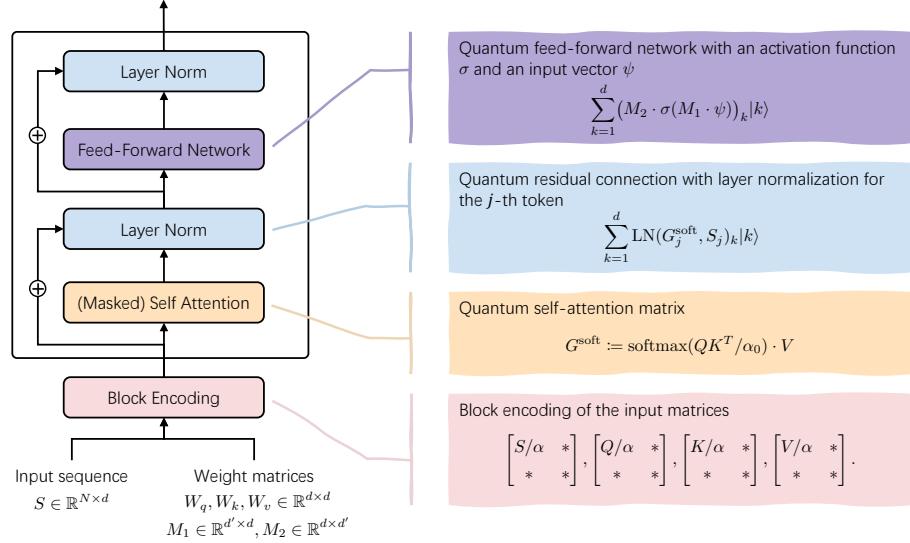


Figure 5.2: **Overview of the single-layer decoder-only quantum transformer.** A quantum transformer consists of a self-attention and a feed-forward network sub-layer, incorporating residual connections with layer normalization. The inputs of the quantum transformer are block encodings of matrices for the input sequence and pre-trained weights, from which the relevant matrices for the transformer are constructed (query  $Q$ , key  $K$ , and value  $V$ ). Each of the components accepts the block encoding from the prior component as the input and prepares a block encoding of the target matrix using quantum linear algebra as the output.

operation  $\text{LN}(\text{Attention}(S, j))$  in Eqn. (5.5)) in Chapter 5.2.2. Last, we exhibit the quantization of the fully connected neural network to complete the computation  $\text{LN}(\text{FFN}(\text{LN}(\text{Attention}(S, j))))$  in Chapter 5.2.3. This end-to-end approach ensures that the generated quantum state corresponds to the one described in Eqn. (5.8).

### 5.2.1 Quantum self-attention

We now describe how to achieve the quantum self-attention block, aiming to complete the computation

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^\top / \alpha_0)V =: G^{\text{soft}}$$

in Eqn. (5.1) on quantum computers. More specifically, under Assumption 5.1, the quantum self-attention block outputs a block encoding of a matrix  $G$  whose  $j$ -th row is the same as the output of the classical attention block, as described in the following theorem.

**Theorem 5.3** (Quantum self-attention). *Let  $\alpha_0 = \alpha_s^2 \alpha_w^2$ . For the index  $j \in [\ell]$ , one can construct a block encoding of a matrix  $G$  such that  $G_{j\star} = G_j^{\text{soft}} := (\text{softmax}(QK^\top/\alpha_0)V)_{j\star}$ .*

Remark

For quantum self-attention, we make a slight change by setting the scaling factor  $\alpha_0 = \alpha_s^2 \alpha_w^2$  for the following reasons. The first is that the usual setting  $\alpha_0 = 1/\sqrt{d}$  is chosen somehow in a heuristic sense, and there are already some classical works considering different scaling coefficients which may even achieve better performance (Yang et al., 2022; Ma et al., 2024). The second, which is more important, is that the quantum input assumption using the block encoding format naturally contains the normalization factor  $\alpha$  which plays a similar role to the scaling factor. Therefore, for the quantum case in the context of our work, it suffices to use  $\alpha$  directly.

The implementation of the quantum self-attention block can be decomposed into three steps:

1. Construction of the block encoding of the matrix  $QK^\top$  and the matrix  $V$  given access to  $U_S$ ,  $U_{W_q}$ ,  $U_{W_k}$ , and  $U_{W_v}$ ;
2. Implementation of the quantum algorithm to compute the softmax function  $\text{softmax}(QK^\top/\alpha_0)$  given access to  $U_{KQ}$ ;
3. Multiply with  $V$  via the product of block encodings.

In what follows, we iteratively detail the implementation of each step.

**Step I.** The construction of the block encoding of matrix  $QK^\top$  and  $V$  builds upon the employment of Fact 2.23. That is, given access to  $(\alpha, a)$ -encoding  $U_A$  of matrix  $A$  and  $(\beta, b)$ -encoding  $U_B$  of matrix  $B$ , an  $(\alpha\beta, a+b)$ -encoding can be constructed for the matrix  $AB$ . This result leads to the efficient construction of the block encodings of  $QK^\top$  and  $V$ , i.e.,

- For the matrix  $V = W_v S$ , it is straightforward to set  $A = W_v$  and  $B = S$  to construct the  $(\alpha_v, a_v)$ -encoding  $U_V$ , where  $\alpha_v = \alpha_s \alpha_w$  and  $a_v = a_s + a_w$ .

- For the matrix  $QK^\top$ , we first use Fact 2.23 to construct the  $(\alpha_q, a_q)$ -encoding  $U_Q$  and  $(\alpha_k, a_k)$ -encoding  $U_K$  with  $Q = W_q S$  and  $K = W_k S$ , respectively. Then, we use Fact 2.23 again to construct the  $(\alpha_0, a_0)$ -encoding  $U_{QK^\top}$  of  $QK^\top$ , where  $\alpha_0 = \alpha_s^2 \alpha_w^2$  and  $a_0 = 2a_s + 2a_w$ . Note that for a real matrix  $M$  and its block encoding unitary  $U_M$ ,  $U_M^\dagger$  is the block encoding of  $M^\top$ .

**Step II.** Once the unitary  $U_{QK^\top}$  is prepared, we move to implement the quantum algorithm corresponding to the softmax function, i.e.,  $\text{softmax}(QK^\top/\alpha_0)$ . Note that the softmax function relies on the exponential function, which is generally resource-intensive to implement on quantum computers. To circumvent this bottleneck, the quantum Transformer uses polynomial functions to approximate the softmax function, as supported by the following fact.

**Fact 5.4.** *For  $x \in [-1, 1]$ , the function  $f(x) := e^x$  can be approximated with error bound  $\epsilon$  with an  $\mathcal{O}(\log(1/\epsilon))$ -degree polynomial function.*

The insight provided by Fact 5.4 is to use polynomial functions to approximate the softmax function, i.e., we first approximate  $\exp \circ (QK^\top/\alpha_0)$  using polynomial functions, then multiply with different coefficients (normalization) for each row.

#### Remark

The notation  $\exp \circ (A)$  indicates that the exponential operation is applied elementwise to each entry of the matrix  $A$ , rather than representing a matrix exponential.

Moreover, the element-wise functions mean that functions are implemented on each matrix element.

In this context, the challenge of implementing a quantized softmax function reduces to the implementation of a quantized polynomial function. The key technique for achieving this lies in applying polynomial functions to each element of block-encoded matrices, as detailed in the following lemma.

**Lemma 5.5** (Element-wise polynomial function of block-encoded matrix). *Let  $N, k \in \mathbb{N}$ . Given access to an  $(\alpha, a)$ -encoding of a matrix  $A \in \mathbb{C}^{2^N \times 2^N}$  and an  $r$ -degree polynomial function  $f_r(x) = \sum_{j=1}^r c_j x^j$ ,  $c_j \in \mathbb{C}$  for  $j \in [r]$ , one can construct a  $(C, b)$ -encoding of  $f_r \circ (A/\alpha)$  by using  $\mathcal{O}(r^2)$  times the input unitary, where  $C := \sum_{j=1}^r |c_j|$ ,  $b := ra + (r-1)N + \lceil \log(r+1) \rceil$ .*

Moreover, for a polynomial function  $g_r(x) = \sum_{j=0}^r c_j x^j$  with constant term  $c_0$ , one can construct a  $(C', b)$ -encoding of  $g_r \circ (A/\alpha)$ , where  $C' = rc_0 + C$ .

*Proof of Lemma 5.5.* To achieve this implementation, we construct two state-preparation unitaries  $P_L$  and  $P_R$ , which act on  $\lceil \log(r+1) \rceil$  qubits such that

$$P_L : |0^{\lceil \log(r+1) \rceil}\rangle \rightarrow \frac{1}{\sqrt{C}} \sum_{j=1}^r \sqrt{|c_j|} |j\rangle, \quad (5.9)$$

$$P_R : |0^{\lceil \log(r+1) \rceil}\rangle \rightarrow \frac{1}{\sqrt{C}} \sum_{j=1}^r \sqrt{|c_j|} e^{i\theta_j} |j\rangle, \quad (5.10)$$

where  $C = \sum_{j=1}^r |c_j|$  and  $|c_j|e^{i\theta_j} = c_j$ . These two unitaries encode the polynomial coefficients  $\{c_j\}$  into the quantum circuit, which is needed for block encoding via the linear combination of unitaries indicated in fact 2.21. Note that the construction of  $P_L$  and  $P_R$  is efficient for small  $r$ , as the corresponding circuit is  $\mathcal{O}(r)$ -depth with only elementary quantum gates (Sun et al., 2023; Zhang et al., 2022a).

For  $j \in [r]$ , let  $U_{A^j}$  be the  $(1, ja + (j - 1)N)$ -encoding of

$$(A/\alpha)^{\circ j} := \underbrace{(A/\alpha) \circ (A/\alpha) \circ \cdots \circ (A/\alpha)}_{j-1 \text{ times of Hadamard product}},$$

which is constructed by iteratively applying lemma 2.25. For simplicity, these block encodings  $U_{A^j}$  can also be considered as  $(1, ra + (r - 1)N)$ -encoding of  $(A/\alpha)^{\circ j}$ . Then, we construct a  $(ra + rN + \lceil \log(r+1) \rceil)$ -qubit unitary  $W = \sum_{j=1}^r |j\rangle\langle j| \otimes U_{A^j} + (\mathbb{I}_{2^{\lceil \log(r+1) \rceil}} - \sum_{j=1}^r |j\rangle\langle j|) \otimes \mathbb{I}_{2^{(ra+rN)}}$ . Taking the linear combination of block encodings via fact 2.21, we can implement a  $(C, ra + (r - 1)N + \lceil \log(r+1) \rceil)$ -encoding of  $f_r \circ (A/\alpha)$ .

To implement element-wise functions including constant terms, we also need access to the block encoding of a matrix whose elements are all 1. Notice that this matrix can be written as the linear combination of the identity matrix and the reflection operator, i.e.,

$$\sum_{k,k' \in [2^N]} |k\rangle\langle k'| = \frac{2^N}{2} \left( \mathbb{I}_{2^N} - \left( \mathbb{I}_{2^N} - \frac{2}{2^N} \sum_{k,k' \in [2^N]} |k\rangle\langle k'| \right) \right) \quad (5.11)$$

$$= \frac{2^N}{2} \left( \mathbb{I}_{2^N} - H^{\otimes N} \left( \mathbb{I}_{2^N} - 2 |0^N\rangle\langle 0^N| \right) H^{\otimes N} \right), \quad (5.12)$$

where  $H$  is the Hadamard gate. Define  $U_{\text{ref}} = |0\rangle\langle 0| \otimes \mathbb{I}_{2^N} + |1\rangle\langle 1| \otimes (H^{\otimes N}(\mathbb{I}_{2^N} - 2|0^N\rangle\langle 0^N|)H^{\otimes N})$ . By direct computation, one can show that  $U_0 = (XH \otimes \mathbb{I}_{2^N})U_{\text{ref}}(H \otimes \mathbb{I}_{2^N})$  is an  $(2^N, 1)$ -encoding of  $\sum_{k,k'} |k\rangle\langle k'|$ . One can achieve the element-wise function by following the same steps as above and taking linear combinations among  $U_0, \dots, U_{A^r}$ . One point to notice is that we can only construct  $(2^N, 1)$ -encoding of the matrix whose elements are all 1 since the spectral norm of this matrix is  $2^N$ . Therefore, we encode  $2^N c_0$  into the state instead of  $c_0$  to amplify the constant term.  $\square$

Supported by the above lemma, we can complete Step II (i.e., the quantum softmax for self-attention), as shown in the following theorem.

**Theorem 5.6** (Quantum softmax for self-attention, informal). *Given an  $(\alpha, a)$ -encoding  $U_A$  of a matrix  $A \in \mathbb{R}^{\ell \times \ell}$ , a positive integer  $d$ , and an index  $j \in [\ell]$ , one can prepare a state-encoding of*

$$|A_j\rangle := \sum_{k=1}^{\ell} \sqrt{\text{softmax}(A/\alpha)_{jk}} |k\rangle = \frac{1}{\sqrt{Z_j}} \sum_{k=1}^{\ell} \exp \circ \left( \frac{A}{2\alpha} \right)_{jk} |k\rangle,$$

where  $Z_j = \sum_{k=1}^{\ell} \exp \circ (A/\alpha)_{jk}$ .

*Proof sketch of Theorem 5.6.* We first construct the block encoding of  $\exp \circ (\frac{A}{2\alpha})$ . Note that Taylor expansion of  $\exp(x)$  contains a constant term 1. This can be achieved with lemma 5.5 and fact 5.4. Here, since we are only focusing on the  $j$ -th row, instead of taking linear combination with the matrix whose elements are all 1, we take sum with the matrix whose  $j$ -th row elements are all 1 and else are 0. This enables us to have a better dependency on  $\ell$ , i.e., from  $\ell$  to  $\sqrt{\ell}$ . For index  $j \in [\ell]$ , let  $U_j : |0\rangle \rightarrow |j\rangle$ . One can achieve this by changing Eqn. (5.12) to the following,

$$\sum_k |j\rangle\langle k| = \frac{\sqrt{\ell}}{2} (U_j H^{\otimes N} - U_j (\mathbb{I}_{2^N} - 2|0^N\rangle\langle 0^N|)H^{\otimes N}). \quad (5.13)$$

Following the same steps in lemma 5.5, one can achieve the construction. There are two error terms in this step. Denote  $U_{f \circ (A)}$  as the constructed block encoding unitary. By lemma 5.5 and some additional calculation, one can show that  $U_{f \circ (A)}$  is a block-encoding of  $\exp \circ (\frac{A}{2\alpha})$ . Note that  $\exp \circ (\frac{A}{2\alpha})_{jk} = \exp \circ (\frac{A}{2\alpha})_{kj}^\top$ . With unitary  $U_{f \circ (A)}^\dagger(I \otimes U_j)$  and amplitude amplification, one can prepare a state-encoding of the target state

$$|A_j\rangle := \frac{1}{\sqrt{Z_j}} \sum_{k=1}^{\ell} \exp \circ \left( \frac{A}{2\alpha} \right)_{jk} |k\rangle, \quad (5.14)$$

where  $Z_j = \sum_{k=1}^{\ell} \exp \circ (A/\alpha)_{jk}$  is the normalization factor of softmax function for the  $j$ -th row.  $\square$

**Step III.** Finally, we implement the matrix multiplication with  $V$ . This can be easily achieved by using Fact 2.23, with  $U_{f(QK^\top)}^\dagger$  and  $U_V$ . Consequently, we obtain an encoded quantum state analogous to

$$\sum_k^{\ell} (\text{softmax}(QK^\top/\alpha_0)V)_{jk} |k\rangle. \quad (5.15)$$

Combining the results of Steps I, II, and III, we are now ready to present the proof of Theorem 5.3.

*Proof of Theorem 5.3.* In the first step, we construct the block encoding of matrix  $QK^\top$  and  $V$ . Note that for a real matrix  $M$  and its block encoding unitary  $U_M$ ,  $U_M^\dagger$  is the block encoding of  $M^\top$ . By Fact 2.23, one can construct an  $(\alpha_0, a_0)$ -encoding  $U_{QK^\top}$  of  $QK^\top$ , where  $\alpha_0 := \alpha_s^2 \alpha_w^2$  and  $a_0 = 2a_s + 2a_w$ . One can also construct an  $(\alpha_v, a_v)$ -encoding  $U_V$  of  $V$ , where  $\alpha_v = \alpha_s \alpha_w$  and  $a_v = a_s + a_w$ .

By theorem 5.6, using  $U_{QK^\top}$  one can prepare a state-encoding of the state

$$\sum_{k=1}^{\ell} \sqrt{\text{softmax}(QK^\top/\alpha_0)_{jk}} |k\rangle,$$

where  $Z_j = \sum_{k=1}^{\ell} \exp \circ (QK^\top/\alpha_0)_{jk}$ . Remember that state encoding is also a block encoding. By lemma 2.25, one can construct a block encoding of a matrix whose  $j$ -th column is

$$[\text{softmax}(QK^\top/\alpha_0)_{j1}, \dots, \text{softmax}(QK^\top/\alpha_0)_{j\ell}]$$

ignoring other columns. Let this block-encoding unitary be  $U_{f(QK^\top)}$ .

Last, by exploiting Fact 2.23 again, with  $U_{f(QK^\top)}^\dagger$  and  $U_V$ , we obtain an encoded quantum state analogous to

$$\sum_k^{\ell} (\text{softmax}(QK^\top/\alpha_0)V)_{jk} |k\rangle.$$

$\square$

### Extension to implement quantum masked self-attention

Now we consider how to implement the *masked* self-attention, which is essential for the decoder-only structure. This can be achieved by slightly changing some steps as introduced in previous theorems.

**Corollary 5.7** (Quantum masked self-attention). *For the index  $j \in [\ell]$ , one can construct a block encoding of a matrix  $G^{\text{mask}}$  such that  $G_{j\star}^{\text{mask}} = (\text{softmax}(\frac{QK^\top}{\alpha_0} + M)V)_{j\star}$ , where  $M$  is the masked matrix as Eqn. (5.2),  $Z_j = \sum_{k=1}^{\ell} \exp \circ (\frac{QK^\top}{\alpha_0} + M)_{jk}$ .*

*Proof of Corollary 5.7.* To achieve masked self-attention, we slightly change the steps mentioned in theorem 5.6.

First, to approximate the exponential function, we move beyond using a matrix where all elements in the  $j$ -th row are set to 1 while other rows remain 0. Instead, we refine this approach by considering only the first  $2^{\lceil \log(j+1) \rceil}$  elements in the  $j$ -th row to be 1. Note that this matrix can be achieved similarly to the original one. The encoding factor of this matrix is  $2^{\lceil \log(j+1) \rceil}/2$ .

Second, after approximating the function, for index  $j \in [\ell]$ , we multiply the block encoding with a projector  $\sum_{k,k \leq j} |k\rangle\langle k|$  to mask the elements. Though the projector  $\sum_{k \in \mathcal{S}} |k\rangle\langle k|$  for  $\mathcal{S} \subseteq [\ell]$  is not unitary in general, one can construct a block encoding of the projector by noticing that it can be written by the linear combination of two unitaries:

$$\sum_{k \in \mathcal{S}} |k\rangle\langle k| = \frac{1}{2}\mathbb{I} + \frac{1}{2}\left(2\sum_{k \in \mathcal{S}} |k\rangle\langle k| - \mathbb{I}\right). \quad (5.16)$$

Define  $U_{\text{proj}} := |0\rangle\langle 0| \otimes \mathbb{I} + |1\rangle\langle 1| \otimes (2\sum_{k \in \mathcal{S}} |k\rangle\langle k| - \mathbb{I})$ . One can easily verify that  $(H \otimes \mathbb{I})U_{\text{proj}}(H \otimes \mathbb{I})$  is a  $(1, 1, 0)$ -encoding of  $\sum_{k \in \mathcal{S}} |k\rangle\langle k|$ , where  $H$  is the Hadamard gate. The following steps follow the same with theorem 5.6 and theorem 5.3.  $\square$

One may further achieve the multi-head self-attention case by using the linear combination of unitaries.

### 5.2.2 Quantum residual connection and layer normalization

In this subsection, we discuss how to implement the residual connection with layer normalization on a quantum computer. We continue based on the result in theorem 5.3 to implement the Layer Norm block shown in Figure 5.2.

**Theorem 5.8** (Quantum residual connection with layer normalization). *Given access to the block encoding of the matrix  $G$  in Theorem 5.3, One can construct a quantum-encoded state*

$$\sum_{k=1}^d \text{LN}(G_j^{\text{soft}}, S_j)_k |k\rangle = \frac{1}{\varsigma} \sum_{k=1}^d (G_{jk}^{\text{soft}} + S_{jk} - \bar{s}_j) |k\rangle,$$

where  $\bar{s}_j := \frac{1}{d} \sum_{k=1}^d (G_{jk}^{\text{soft}} + S_{jk})$  and  $\varsigma := \sqrt{\sum_{k=1}^d (G_{jk}^{\text{soft}} + S_{jk} - \bar{s}_j)^2}$ .

*Proof of sketch of Theorem 5.8.* As shown in theorem 5.3, we can construct a block-encoding of a matrix  $G$  whose  $j$ -th row is the same row as that of  $G^{\text{soft}}$ . By Assumption 5.1, we are given  $U_s$  which is an  $(\alpha_s, a_s)$ -encoding of  $S$ . By Lemma 2.24 with the state preparation pair  $(P, P)$  such that

$$P|0\rangle = \frac{1}{\sqrt{\alpha_g + \alpha_s}} (\sqrt{\alpha_g}|0\rangle + \sqrt{\alpha_s}|1\rangle), \quad (5.17)$$

one can construct a quantum circuit  $U_{\text{res}}$  which is an  $(\alpha_g + \alpha_s, a_g + 1)$ -encoding of an  $\ell \times d$  matrix whose  $j$ -th row is the same as that of  $G^{\text{soft}} + S$ .

Now we consider how to create a block encoding of a diagonal matrix  $\bar{s}_j \cdot \mathbb{I}$ , where  $\bar{s}_j := \frac{1}{d} \sum_{k=1}^d (G_{jk}^{\text{soft}} + S_{jk})$ . Let us define a unitary  $H^{\log d} := H^{\otimes \log d}$ . Note that  $H^{\log d}$  is a  $(1, 0, 0)$ -encoding of itself, and the first column of  $H^{\log d}$  is  $\frac{1}{\sqrt{d}}(1, \dots, 1)^\top$ . By fact 2.23, one can multiply  $G^{\text{soft}} + S$  with  $H^{\log d}$  to construct a block encoding of an  $\ell \times d$  matrix, whose  $(j, 1)$ -element is  $\sqrt{d}\bar{s}_j$ . One can further move this element to  $(1, 1)$  by switching the first row with the  $j$ -th row. By tensor product with the identity  $\mathbb{I}$  of  $\log d$  qubits, one can construct a block encoding of  $\sqrt{d}\bar{s}_j \cdot \mathbb{I}$ .

With  $U_j : |0\rangle \rightarrow |j\rangle$ , one can prepare the state

$$U_{\text{res}}^\dagger(\mathbb{I} \otimes U_j)|0\rangle|0\rangle = \frac{1}{\alpha_g + \alpha_s} |0\rangle \sum_{k=1}^d \psi_k |k\rangle + \sqrt{1 - \frac{\sum_k \psi_k^2}{(\alpha_g + \alpha_s)^2}} |1\rangle |\text{bad}\rangle. \quad (5.18)$$

By the diagonal block encoding of amplitudes mentioned as fact 2.28, this can be converted to a block encoding of the diagonal matrix  $\text{diag}(G_{j1} + S_{j1}, \dots, G_{jd} + S_{jd})$ .

By taking the linear combination as fact 2.24 with state preparation pair  $(P_1, P_2)$ , where

$$P_1|0\rangle = \frac{1}{\sqrt{1 + 1/\sqrt{d}}} (|0\rangle + \frac{1}{\sqrt{d}}|1\rangle) \quad (5.19)$$

and

$$P_2 |0\rangle = \frac{1}{\sqrt{1+1/\sqrt{d}}}(|0\rangle - \frac{1}{\sqrt{d}}|1\rangle), \quad (5.20)$$

one can construct a block encoding of  $\text{diag}(G_{j1} + S_{j1} - \bar{s}_j, \dots, G_{jd} + S_{jd} - \bar{s}_j)$ , and we call it  $U_{\text{LN}}$ . Then the unitary  $U_{\text{LN}}(\mathbb{I} \otimes H^{\log d})$  is an state-encoding of the state

$$\frac{1}{\varsigma} \sum_{k=1}^d (G_{jk}^{\text{soft}} + S_{jk} - \bar{s}_j) |k\rangle,$$

where  $\varsigma := \sqrt{\sum_{k=1}^d (G_{jk}^{\text{soft}} + S_{jk} - \bar{s}_j)^2}$ .  $\square$

### 5.2.3 Quantum feedforward neural network

We turn our attention to the third main building block of the transformer architecture, the feed-forward neural network. This block often is a relatively shallow neural network with linear transformations and ReLU activation functions (Vaswani, 2017). More recently, activation functions such as the GELU have become popular, being continuously differentiable. We highlight that they are ideal for quantum Transformers, since the QSVT framework requires functions that are well approximated by polynomial functions. Functions like  $\text{ReLU}(x) = \max(0, x)$  can not be efficiently approximated. The GELU is constructed from the error function, which is efficiently approximated as follows.

**Fact 5.9** (Polynomial approximation of error function (Low, 2017)). *Let  $\epsilon > 0$ . For every  $k > 0$ , the error function  $\text{erf}(kx) := \frac{2}{\sqrt{\pi}} \int_0^{kx} e^{-t^2} dt$  can be approximated with error up to  $\epsilon$  by a polynomial function with degree  $\mathcal{O}(k \log(\frac{1}{\epsilon}))$ .*

This lemma implies the following efficient approximation of the GELU function with polynomials.

**Corollary 5.10** (Polynomial approximation of GELU function). *Let  $\epsilon > 0$  and  $\lambda \in \mathcal{O}(1)$ . For every  $k > 0$  and  $x \in [-\lambda, \lambda]$ , the GELU function  $\text{GELU}(kx) := kx \cdot \frac{1}{2}(1 + \text{erf}(\frac{kx}{\sqrt{2}}))$  can be approximated with error up to  $\epsilon$  by a polynomial function with degree  $\mathcal{O}(k \log(\frac{k\lambda}{\epsilon}))$ .*

*Proof of Corollary 5.10.* It suffices to approximate the error function with precision  $\frac{\epsilon}{k\lambda}$  by fact 5.9.  $\square$

In the following theorem, we consider how to implement the two-layer feedforward network on quantum computers. As mentioned, the GELU function is widely used in transformer-based models and we explicitly consider it as the activation function in the theorem. Cases for other activation functions like sigmoid follow the same analysis. An example is the  $\tanh(x)$  function, which can be well approximated by a polynomial for  $x \in [-\pi/2, \pi/2]$  (Guo et al., 2024b).

**Theorem 5.11** (Two-layer feedforward network with GELU function, informal). *Assume we have access to  $(\alpha, a)$ -state encoding of an  $N$ -qubit state  $|\psi\rangle = \sum_{k=1}^{2^N} \psi_k |k\rangle$ , where  $\{\psi_k\}$  are real and  $\|\psi\|_2 = 1$ . Further, assume access to  $(\alpha_m, a_m)$ -encodings  $U_{M_1}$  and  $U_{M_2}$  of weight matrices  $M_1 \in \mathbb{R}^{d' \times d}$  and  $M_2 \in \mathbb{R}^{d \times d'}$ . Let the activation function be  $\text{GELU}(x) := x \cdot \frac{1}{2}(1 + \text{erf}(\frac{x}{\sqrt{2}}))$ . One can prepare a state-encoding of the state*

$$|\phi\rangle = \frac{1}{C} \sum_{k=1}^d \left( M_2 \cdot \text{GELU}(M_1 \cdot \psi) \right)_k |k\rangle, \quad (5.21)$$

where  $C$  is the normalization factor.

*Proof of sketch of Theorem 5.11.* We have

$$(\mathbb{I}_{2^a} \otimes U_{M_1})(\mathbb{I}_{2^{a_m}} \otimes U_\psi) |0^{a+a_m+N}\rangle = \frac{1}{\alpha\alpha_m} |0^{a+a_m}\rangle M_1 |\psi\rangle + |\tilde{\perp}\rangle, \quad (5.22)$$

where  $|\tilde{\perp}\rangle$  is an unnormalized orthogonal state. For the case  $d' \geq \ell$ , this can be achieved by padding ancilla qubits to the initial state. By fact 2.28, one can construct a block encoding of the diagonal matrix  $\text{diag}((M_1\psi)_1, \dots, (M_1\psi)_{d'})$ . Note that the GELU function does not have a constant term, and is suitable to use the importance-weighted amplitude transformation as in Rattew and Rebentrost (2023). Instead of directly implementing the GELU function, we first implement the function  $f(x) = \frac{1}{2}(1 + \text{erf}(\frac{x}{\sqrt{2}}))$ . Note that the value of  $|\text{erf}(x)|$  is upper bounded by 1. By fact 2.28 with function  $\frac{1}{4}(1 + \text{erf}(\alpha\alpha_m\frac{x}{\sqrt{2}}))$ , one can construct a block encoding of matrix  $\text{diag}(f(M_1\psi)_1, \dots, f(M_1\psi)_{d'})$ .

Let the previously constructed block-encoding unitary be  $U_{f(x)}$ . We have

$$U_{f(x)}(\mathbb{I} \otimes U_{M_1})(\mathbb{I} \otimes U_\psi) |0\rangle |0\rangle = \frac{1}{2\alpha\alpha_m} |0\rangle \sum_k \text{GELU}(M_1\psi)_k |k\rangle + |\tilde{\perp}'\rangle, \quad (5.23)$$

where  $|\widetilde{\perp}'\rangle$  is an unnormalized orthogonal state. Finally, by implementing the block-encoding unitary  $U_{M_2}$ , we have

$$\begin{aligned} & (\mathbb{I} \otimes U_{M_2})(\mathbb{I} \otimes U_{f(x)})(\mathbb{I} \otimes U_{M_1})(\mathbb{I} \otimes U_\psi) |0\rangle |0\rangle \\ &= \frac{C}{2\alpha_m^2} |0\rangle \sum_{k=1}^d \left( M_2 \cdot \text{GELU}(M_1 \cdot \psi) \right)_k |k\rangle + |\widetilde{\perp}''\rangle, \end{aligned} \quad (5.24)$$

where  $C$  is the normalization factor, and  $|\widetilde{\perp}''\rangle$  is an unnormalized orthogonal state.  $\square$

#### Remark

The quantum feedforward network discussed in this subsection is a quantum implementation of the classical feedforward network under the input assumption of block encoding, which is essentially different from the quantum analog of neural networks introduced in chapter 4.

## 5.3 Runtime Analysis with Quadratic Speedups

In this section, we provide a combined analytical and numerical analysis to explore the potential of a quantum speedup in time complexity.

### 5.3.1 Overview

Having the quantum implementation of self-attention, residual connection, layer normalization, and feed-forward networks, we are able to construct the quantum transformer by combining these building blocks as in theorem 5.2.

We obtain this final complexity on the basis of the following considerations: the single-head and single-block transformer architecture includes one self-attention, one feed-forward network, and two residual connections with layer normalization, as shown in Figure 5.2.

Starting from the input assumption as Assumption 5.1, for the index  $j \in [\ell]$ , we first construct the block encoding of self-attention matrix, as described in Section 5.2.1. This output can be directly the input of the quantum residual connection and layer normalization, as Section 5.2.2, whose output is a state encoding. Remind the definition of state encoding as definition 2.22. The state encoding can directly be used as the input of the feed-forward network, as section 5.2.3. Finally, we put the output of the feed-forward network, which is a state encoding, into the residual connection block. This is possible by noticing that state encoding is a specific

kind of block encoding. Multiplying the query complexity of these computational blocks, one can achieve final result. The detailed analysis of runtime is referred to [Guo et al. \(2024a\)](#)

As theorem 5.2 shows, the quantum transformer uses  $\tilde{\mathcal{O}}(dN^2\alpha_s\alpha_w)$  times the input block encodings, where  $\alpha_s$  and  $\alpha_w$  are encoding factors. By analyzing naive matrix multiplication, the runtime of classical single-head and single-block Transformer during the inference stage is  $\mathcal{O}(\ell d + d^2)$ , where  $d$  is the embedding dimension and  $\ell = 2^N$  is the input sequence length. From the comparison, we can see that  $\alpha_s$  and  $\alpha_w$  are the dominant factors that affect the potential quantum speedup. We will explore the properties of these two factors via numerical studies.

### 5.3.2 Numerical evidence

The encoding factors  $\alpha_s$  and  $\alpha_w$  appear in the block encodings of matrices  $S$  and  $W_q, W_k, W_v$ . Recall that the encoding factor  $\alpha$  is lower bounded by the spectral norm of a block-encoded matrix  $A$ , i.e.,  $\alpha \geq \|A\|$ . Given access to quantum Random Access Memory (QRAM) and a quantum data structure ([Lloyd et al., 2014](#); [Kerenidis and Prakash, 2016](#)), there are well-known implementations that enable the construction of a block encoding for an arbitrary matrix  $A$  where the encoding factor is upper bounded by the Frobenius norm  $\|A\|_F$ . Based on these considerations, we numerically study these two norms of the input matrices of several open-source large language models<sup>1</sup> to provide upper and lower bound of  $\alpha_s$  and  $\alpha_w$ .

We first investigate the input sequence matrix  $S$ , which introduces the dependency on  $\ell$ . We consider input data in real-world applications sampled from the widely-used Massive Multitask Language Understanding (MMLU) dataset ([Hendrycks et al., 2021](#)) covering 57 subjects across science, technology, engineering, mathematics, the humanities, the social sciences, and more. The scaling of the spectral norm and Frobenius norm of  $S$  on the MMLU dataset is demonstrated in Figure 5.3. We can find that the matrix norms of the input matrix of all LLMs scale at most as  $\mathcal{O}(\sqrt{\ell})$ .

As additional interest, this analysis of the matrix norm provides new insights for the classical tokenization and embedding design. We also observe that comparatively more advanced LLM models like *Llama2-7b* and *Mistral-7b* have large variances of matrix norms. This phenomenon is arguably the consequence of the training in those models; the embeddings that frequently appear in the real-world dataset are actively updated at the pre-training stage, and therefore, are more broadly distributed.

We then compute the spectral and Frobenius norms of weight matrices

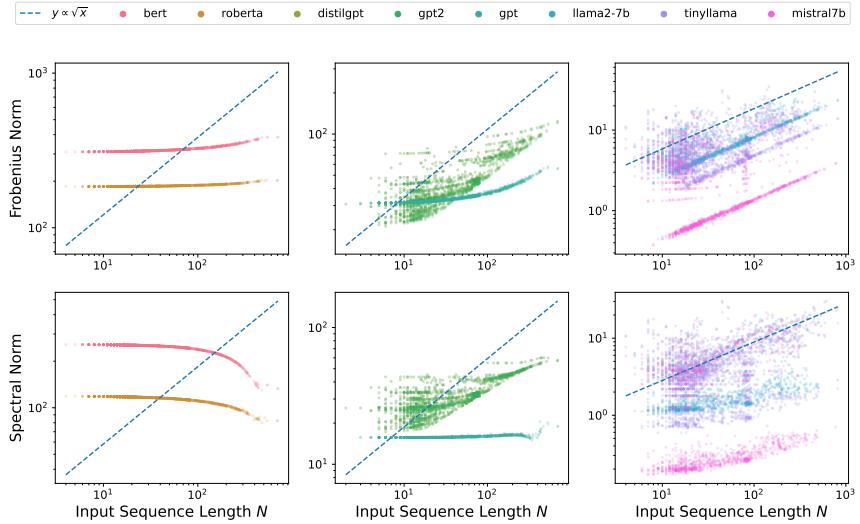
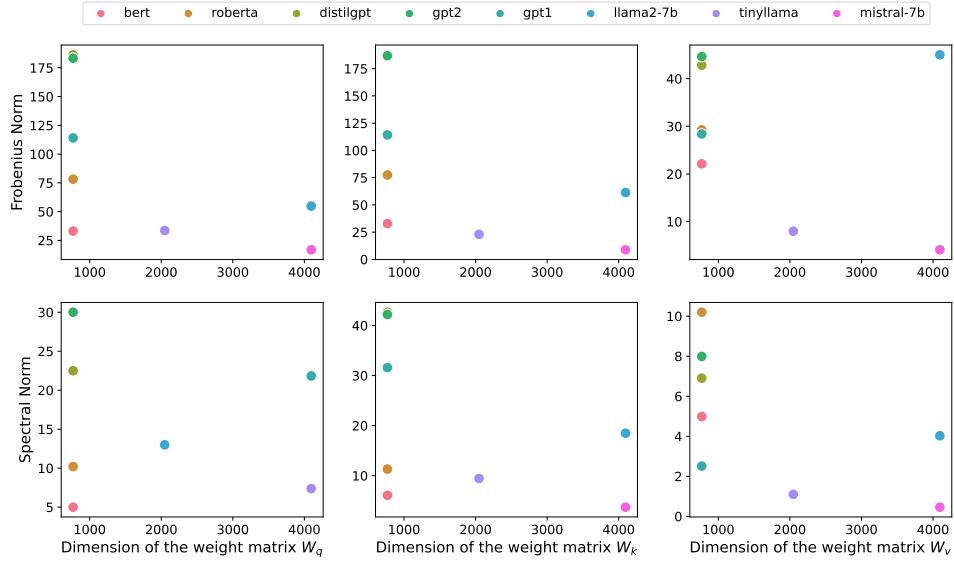


Figure 5.3: Scaling of the spectral norm  $\|S\|$  and the Frobenius norm  $\|S\|_F$  with  $\ell$  for each model, displayed on logarithmic scales for both axes. For reference, the line  $y \propto \sqrt{x}$  is also shown. We use tokens in MMLU dataset and convert them to  $S$ .

$(W_q, W_k, W_v)$  for the large language models. The result can be seen in Figure 5.4. Many of the LLMs below a dimension  $d$  of  $10^3$  that we have checked have substantially different norms. We observe that for larger models such as *Llama2-7b* and *Mistral-7b*, which are close to the current state-of-the-art open-source models, the norms do not change dramatically. Therefore, it is reasonable to assume that the spectral norm and the Frobenius norm of the weight matrices are at most  $\mathcal{O}(\sqrt{d})$  for advanced LLMs.

Given these numerical experiments it is reasonable to assume that  $\alpha_s = \mathcal{O}(\sqrt{\ell})$  and  $\alpha_w = \mathcal{O}(\sqrt{d})$ , and we obtain a query complexity of the quantum transformer in  $\tilde{\mathcal{O}}(d^{\frac{3}{2}}\sqrt{\ell})$ . We continue with a discussion of the possible time complexity. With the QRAM assumption, the input block encodings can be implemented in a polynomially logarithmic time of  $\ell$ . Even without a QRAM assumption, there can be cases when the input sequence is generated efficiently, for example when the sequence is generated from a differential equation, see additional discussions in the supplementary material. In these cases, we demonstrate that a quadratic speedup of the runtime of a single-head and single-block Transformer can be expected.



**Figure 5.4: Norms of weight matrices across open-source LLMs.**  
The figure shows the maximum Frobenius norm and spectral norm values among weight matrices  $W_q$ ,  $W_k$ , and  $W_v$  across the eight LLM models.

## 5.4 Code Demonstration

We explain how self-attention works with a simple concrete example. Consider a short sequence of three words: “The cat sleeps”.

First, we convert each word into an embedding vector. For this toy example, we use very small 4-dimensional embeddings:

```

1 The      = [1, 0, 1, 0]
2 cat      = [0, 1, 1, 1]
3 sleeps  = [1, 1, 0, 1]
```

In self-attention, each word needs to “attend” to all other words in the sequence. This happens through three key steps using learned weight matrices ( $W_q$ ,  $W_k$ ,  $W_v$ ) to transform the embeddings into queries, keys, and values. When we multiply our word embeddings by these matrices, we get

```

1 import numpy as np
2
3 # Input tokens (3 tokens, each with 3 features)
4 S = np.array([
    [1, 0, 1, 0],
    [0, 1, 1, 1],
    [1, 1, 0, 1]]))
```

```

5      [1, 0, 1, 0], # for "The"
6      [0, 1, 1, 1], # for "cat"
7      [1, 1, 0, 1] # for "sleeps"
8  ])
9
10 # Initialize weights for Query, Key, and Value (4x3
11 #   matrices)
12 W_q = np.array(
13   [
14     [0.2, 0.4, 0.6, 0.8],
15     [0.1, 0.3, 0.5, 0.7],
16     [0.9, 0.8, 0.7, 0.6],
17     [0.5, 0.4, 0.3, 0.2],
18   ]
19 )
20 W_k = np.array(
21   [
22     [0.1, 0.3, 0.5, 0.7],
23     [0.6, 0.4, 0.2, 0.1],
24     [0.8, 0.9, 0.7, 0.6],
25     [0.2, 0.1, 0.3, 0.4],
26   ]
27 )
28 W_v = np.array(
29   [
30     [0.3, 0.5, 0.7, 0.9],
31     [0.6, 0.4, 0.2, 0.1],
32     [0.8, 0.9, 0.7, 0.6],
33     [0.5, 0.4, 0.3, 0.2],
34   ]
35 )
36
37 # Compute Query, Key, and Value matrices
38 Q = S @ W_q
39 K = S @ W_k
40 V = S @ W_v

```

Next, we compute attention scores by multiplying  $Q$  and  $K^\top$ , then applying softmax.

```

1 # Compute scaled dot-product attention
2 d = Q.shape[1] # Feature dimension
3 attention_scores = Q @ K.T / np.sqrt(d)

```

```

4
5     def softmax(x):
6         """Compute softmax values for each set of scores in x
7             """
8         return np.exp(x) / np.sum(np.exp(x), axis=1, keepdims
9             =True)
10
11 attention_weights = softmax(attention_scores)

```

The attention weight would be:

$$\text{softmax}\left(\frac{Q \cdot K^\top}{\sqrt{4}}\right) = \begin{bmatrix} 0.324 & 0.467 & 0.209 \\ 0.305 & 0.515 & 0.180 \\ 0.346 & 0.432 & 0.222 \end{bmatrix}.$$

The final output captures how each word relates to every other word in the sentence. In this case, “sleeps” pays most attention to “cat” (0.432), some attention to “The” (0.346), and less attention to itself (0.222). Finally, we use these scores to create a weighted sum of the values:

```

1 output = attention_weights @ V

```

The final output is

$$\text{output} = \begin{bmatrix} 1.536 & 1.519 & 1.265 & 1.157 \\ 1.566 & 1.536 & 1.261 & 1.137 \\ 1.512 & 1.507 & 1.269 & 1.174 \end{bmatrix}.$$

## 5.5 Bibliographic Remarks

Transformer architecture has profoundly revolutionized AI and has broad impacts. As classical computing approaches its physical limitations, it is important to ask how we can leverage quantum computers to advance Transformers with better performance and energy efficiency. Besides the fault-tolerant quantum Transformers introduced in Chapter 5.2, multiple works are advancing this frontier from various perspectives.

One aspect is to design novel quantum neural network architectures introduced in Chapter 4 with the intuition from the Transformer, especially the design of the self-attention block. In particular, Li et al. (2023b) proposes the quantum self-attention neural networks and verifies their effectiveness with the synthetic data sets of quantum natural language processing. There are several follow-up works along this direction (Cherrat et al., 2024; Evans et al., 2024; Widdows et al., 2024).

Another research direction is exploring how to utilize quantum processors to advance certain parts of the transformer architecture. Specifically, Gao et al. (2023) considers how to compute the self-attention matrix under sparse assumption and shows a quadratic quantum speedup. Liu et al. (2024b) harnesses quantum neural networks to generate weight parameters for the classical model. In addition, Liu et al. (2024a) devises a quantum algorithm for the training process of large-scale neural networks, implying an exponential speedup under certain conditions. There are several other works that consider machine learning related optimization problems (Yang et al., 2023; Zhang and Li, 2024; Wang et al., 2024c; Rebentrost et al., 2018b).

Despite the progress made, several important questions remain unresolved. Among them, one key challenge is devising efficient methods to encode classical data or parameters onto quantum computers. Currently, most quantum algorithms can only implement one or at most constant layers of Transformers (Guo et al., 2024a; Liao and Ferrie, 2024; Khatri et al., 2024) without quantum tomography. Are there effective methods that can be generalized to multiple layers, or is achieving this even necessary? Moreover, given the numerous variants of the classical Transformer architecture, can these variants also benefit from the capabilities of quantum computers? Lastly, if one considers training a model directly on a quantum computer, is it possible to do so in a “quantum-native” manner—avoiding excessive data read-in and read-out operations?

# Chapter 6

## Conclusion

In this tutorial, we systematically explored the landscape of quantum machine learning, covering foundational principles, the adaptation of classical models to quantum frameworks, and the theoretical underpinnings of quantum algorithms. By providing practical implementations and discussing emerging research directions, we aimed to bridge the gap between classical AI and quantum computing for researchers and practitioners.

The insights gained from this tutorial highlight how quantum machine learning has the potential to revolutionize various domains, from fundamental scientific research to practical applications in industry. As quantum hardware continues to evolve, quantum machine learning will likely play a central role in harnessing the computational advantages of quantum systems. In the meantime, the field of quantum machine learning faces challenges, as discussed at the end of each chapter. Overcoming these barriers will be critical for unlocking the full potential of quantum machine learning.

Moving forward, interdisciplinary collaboration between quantum computing and AI researchers will be essential for addressing these challenges and realizing the transformative potential of QML. All in all, we hope this tutorial serves as a valuable resource for those eager to contribute to this rapidly evolving discipline.



# Appendix A

## Notations Summary

Notation	Concept
$a, b, \mathbf{a}_j, \mathbf{b}_j, \alpha, \beta$	Scalars
$\mathbf{x}, \mathbf{y}$	Vectors
$W, A$	Matrices
$\mathbb{R}$	Real Euclidean space
$\mathbb{C}$	Complex Euclidean space
$\mathbb{N}$	The set of natural numbers
$[a]$	The set of integers $\{1, 2, \dots, a\}$
$\mathbb{E}[\cdot]$	Expectation value of a random variable
$\text{Var}[\cdot]$	Variance of a random variable
$\mathcal{O}$	Asymptotic upper bound notation
$\Omega$	Asymptotic lower bound notation
$\top$	Transpose operation
$a^*$	complex conjugate of the number $a$
$\dagger$	Conjugate Transpose operation
$ \cdot\rangle\langle\cdot $	Outer product operation
$A \odot B$	element-wise multiplication (Hadamard product) of matrices $A$ and $B$
$f \circ g$	composition of functions $f$ and $g$
$ 0\rangle,  1\rangle,  \psi\rangle$	Pure quantum state in Dirac notation
$N$	Number of qubits
$ 0^N\rangle,  0\rangle^{\otimes N}$	Zero state with $N$ -qubits
$\mathbb{I}_d$	Identity matrix with the size $d \times d$
$\rho$	Quantum state in density matrix representation
$H$	Hamiltonian

$U, V$	Unitary operator
$X, Y, Z$	Pauli operators
$\text{RX}, \text{RY}, \text{RZ}$	Rotation gates along the $x, y$ and $z$ axes, respectively
$CX, CZ$	Controlled-X gate and controlled-Z gate
$\mathcal{E}, \mathcal{N}$	Quantum channel
$O$	Observable
$\langle O \rangle$	Expectation of observable $O$
$\ \cdot\ _{op}$	Operator norm
<hr/>	
$n$	Number of training examples
$\mathcal{D}$	Dataset
$\text{Tr}(O\rho)$	Expectation value of an observable $O$
$U(\boldsymbol{\theta})$	Parameterized quantum circuit
$\mathcal{L}(\boldsymbol{\theta})$	Loss function
$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$	Gradient of loss function $\mathcal{L}$ w.r.t parameters $\boldsymbol{\theta}$

Table A.1: Notations used in this work.

## Appendix B

# Concentration Inequality

In this section, we introduce some of the most common concentration inequalities in statistical learning theory. These inequalities are widely used in deriving generalization error bounds for learning models. In practical scenarios, one often needs to infer properties of an unknown distribution based on finite data samples drawn from that distribution. Concentration inequalities address the deviations of functions of independent random variables from their expectations. They provide tools to analyze the difference between the empirical mean (or some estimate) and the true expectation of random variables that follow a probability distribution.

We begin by recalling some basic tools that will be used throughout this section. For any nonnegative random variable  $X$  following the probability distribution  $p(x)$ , its expectation can be written as

$$\mathbb{E}X = \int_0^\infty xp(x)dx. \quad (\text{B.1})$$

This leads directly to a fundamental building block for concentration inequalities, namely Markov's inequality.

**Lemma B.1** (Markov's inequality). *For any nonnegative random variable  $X$ , and a positive constant  $t > 0$ , we have*

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}X}{t}. \quad (\text{B.2})$$

*Proof.* Employing the definition of cumulative distribution function, we have

$$\begin{aligned}\mathbb{P}\{X \geq t\} &= \int_t^\infty p(x)dt \\ &\leq \int_t^\infty p(x)\frac{x}{t}dt \\ &\leq \frac{\int_0^\infty p(x)xdt}{t} \\ &\leq \frac{\mathbb{E}X}{t},\end{aligned}\tag{B.3}$$

where the first inequality follows that  $x/t > 1$  in the interval  $x \in [t, \infty]$ , and the second inequality employs the positiveness of integral term  $p(x)x/t$ .  $\square$

Using Markov's inequality, it follows that if  $\phi$  is a strictly monotonically increasing, nonnegative function, then for any random variable  $X$  and real number  $t > 0$ , we have

$$\mathbb{P}\{X \geq t\} = \mathbb{P}\{\phi(X) \geq \phi(t)\} \leq \frac{\mathbb{E}\phi(X)}{\phi(t)}.\tag{B.4}$$

An application of this result with  $\phi(x) = x^2$  leads to the simplest concentration inequality, i.e., Chebyshev's inequality.

**Lemma B.2.** *Let  $X$  be an arbitrary random variable and the real number  $t > 0$ , then*

$$\mathbb{P}\{|X - \mathbb{E}X| \geq t\} \leq \frac{\text{Var}(X)}{t^2}.\tag{B.5}$$

*Proof.* Utilizing the extention of Markov's inequality in Eqn. (B.4) with setting  $\phi(x) = x^2$  yields

$$\begin{aligned}\mathbb{P}\{|X - \mathbb{E}X| \geq t\} &= \mathbb{P}\{|X - \mathbb{E}X|^2 \geq t^2\} \\ &\leq \frac{\mathbb{E}(X - \mathbb{E}X)^2}{t^2} \\ &= \frac{\text{Var}(X)}{t^2}.\end{aligned}\tag{B.6}$$

$\square$

More generally, by taking  $\phi(x) = x^q$  ( $x \geq 0$ ), then for any  $q > 0$ , we obtain the following moment-based inequality

$$\mathbb{P}\{|X - \mathbb{E}X| \geq t\} \leq \frac{\mathbb{E}(X - \mathbb{E}X)^q}{t^q}.\tag{B.7}$$

Here, the parameter  $q$  can be chosen to optimize the upper bound in specific examples. Such moment bounds often provide sharp estimates for tail probabilities. A related idea forms the basis of Chernoff's bounding method. In particular, by setting  $\phi(x) = e^{sx}$  for some  $s > 0$ , we can derive a useful upper bound for any random variable  $X$  and  $t > 0$ ,

$$\mathbb{P}\{X \geq t\} = \mathbb{P}\{e^{sX} \geq e^{st}\} \leq \frac{\mathbb{E}e^{sX}}{e^{st}}. \quad (\text{B.8})$$

In Chernoff's method, the goal is to choose an appropriate  $s > 0$  to minimize the upper bound or make it as small as possible.

Now, we turn to concentration inequalities for sums of independent random variables. Specifically, we aim to bound probabilities of deviations from the mean, i.e.,  $\mathbb{P}\{|S_n - \mathbb{E}S_n| \geq t\}$ , where  $S_n = \sum_{i=1}^n X_i$ , and  $X_1, \dots, X_n$  are independent real-valued random variables.

By applying Chebyshev's inequality to  $S_n$ , we obtain

$$\mathbb{P}\{|S_n - \mathbb{E}S_n| \geq t\} \leq \frac{\text{Var}(S_n)}{t^2} = \frac{\sum_{i=1}^n \text{Var}(X_i)}{t^2}. \quad (\text{B.9})$$

In terms of the sample mean, this can be rewritten as,

$$\mathbb{P}\left\{\left|\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}X_i\right| \geq \varepsilon\right\} \leq \frac{\sigma^2}{n\varepsilon^2}, \quad (\text{B.10})$$

where  $\sigma^2 = \frac{1}{n} \sum_{i=1}^n \text{Var}(X_i)$ . Chernoff's bounding method is particularly useful for bounding tail probabilities of sums of independent random variables. By exploiting the independence property (i.e., the expected value of a product of independent random variables equals the product of their expected values), Chernoff's bound can be expressed as

$$\begin{aligned} \mathbb{P}\{S_n - \mathbb{E}S_n \geq t\} &\leq e^{-st} \mathbb{E}\left[\exp\left(s \sum_{i=1}^n (X_i - \mathbb{E}X_i)\right)\right] \\ &= e^{-st} \prod_{i=1}^n \mathbb{E}[\exp(s(X_i - \mathbb{E}X_i))] \quad (\text{by independence}). \end{aligned} \quad (\text{B.11})$$

Now, the challenge then becomes finding a good upper bound for the moment generating function of the random variables  $X_i - \mathbb{E}X_i$ . For bounded random variables, one of the most elegant results is Hoeffding's inequality [Hoeffding \(1994\)](#).

**Lemma B.3** (Hoeffding's inequality). *Let  $X$  be a random variable with  $\mathbb{E}X = 0, a \leq X \leq b$ . Then for  $s > 0$ ,*

$$\mathbb{E}[e^{sx}] \leq \exp\left(\frac{s^2(b-a)^2}{8}\right) \quad (\text{B.12})$$

This lemma, combined with Eqn. (B.11) immediately implies Hoeffding's tail inequality (Hoeffding, 1994).

**Theorem B.4.** *Let  $X_1, \dots, X_n$  be independent bounded random variables such that  $X_i$  falls in the interval  $[a_i, b_i]$  with probability one. Then for any real number  $t > 0$ , we have*

$$\mathbb{P}\{S_n - \mathbb{E}S_n \geq t\} \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \quad (\text{B.13})$$

and

$$\mathbb{P}\{S_n - \mathbb{E}S_n \leq -t\} \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \quad (\text{B.14})$$

Hoeffding's inequality, first proven for binomial random variables by Chernoff (1952) and Okamoto (1959), provides a powerful tool for bounding tail probabilities. However, a limitation is that it does not take into account the variance of the  $X_i$ 's, which can sometimes yield loose bounds.

## Appendix C

# Haar Measure and Unitary t-design

In this section, we introduce some basic knowledge of Haar measure (Haar, 1933) and unitary  $t$ -design (Dankert et al., 2009), which are extensively employed in group representation theory and quantum information (Nagy, 1993; Adam, 2013), especially in the analysis of barren plateaus and the trainability of variational quantum algorithms (Larocca et al., 2024).

We begin with the Haar measure. Roughly speaking, Haar measure is a unique probability measure that generates the uniform distribution over a compact group. In this chapter, we focus on the Haar measure on the unitary space  $\mathcal{U}(d)$  for convenience. Mathematically, the Haar measure is uniform given by invariant properties in Definition C.1.

**Definition C.1** (Haar measure on  $\mathcal{U}(d)$ ). *A measure  $\mu$  is the Haar measure on the unitary space  $\mathcal{U}(d)$  if and only if  $\mu$  is*

1. *left-invariant*, i.e.  $\mu(U\mathcal{S}) = \mu(\mathcal{S})$  for any measurable set  $\mathcal{S} \subseteq \mathcal{U}(d)$  and any unitary  $U \in \mathcal{U}(d)$ .
2. *right-invariant*, i.e.  $\mu(\mathcal{S}U) = \mu(\mathcal{S})$  for any measurable set  $\mathcal{S} \subseteq \mathcal{U}(d)$  and any unitary  $U \in \mathcal{U}(d)$ .
3. *a probability measure*, i.e.  $\int d\mu(U) = 1$ .

As provided in Definition C.1, the Haar measure is continuous on the whole space due to left- and right-invariant properties. Therefore, researchers are interested in approximating the Haar measure with the uniform distribution over some finite sets. Specifically, the unitary set whose uniform

distribution shares the same  $t$ -th moment with the Haar measure is defined as the unitary  $t$ -design.

**Definition C.2** (Unitary 2-design). *Let  $\mu$  be the Haar measure on the space  $\mathcal{U}(d)$ . Then, a finite set  $\mathcal{S}$  forms a unitary  $t$ -design if and only if it fulfills one of the following equivalent conditions:*

1.

$$\frac{1}{|\mathcal{S}|} \sum_{U \in \mathcal{S}} U^{\otimes t} \otimes (U^\dagger)^{\otimes t} = \int_{\mathcal{U}(d)} U^{\otimes t} \otimes (U^\dagger)^{\otimes t} d\mu(U).$$

2. Let  $P_{t,t}(U)$  be the polynomial with at most  $t$  degrees of elements from  $U$  and at most  $t$  degrees of elements from  $U^\dagger$ . Then

$$\frac{1}{|\mathcal{S}|} \sum_{U \in \mathcal{S}} P_{t,t}(U) = \int_{\mathcal{U}(d)} P_{t,t}(U) d\mu(U).$$

**Corollary C.3.** *A unitary  $t$ -design is also a unitary  $(t-1)$ -design.*

We remark that invariant properties of the Haar measure lead to several useful formulations of unitary  $t$ -designs provided in Facts C.4 and C.5.

**Fact C.4** (Average over unitary 1-design (Puchała and Miszczak, 2017)). *Let  $\mathcal{S}$  be a set of unitary 1-design on  $\mathcal{U}(d)$  and  $\mu$  be the corresponding Haar measure. Then*

$$\frac{1}{|\mathcal{S}|} \sum_{U \in \mathcal{S}} U_{ij} U_{i'j'}^* = \int_{\mathcal{U}(d)} U_{ij} U_{i'j'}^* d\mu(U) = \frac{1}{d} \delta_{ii'} \delta_{jj'}.$$

**Fact C.5** (Average over unitary 2-design (Puchała and Miszczak, 2017)). *Let  $\mathcal{S}$  be a set of unitary 2-design on  $\mathcal{U}(d)$  and  $\mu$  be the corresponding Haar measure. Then*

$$\begin{aligned} & \frac{1}{|\mathcal{S}|} \sum_{U \in \mathcal{S}} U_{i_1 j_1} U_{i_2 j_2} U_{i'_1 j'_1}^* U_{i'_2 j'_2}^* = \int_{\mathcal{U}(d)} U_{i_1 j_1} U_{i_2 j_2} U_{i'_1 j'_1}^* U_{i'_2 j'_2}^* d\mu(U) \\ &= \frac{1}{d^2 - 1} \left( \delta_{i_1 i'_1} \delta_{i_2 i'_2} \delta_{j_1 j'_1} \delta_{j_2 j'_2} + \delta_{i_1 i'_2} \delta_{i_2 i'_1} \delta_{j_1 j'_2} \delta_{j_2 j'_1} \right) \\ & \quad - \frac{1}{d(d^2 - 1)} \left( \delta_{i_1 i'_1} \delta_{i_2 i'_2} \delta_{j_1 j'_2} \delta_{j_2 j'_1} + \delta_{i_1 i'_2} \delta_{i_2 i'_1} \delta_{j_1 j'_1} \delta_{j_2 j'_2} \right). \end{aligned}$$

How big is a unitary  $t$ -design? Roy and Scott (2009) have proved that, for instance, the size of unitary 1-design and unitary 2-design scale polynomially to the dimension of the unitary space.

**Fact C.6** (The size of a unitary 2-design (Roy and Scott, 2009)). *A unitary 1-design on  $\mathcal{U}(d)$  has no fewer than  $d^2$  elements. A unitary 2-design on  $\mathcal{U}(d)$  has no fewer than  $d^4 - 2d^2 + 2$  elements.*

For a system with  $N$  qubits, the dimension of the unitary space is  $d = 2^N$ . Therefore, an exact  $t$ -design could involve exponential numbers of ensembles with increased qubits. Could we obtain an approximation to the unitary  $t$ -design, which can be generated in polynomial times with less degree of freedom? Haferkamp (2022) has proved that random quantum circuits with linear depths could form an approximate unitary  $t$ -design.

**Definition C.7** (Approximate unitary designs). *Let  $\mu$  be the Haar measure on the space  $\mathcal{U}(d)$ . We denote the moment superoperator  $\Phi_\nu^{(t)}(A) := \int_{\mathcal{U}(d)} U^{\otimes t} A (U^\dagger)^{\otimes t}$ . Denote by  $M_n(\mathcal{C})$  the  $n \times n$  complex matrices. Denote by  $\|\Phi\|_\diamond := \max_{X; \|X\|_1 \leq 1} \|(\Phi \otimes \mathcal{I}_n)X\|_1$  the diamond norm for the linear transformation  $\Phi : M_n(\mathcal{C}) \rightarrow M_m(\mathcal{C})$  and  $X \in M_{n^2}(\mathcal{C})$ . Then, a probability distribution  $\nu$  on  $\mathcal{U}(d)$  is an  $\epsilon$ -approximate unitary  $t$ -design if*

$$\left\| \Phi_\nu^{(t)} - \Phi_\mu^{(t)} \right\|_\diamond \leq \frac{\epsilon}{d^t}.$$

**Fact C.8** (Random quantum circuits form approximate unitary designs, informal version from Haferkamp (2022)). *For the number of qubits  $N \geq \mathcal{O}(\log t)$ , alternative layered random quantum circuits with Haar-random unitary gates sampled from  $\mathcal{U}(4)$  lead to an  $\epsilon$ -approximate unitary  $t$ -design when the circuit depth*

$$k \geq \mathcal{O}\left(t^{4+o(1)} \left(Nt + \log \frac{1}{\epsilon}\right)\right),$$

where the term  $o(1) \rightarrow 0$  when  $t \rightarrow \infty$ .



## Notes

<sup>1</sup>Parameters are obtained from the [Hugging Face](#) website, which is an open-source platform for machine learning models.



# Bibliography

- Richard P Feynman. Quantum mechanical computers. *Between Quantum and Cosmos*, pages 523–548, 2017.
- John Watrous. Quantum computational complexity. *arXiv preprint arXiv:0804.3401*, 2008.
- Peter W Shor. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM review*, 41(2):303–332, 1999.
- Craig Gidney and Martin Ekerå. How to factor 2048 bit rsa integers in 8 hours using 20 million noisy qubits. *Quantum*, 5:433, 2021.
- Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- Paul Benioff. Quantum mechanical hamiltonian models of turing machines. *Journal of Statistical Physics*, 29:515–546, 1982.
- Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, USA, 2 edition, 2011. ISBN 1107002176.
- Robert Raussendorf and Hans J Briegel. A one-way quantum computer. *Physical review letters*, 86(22):5188, 2001.
- Tameem Albash and Daniel A Lidar. Adiabatic quantum computation. *Reviews of Modern Physics*, 90(1):015002, 2018.

- A Yu Kitaev. Fault-tolerant quantum computation by anyons. *Annals of physics*, 303(1):2–30, 2003.
- Lov K Grover. A fast quantum mechanical algorithm for database search. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 212–219, 1996.
- Dylan Herman, Cody Gogin, Xiaoyuan Liu, Yue Sun, Alexey Galda, Ilya Safro, Marco Pistoia, and Yuri Alexeev. Quantum computing for finance. *Nature Reviews Physics*, 5(8):450–465, 2023.
- Raffaele Santagati, Alan Aspuru-Guzik, Ryan Babbush, Matthias Degroote, Leticia Gonzalez, Elica Kyoseva, Nikolaj Moll, Markus Oppel, Robert M Parrish, Nicholas C Rubin, et al. Drug design on quantum computers. *Nature Physics*, 20(4):549–557, 2024.
- Amira Abbas, Andris Ambainis, Brandon Augustino, Andreas Bärtschi, Harry Buhrman, Carleton Coffrin, Giorgio Cortiana, Vedran Dunjko, Daniel J Egger, Bruce G Elmegreen, et al. Challenges and opportunities in quantum optimization. *Nature Reviews Physics*, pages 1–18, 2024.
- Andrew W Cross, Lev S Bishop, Sarah Sheldon, Paul D Nation, and Jay M Gambetta. Validating quantum computers using randomized model circuits. *Physical Review A*, 100(3):032328, 2019.
- Andrew Wack, Hanhee Paik, Ali Javadi-Abhari, Petar Jurcevic, Ismael Faro, Jay M Gambetta, and Blake R Johnson. Quality, speed, and scale: three key attributes to measure the performance of near-term quantum computers. *arXiv preprint arXiv:2110.14108*, 2021.
- Kostyantyn Kechedzhi, Sergei V Isakov, Salvatore Mandrà, Benjamin Villalonga, Xiao Mi, Sergio Boixo, and Vadim Smelyanskiy. Effective quantum volume, fidelity and computational cost of noisy quantum processing experiments. *Future Generation Computer Systems*, 153:431–441, 2024.
- Srinivasan Arunachalam and Ronald De Wolf. Guest column: A survey of quantum learning theory. *ACM Sigact News*, 48(2):41–67, 2017.
- Anurag Anshu and Srinivasan Arunachalam. A survey on the complexity of learning quantum states. *Nature Reviews Physics*, 6(1):59–69, 2024.
- Aram W Harrow and Ashley Montanaro. Quantum computational supremacy. *Nature*, 549(7671):203–209, 2017.

- Youngseok Kim, Andrew Eddins, Sajant Anand, Ken Xuan Wei, Ewout Van Den Berg, Sami Rosenblatt, Hasan Nayfeh, Yantao Wu, Michael Zaletel, Kristan Temme, et al. Evidence for the utility of quantum computing before fault tolerance. *Nature*, 618(7965):500–505, 2023.
- Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. An introduction to quantum machine learning. *Contemporary Physics*, 56(2):172–185, 2015.
- Vedran Dunjko and Hans J Briegel. Machine learning & artificial intelligence in the quantum domain: a review of recent progress. *Reports on Progress in Physics*, 81(7):074001, 2018.
- Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):045002, 2019.
- Bin Cheng, Xiu-Hao Deng, Xiu Gu, Yu He, Guangchong Hu, Peihao Huang, Jun Li, Ben-Chuan Lin, Dawei Lu, Yao Lu, et al. Noisy intermediate-scale quantum computers. *Frontiers of Physics*, 18(2):21308, 2023.
- He-Liang Huang, Dachao Wu, Daojin Fan, and Xiaobo Zhu. Superconducting quantum computing: a review. *Science China Information Sciences*, 63:1–32, 2020a.
- Colin D Bruzewicz, John Chiaverini, Robert McConnell, and Jeremy M Sage. Trapped-ion quantum computing: Progress and challenges. *Applied Physics Reviews*, 6(2), 2019.
- M Morgado and S Whitlock. Quantum simulation and computing with rydberg-interacting qubits. *AVS Quantum Science*, 3(2), 2021.
- John Preskill. Quantum computing in the nisq era and beyond. *Quantum*, 2:79, 2018.
- Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando GSL Brandao, David A Buell, et al. Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510, 2019.
- Yulin Wu, Wan-Su Bao, Sirui Cao, Fusheng Chen, Ming-Cheng Chen, Xiewei Chen, Tung-Hsun Chung, Hui Deng, Yajie Du, Daojin Fan, et al. Strong quantum computational advantage using a superconducting quantum processor. *Physical review letters*, 127(18):180501, 2021.

- Aram W Harrow, Avinatan Hassidim, and Seth Lloyd. Quantum algorithm for linear systems of equations. *Physical review letters*, 103(15):150502, 2009.
- Ashley Montanaro. Quantum algorithms: an overview. *npj Quantum Information*, 2(1):1–8, 2016.
- Alexander M Dalzell, Sam McArdle, Mario Berta, Przemyslaw Bienias, Chi-Fang Chen, András Gilyén, Connor T Hann, Michael J Kastoryano, Emil T Khabiboulline, Aleksander Kubica, et al. Quantum algorithms: A survey of applications and end-to-end complexities. *arXiv preprint arXiv:2310.03011*, 2023.
- Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost. Quantum principal component analysis. *Nature Physics*, 10(9):631–633, September 2014. ISSN 1745-2481. doi: 10.1038/nphys3029.
- Patrick Rebentrost, Masoud Mohseni, and Seth Lloyd. Quantum support vector machine for big data classification. *Physical review letters*, 113(13):130503, 2014.
- András Gilyén, Yuan Su, Guang Hao Low, and Nathan Wiebe. Quantum singular value transformation and beyond: exponential improvements for quantum matrix arithmetics. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC ’19. ACM, June 2019. doi: 10.1145/3313276.3316366. URL <http://dx.doi.org/10.1145/3313276.3316366>.
- Zherui Chen, Yuchen Lu, Hao Wang, Yizhou Liu, and Tongyang Li. Quantum langevin dynamics for optimization. *arXiv preprint arXiv:2311.15587*, 2023.
- Junyu Liu, Minzhao Liu, Jin-Peng Liu, Ziyu Ye, Yunfei Wang, Yuri Alexeev, Jens Eisert, and Liang Jiang. Towards provably efficient quantum algorithms for large-scale machine-learning models. *Nature Communications*, 15(1):434, 2024a.
- Scott Aaronson. Read the fine print. *Nature Physics*, 11(4):291–293, 2015.
- Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. Quantum random access memory. *Physical review letters*, 100(16):160501, 2008.

- Vojtěch Havlíček, Antonio D Córcoles, Kristan Temme, Aram W Harrow, Abhinav Kandala, Jerry M Chow, and Jay M Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212, 2019.
- Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, 2021a.
- Weikang Li and Dong-Ling Deng. Recent advances for quantum classifiers. *Science China Physics, Mechanics & Astronomy*, 65(2):220301, 2022.
- Jinkai Tian, Xiaoyu Sun, Yuxuan Du, Shanshan Zhao, Qing Liu, Kaining Zhang, Wei Yi, Wanrong Huang, Chaoyue Wang, Xingyao Wu, et al. Recent advances for quantum neural networks in generative learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12321–12340, 2023.
- Kishor Bharti, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermanni Heimonen, Jakob S Kottmann, Tim Menke, et al. Noisy intermediate-scale quantum algorithms. *Reviews of Modern Physics*, 94(1):015004, 2022.
- Marco Cerezo, Guillaume Verdon, Hsin-Yuan Huang, Lukasz Cincio, and Patrick J Coles. Challenges and opportunities in quantum machine learning. *Nature Computational Science*, 2(9):567–576, 2022.
- Yuxuan Du, Tao Huang, Shan You, Min-Hsiu Hsieh, and Dacheng Tao. Quantum circuit architecture search for variational quantum algorithms. *npj Quantum Information*, 8(1):62, 2022a.
- James Stokes, Josh Izaac, Nathan Killoran, and Giuseppe Carleo. Quantum natural gradient. *Quantum*, 4:269, 2020.
- Sukin Sim, Jonathan Romero, Jérôme F Gonthier, and Alexander A Kunitsa. Adaptive pruning-based optimization of parameterized quantum circuits. *Quantum Science and Technology*, 6(2):025019, 2021.
- Xinbiao Wang, Junyu Liu, Tongliang Liu, Yong Luo, Yuxuan Du, and Dacheng Tao. Symmetric pruning in quantum neural networks. In *The Eleventh International Conference on Learning Representations*, 2023a. URL <https://openreview.net/forum?id=K96AogLDT2K>.

- Sirui Lu, Lu-Ming Duan, and Dong-Ling Deng. Quantum adversarial machine learning. *Phys. Rev. Res.*, 2:033212, Aug 2020. doi: 10.1103/PhysRevResearch.2.033212. URL <https://link.aps.org/doi/10.1103/PhysRevResearch.2.033212>.
- Wenjie Jiang, Zhide Lu, and Dong-Ling Deng. Quantum continual learning overcoming catastrophic forgetting. *Chinese Physics Letters*, 39(5):050303, may 2022. doi: 10.1088/0256-307X/39/5/050303. URL <https://dx.doi.org/10.1088/0256-307X/39/5/050303>.
- Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, Dacheng Tao, and Nana Liu. Quantum noise protects quantum classifiers against adversaries. *Physical Review Research*, 3(2):023153, 2021a.
- William M. Watkins, Samuel Yen-Chi Chen, and Shinjae Yoo. Quantum machine learning with differential privacy. *Scientific Reports*, 13(1):2453, 2023. doi: 10.1038/s41598-022-24082-z. URL <https://doi.org/10.1038/s41598-022-24082-z>.
- Yuxuan Du, Yang Qian, Xingyao Wu, and Dacheng Tao. A distributed learning scheme for variational quantum algorithms. *IEEE Transactions on Quantum Engineering*, 3:1–16, 2022b.
- Chao Ren, Rudai Yan, Huihui Zhu, Han Yu, Minrui Xu, Yuan Shen, Yan Xu, Ming Xiao, Zhao Yang Dong, Mikael Skoglund, et al. Towards quantum federated learning. *arXiv preprint arXiv:2306.09912*, 2023.
- Lirandë Pira and Chris Ferrie. On the interpretability of quantum neural networks. *Quantum Machine Intelligence*, 6(2):52, 2024.
- Leonardo Banchi, Jason Luke Pereira, Sharu Theresa Jose, and Osvaldo Simeone. Statistical complexity of quantum learning. *Advanced Quantum Technologies*, page 2300311, 2023.
- Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, Shah-nawaz Ahmed, Vishnu Ajith, M Sohaib Alam, Guillermo Alonso-Linaje, B AkashNarayanan, Ali Asadi, et al. Pennylane: Automatic differentiation of hybrid quantum-classical computations. *arXiv preprint arXiv:1811.04968*, 2018.
- Ali Javadi-Abhari, Matthew Treinish, Kevin Krsulich, Christopher J Wood, Jake Lishman, Julien Gacon, Simon Martiel, Paul D Nation, Lev S Bishop, Andrew W Cross, et al. Quantum computing with qiskit. *arXiv preprint arXiv:2405.08810*, 2024.

Cirq Developers. Cirq, May 2024. URL <https://doi.org/10.5281/zenodo.11398048>.

Michael Broughton, Guillaume Verdon, Trevor McCourt, Antonio J Martinez, Jae Hyeon Yoo, Sergei V Isakov, Philip Massey, Ramin Halavati, Murphy Yuezhen Niu, Alexander Zlokapa, et al. Tensorflow quantum: A software framework for quantum machine learning. *arXiv preprint arXiv:2003.02989*, 2020.

Richard Jozsa and Noah Linden. On the role of entanglement in quantum-computational speed-up. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 459(2036): 2011–2032, 2003.

Daniel M Greenberger, Michael A Horne, and Anton Zeilinger. Going beyond bell’s theorem. In *Bell’s theorem, quantum theory and conceptions of the universe*, pages 69–72. Springer, 1989.

Donald P Leach and Albert P Malvino. *Digital principles and Applications*. Glencoe/McGraw-Hill, 1994.

Christopher M Dawson and Michael A Nielsen. The solovay-kitaev algorithm. *arXiv preprint quant-ph/0505030*, 2005.

Mark M Wilde. From classical to quantum shannon theory. *arXiv preprint arXiv:1106.1445*, 2011.

Steven T Flammia and Joel J Wallman. Efficient estimation of pauli channels. *ACM Transactions on Quantum Computing*, 1(1):1–32, 2020.

John Preskill. Lecture notes for physics 219: Quantum computation. *Caltech Lecture Notes*, 7:1, 1999.

K Vogel and H Risken. Determination of quasiprobability distributions in terms of probability distributions for the rotated quadrature phase. *Physical Review A*, 40(5):2847, 1989.

Bo Qi, Zhibo Hou, Li Li, Daoyi Dong, Guoyong Xiang, and Guangcan Guo. Quantum state tomography via linear regression estimation. *Scientific reports*, 3(1):3496, 2013.

Zdenek Hradil. Quantum-state estimation. *Physical Review A*, 55(3):R1561, 1997.

- David Layden, Guglielmo Mazzola, Ryan V Mishmash, Mario Motta, Paweł Wocjan, Jin-Sung Kim, and Sarah Sheldon. Quantum-enhanced markov chain monte carlo. *Nature*, 619(7969):282–287, 2023.
- Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm. *arXiv preprint arXiv:1411.4028*, 2014.
- Sergio Boixo, Sergei V Isakov, Vadim N Smelyanskiy, Ryan Babbush, Nan Ding, Zhang Jiang, Michael J Bremner, John M Martinis, and Hartmut Neven. Characterizing quantum supremacy in near-term devices. *Nature Physics*, 14(6):595–600, 2018.
- Adam Bouland, Bill Fefferman, Chinmay Nirkhe, and Umesh Vazirani. On the complexity and verification of quantum random circuit sampling. *Nature Physics*, 15(2):159–163, 2019.
- Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M Chow, and Jay M Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *nature*, 549(7671):242–246, 2017a.
- Jules Tilly, Hongxiang Chen, Shuxiang Cao, Dario Picozzi, Kanav Setia, Ying Li, Edward Grant, Leonard Wossnig, Ivan Rungger, George H Booth, et al. The variational quantum eigensolver: a review of methods and best practices. *Physics Reports*, 986:1–128, 2022.
- Scott Aaronson. Shadow tomography of quantum states. In *Proceedings of the 50th annual ACM SIGACT symposium on theory of computing*, pages 325–338, 2018.
- Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, 2020b.
- Naixu Guo, Zhan Yu, Matthew Choi, Aman Agrawal, Kouhei Nakaji, Alán Aspuru-Guzik, and Patrick Rebentrost. Quantum linear algebra is all you need for transformer architectures, 2024a. URL <https://arxiv.org/abs/2402.16714>.
- Gilles Brassard, Peter Høyer, Michele Mosca, and Alain Tapp. Quantum amplitude amplification and estimation. In *Quantum Computation and Information (Washington, DC, 2000)*, volume 305 of *Contemporary Mathematics*, pages 53–74. American Mathematical Society, Providence, RI, 2002. ISBN 978-0-8218-2140-4. doi: 10.1090/conm/305/05215.

- Liming Zhao, Zhikuan Zhao, Patrick Rebentrost, and Joseph Fitzsimons. Compiling basic linear algebra subroutines for quantum computers. *Quantum Machine Intelligence*, 3(2):21, June 2021. ISSN 2524-4914. doi: 10.1007/s42484-021-00048-8. URL <https://doi.org/10.1007/s42484-021-00048-8>.
- Naixu Guo, Kosuke Mitarai, and Keisuke Fujii. Nonlinear transformation of complex amplitudes via quantum singular value transformation. *Phys. Rev. Res.*, 6:043227, Dec 2024b. doi: 10.1103/PhysRevResearch.6.043227. URL <https://link.aps.org/doi/10.1103/PhysRevResearch.6.043227>.
- Arthur G. Rattew and Patrick Rebentrost. Non-linear transformations of quantum amplitudes: Exponential improvement, generalization, and applications, 2023. URL <https://arxiv.org/abs/2309.09839>.
- Kouhei Nakaji, Shunpei Uno, Yohichi Suzuki, Rudy Raymond, Tamiya Onodera, Tomoki Tanaka, Hiroyuki Tezuka, Naoki Mitsuda, and Naoki Yamamoto. Approximate amplitude encoding in shallow parameterized quantum circuits and its application to financial market indicators. *Physical Review Research*, 4(2):023136, 2022.
- Adrián Pérez-Salinas, Alba Cervera-Lierta, Elies Gil-Fuster, and José I Latorre. Data re-uploading for a universal quantum classifier. *Quantum*, 4:226, 2020.
- Tak Hur, Israel F Araujo, and Daniel K Park. Neural quantum embedding: Pushing the limits of quantum supervised learning. *Physical Review A*, 110(2):022411, 2024.
- Maria Schuld, Francesco Petruccione, Maria Schuld, and Francesco Petruccione. Information encoding. *Supervised Learning with Quantum Computers*, pages 139–171, 2018.
- Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.
- Seth Lloyd, Maria Schuld, Aroosa Ijaz, Josh Izaac, and Nathan Killoran. Quantum embeddings for machine learning. *arXiv preprint arXiv:2001.03622*, 2020.
- Minzhao Liu, Junyu Liu, Rui Liu, Henry Makhanov, Danylo Lykov, Anuj Apte, and Yuri Alexeev. Embedding learning in hybrid quantum-classical

- neural networks. In *2022 IEEE International Conference on Quantum Computing and Engineering (QCE)*, pages 79–86. IEEE, 2022.
- Patrick Rebentrost, Adrian Steffens, Iman Marvian, and Seth Lloyd. Quantum singular-value decomposition of nonsparse low-rank matrices. *Physical review A*, 97(1):012327, 2018a.
- Kaining Zhang, Min-Hsiu Hsieh, Liu Liu, and Dacheng Tao. Quantum gram-schmidt processes and their application to efficient state readout for quantum algorithms. *Physical Review Research*, 3(4):043095, 2021a.
- Ben P Lanyon, Christine Maier, Milan Holzäpfel, Tillmann Baumgratz, Cornelius Hempel, Petar Jurcevic, Ish Dhand, AS Buyskikh, Andrew J Daley, Marcus Cramer, et al. Efficient tomography of a quantum many-body system. *Nature Physics*, 13(12):1158–1162, 2017.
- Román Orús. Tensor networks for complex quantum systems. *Nature Reviews Physics*, 1(9):538–550, 2019.
- Asja Fischer and Christian Igel. An introduction to restricted boltzmann machines. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 17th Iberoamerican Congress, CIARP 2012, Buenos Aires, Argentina, September 3-6, 2012. Proceedings 17*, pages 14–36. Springer, 2012.
- A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- Giacomo Torlai, Guglielmo Mazzola, Juan Carrasquilla, Matthias Troyer, Roger Melko, and Giuseppe Carleo. Neural-network quantum state tomography. *Nature Physics*, 14(5):447–450, 2018. doi: 10.1038/s41567-018-0048-5. URL <https://doi.org/10.1038/s41567-018-0048-5>.
- Tobias Schmale, Moritz Reh, and Martin Gärttner. Efficient quantum state tomography with convolutional neural networks. *npj Quantum Information*, 8(1), September 2022. ISSN 2056-6387. doi: 10.1038/s41534-022-00621-4. URL <http://dx.doi.org/10.1038/s41534-022-00621-4>.
- Haoxiang Wang, Maurice Weber, Josh Izaac, and Cedric Yen-Yu Lin. Predicting properties of quantum systems with conditional generative models. *arXiv preprint arXiv:2211.16943*, 2022a.

Liming Zhao, Naixu Guo, Ming-Xing Luo, and Patrick Rebentrost. Provable learning of quantum states with graphical models, 2023. URL <https://arxiv.org/abs/2309.09235>.

Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M. Chow, and Jay M. Gambetta. Hardware-efficient Variational Quantum Eigensolver for Small Molecules and Quantum Magnets. *Nature*, 549(7671):242–246, September 2017b. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature23879.

Vladyslav Verteletskyi, Tzu-Ching Yen, and Artur F Izmaylov. Measurement optimization in the variational quantum eigensolver using a minimum clique cover. *The Journal of chemical physics*, 152(12), 2020.

Nicholas C Rubin, Ryan Babbush, and Jarrod McClean. Application of fermionic marginal constraints to hybrid quantum algorithms. *New Journal of Physics*, 20(5):053020, 2018.

Andrew Arrasmith, Lukasz Cincio, Rolando D Somma, and Patrick J Coles. Operator sampling for shot-frugal optimization in variational algorithms. *arXiv preprint arXiv:2004.06252*, 2020.

Yang Qian, Yuxuan Du, and Dacheng Tao. Shuffle-qudio: accelerate distributed vqe with trainability enhancement and measurement reduction. *Quantum Machine Intelligence*, 6(1):1–22, 2024.

A. Yu. Kitaev. Quantum measurements and the abelian stabilizer problem, 1995. URL <https://arxiv.org/abs/quant-ph/9511026>.

Lidia Ruiz-Perez and Juan Carlos Garcia-Escartin. Quantum arithmetic with the quantum fourier transform. *Quantum Information Processing*, 16(6), April 2017. ISSN 1573-1332. doi: 10.1007/s11128-017-1603-1. URL <http://dx.doi.org/10.1007/s11128-017-1603-1>.

Jin-Peng Liu, Herman Øie Kolden, Hari K. Krovi, Nuno F. Loureiro, Konstantina Trivisa, and Andrew M. Childs. Efficient quantum algorithm for dissipative nonlinear differential equations. *Proceedings of the National Academy of Sciences*, 118(35), August 2021a. ISSN 1091-6490. doi: 10.1073/pnas.2026805118. URL <http://dx.doi.org/10.1073/pnas.2026805118>.

Andrew M. Childs, Jim-Peng Liu, and Aaron Ostrander. High-precision quantum algorithms for partial differential equations. *Quantum*, 5:574,

- November 2021. ISSN 2521-327X. doi: 10.22331/q-2021-11-10-574. URL <https://doi.org/10.22331/q-2021-11-10-574>.
- Dong An, Noah Linden, Jin-Peng Liu, Ashley Montanaro, Changpeng Shao, and Jiasu Wang. Quantum-accelerated multilevel Monte Carlo methods for stochastic differential equations in mathematical finance. *Quantum*, 5: 481, June 2021. ISSN 2521-327X. doi: 10.22331/q-2021-06-24-481. URL <https://doi.org/10.22331/q-2021-06-24-481>.
- Shi Jin, Nana Liu, and Yue Yu. Quantum simulation of partial differential equations via schrodingerisation, 2022. URL <https://arxiv.org/abs/2212.13969>.
- Zhong-Xia Shang, Naixu Guo, Dong An, and Qi Zhao. Design nearly optimal quantum algorithm for linear differential equations via lindbladians, 2024. URL <https://arxiv.org/abs/2410.19628>.
- Guang Hao Low and Yuan Su. Quantum eigenvalue processing, 2024. URL <https://arxiv.org/abs/2401.06240>.
- Dong An, Jin-Peng Liu, and Lin Lin. Linear combination of hamiltonian simulation for nonunitary dynamics with optimal state preparation cost. *Phys. Rev. Lett.*, 131:150603, Oct 2023. doi: 10.1103/PhysRevLett.131.150603. URL <https://link.aps.org/doi/10.1103/PhysRevLett.131.150603>.
- Dong An, Andrew M. Childs, Lin Lin, and Lexing Ying. Laplace transform based quantum eigenvalue transformation via linear combination of hamiltonian simulation, 2024. URL <https://arxiv.org/abs/2411.04010>.
- Youle Wang, Lei Zhang, Zhan Yu, and Xin Wang. Quantum phase processing and its applications in estimating phase and entropies. *Physical Review A*, 108(6), December 2023b. ISSN 2469-9934. doi: 10.1103/physreva.108.062413. URL <http://dx.doi.org/10.1103/PhysRevA.108.062413>.
- Danial Motlagh and Nathan Wiebe. Generalized quantum signal processing, 2024. URL <https://arxiv.org/abs/2308.01501>.
- Zane M. Rossi and Isaac L. Chuang. Multivariable quantum signal processing (m-qsp): prophecies of the two-headed oracle. *Quantum*, 6:811, September 2022. ISSN 2521-327X. doi: 10.22331/q-2022-09-20-811. URL <http://dx.doi.org/10.22331/q-2022-09-20-811>.

- Hitomi Mori, Kosuke Mitarai, and Keisuke Fujii. Efficient state preparation for multivariate monte carlo simulation, 2024. URL <https://arxiv.org/abs/2409.07336>.
- Xiao-Ming Zhang, Tongyang Li, and Xiao Yuan. Quantum State Preparation with Optimal Circuit Depth: Implementations and Applications. *Physical Review Letters*, 129(23):230504, November 2022a. doi: 10.1103/PhysRevLett.129.230504. URL <https://link.aps.org/doi/10.1103/PhysRevLett.129.230504>.
- Xiaoming Sun, Guojing Tian, Shuai Yang, Pei Yuan, and Shengyu Zhang. Asymptotically Optimal Circuit Depth for Quantum State Preparation and General Unitary Synthesis. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 42(10):3301–3314, October 2023. ISSN 1937-4151. doi: 10.1109/TCAD.2023.3244885. URL <https://ieeexplore.ieee.org/document/10044235>.
- Andrew M. Childs and Nathan Wiebe. Hamiltonian simulation using linear combinations of unitary operations. *Quantum Info. Comput.*, 12(11–12):901–924, November 2012. ISSN 1533-7146.
- Nikita Guseynov and Nana Liu. Efficient explicit circuit for quantum state preparation of piece-wise continuous functions, 2024. URL <https://arxiv.org/abs/2411.01131>.
- Nikita Guseynov, Xiajie Huang, and Nana Liu. Explicit gate construction of block-encoding for hamiltonians needed for simulating partial differential equations, 2024. URL <https://arxiv.org/abs/2405.12855>.
- Daan Camps, Lin Lin, Roel Van Beeumen, and Chao Yang. Explicit quantum circuits for block encodings of certain sparse matrices, 2023. URL <https://arxiv.org/abs/2203.10236>.
- Mehryar Mohri. Foundations of machine learning, 2018.
- Takeru Kusumoto, Kosuke Mitarai, Keisuke Fujii, Masahiro Kitagawa, and Makoto Negoro. Experimental quantum kernel trick with nuclear spins in a solid. *npj Quantum Information*, 7(1):94, 2021.
- Carsten Blank, Daniel K Park, June-Koo Kevin Rhee, and Francesco Petruccione. Quantum classifier with tailored quantum kernel. *npj Quantum Information*, 6(1):41, 2020.

- Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme. A rigorous and robust quantum speed-up in supervised machine learning. *Nature Physics*, 17(9):1013–1017, 2021b.
- Elies Gil-Fuster, Jens Eisert, and Vedran Dunjko. On the expressivity of embedding quantum kernels. *Machine Learning: Science and Technology*, 5(2):025003, 2024.
- Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R McClean. Power of data in quantum machine learning. *Nature communications*, 12(1):2631, 2021a.
- Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- Maria Schuld, Mark Fingerhuth, and Francesco Petruccione. Implementing a distance-based classifier with a quantum interference circuit. *Europhysics Letters*, 119(6):60002, 2017.
- Maria Schuld and Nathan Killoran. Quantum machine learning in feature hilbert spaces. *Physical review letters*, 122(4):040504, 2019.
- Thomas Hubregtsen, David Wierichs, Elies Gil-Fuster, Peter-Jan HS Derkx, Paul K Faehrmann, and Johannes Jakob Meyer. Training quantum embedding kernels on near-term quantum computers. *Physical Review A*, 106(4):042431, 2022.
- Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103(3):032430, 2021.
- Seyed Shakib Vedaie, Moslem Noori, Jaspreet S Oberoi, Barry C Sanders, and Ehsan Zahedinejad. Quantum multiple kernel learning. *arXiv preprint arXiv:2011.09694*, 2020.
- Cong Lei, Yuxuan Du, Peng Mi, Jun Yu, and Tongliang Liu. Neural auto-designer for enhanced quantum kernels. *arXiv preprint arXiv:2401.11098*, 2024.
- Jennifer R Glick, Tanvi P Gujarati, Antonio D Corcoles, Youngseok Kim, Abhinav Kandala, Jay M Gambetta, and Kristan Temme. Covariant quantum kernels for data with group structure. *Nature Physics*, 20(3):479–483, 2024.

- Supanut Thanasilp, Samson Wang, Marco Cerezo, and Zoë Holmes. Exponential concentration and untrainability in quantum kernel methods. *arXiv preprint arXiv:2208.11060*, 2022.
- Yudai Suzuki, Hideaki Kawaguchi, and Naoki Yamamoto. Quantum fisher kernel for mitigating the vanishing similarity issue. *Quantum Science and Technology*, 2022.
- Ruslan Shaydulin and Stefan M Wild. Importance of kernel bandwidth in quantum machine learning. *Physical Review A*, 106(4):042407, 2022.
- Abdulkadir Canatar, Evan Peters, Cengiz Pehlevan, Stefan M Wild, and Ruslan Shaydulin. Bandwidth enables generalization in quantum kernel models. *arXiv preprint arXiv:2206.06686*, 2022.
- Maria Schuld. Supervised quantum machine learning models are kernel methods. *arXiv preprint arXiv:2101.11020*, 2021.
- Sofiene Jerbi, Lukas J Fiderer, Hendrik Poulsen Nautrup, Jonas M Kübler, Hans J Briegel, and Vedran Dunjko. Quantum machine learning beyond kernel methods. *Nature Communications*, 14(1):1–8, 2023.
- Xinbiao Wang, Yuxuan Du, Yong Luo, and Dacheng Tao. Towards understanding the power of quantum kernels in the nisq era. *Quantum*, 5:531, 2021a.
- Jonas Kübler, Simon Buchholz, and Bernhard Schölkopf. The inductive bias of quantum kernels. *Advances in Neural Information Processing Systems*, 34:12661–12673, 2021.
- Ryan Sweke, Jean-Pierre Seifert, Dominik Hangleiter, and Jens Eisert. On the quantum versus classical learnability of discrete distributions. *Quantum*, 5:417, 2021.
- Daniel Beaulieu, Dylan Miracle, Anh Pham, and William Scherr. Quantum kernel for image classification of real world manufacturing defects. *arXiv preprint arXiv:2212.08693*, 2022.
- Pablo Rodriguez-Grasa, Robert Farzan-Rodriguez, Gabriele Novelli, Yue Ban, and Mikel Sanz. Satellite image classification with neural quantum kernels. *arXiv preprint arXiv:2409.20356*, 2024.
- Teresa Sancho-Lorente, Juan Román-Roche, and David Zueco. Quantum kernels to learn the phases of quantum matter. *Physical Review A*, 105(4):042432, 2022.

- Yusen Wu, Bujiao Wu, Jingbo Wang, and Xiao Yuan. Quantum phase recognition via quantum kernel methods. *Quantum*, 7:981, 2023.
- Kushal Batra, Kimberley M Zorn, Daniel H Foil, Eni Minerali, Victor O Gawriljuk, Thomas R Lane, and Sean Ekins. Quantum machine learning algorithms for drug discovery applications. *Journal of chemical information and modeling*, 61(6):2641–2647, 2021.
- Nana Liu and Patrick Rebentrost. Quantum machine learning for quantum anomaly detection. *Physical Review A*, 97(4):042315, 2018.
- Michele Grossi, Noelle Ibrahim, Voica Radescu, Robert Loredo, Kirsten Voigt, Constantin Von Altrock, and Andreas Rudnik. Mixed quantum-classical method for fraud detection with quantum feature selection. *IEEE Transactions on Quantum Engineering*, 3:1–12, 2022.
- Shungo Miyabe, Brian Quanz, Noriaki Shimada, Abhijit Mitra, Takahiro Yamamoto, Vladimir Rastunkov, Dimitris Alevras, Mekena Metcalf, Daniel JM King, Mohammad Mamouei, et al. Quantum multiple kernel learning in financial classification tasks. *arXiv preprint arXiv:2312.00260*, 2023.
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- Athanasis Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018(1):7068349, 2018.
- Daniel W Otter, Julian R Medina, and Jugal K Kalita. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624, 2020.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 30016–30030, 2022.
- Alex de Vries. The growing energy footprint of artificial intelligence. *Joule*, 7(10):2191–2194, 2023.

- SK Jeswal and S Chakraverty. Recent developments and applications in quantum neural network: A review. *Archives of Computational Methods in Engineering*, 26(4):793–807, 2019.
- Hsin-Yuan Huang, Michael Broughton, Jordan Cotler, Sitan Chen, Jerry Li, Masoud Mohseni, Hartmut Neven, Ryan Babbush, Richard Kueng, John Preskill, et al. Quantum advantage in learning from experiments. *Science*, 376(6598):1182–1186, 2022.
- Evan Peters, João Caldeira, Alan Ho, Stefan Leichenauer, Masoud Mohseni, Hartmut Neven, Panagiotis Spentzouris, Doug Strain, and Gabriel N Perdue. Machine learning of high dimensional data on a noisy quantum processor. *npj Quantum Information*, 7(1):161, 2021.
- Rajeev Acharya, Laleh Aghababaie-Beni, Igor Aleiner, Trond I Andersen, Markus Ansmann, Frank Arute, Kunal Arya, Abraham Asfaw, Nikita Astrakhantsev, Juan Atalaya, et al. Quantum error correction below the surface code threshold. *arXiv preprint arXiv:2408.13687*, 2024.
- Bartłomiej Gardas, Marek M Rams, and Jacek Dziarmaga. Quantum neural networks to simulate many-body quantum systems. *Physical Review B*, 98(18):184304, 2018.
- Yudong Cao, Jonathan Romero, Jonathan P Olson, Matthias Degroote, Peter D Johnson, Mária Kieferová, Ian D Kivlichan, Tim Menke, Borja Peropadre, Nicolas PD Sawaya, et al. Quantum chemistry in the age of quantum computing. *Chemical reviews*, 119(19):10856–10915, 2019.
- Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL <http://archive.ics.uci.edu/ml>.
- Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5: 115–133, 1943.
- Matt W Gardner and SR Dorling. Artificial neural networks (the multi-layer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636, 1998.

- Kurt Hornik. Some new results on neural network approximation. *Neural networks*, 6(8):1069–1072, 1993.
- Shun-ichi Amari. Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4-5):185–196, 1993.
- Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Albert BJ Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, volume 12, pages 615–622. New York, NY, 1962.
- Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.
- Yann LeCun, D Touresky, G Hinton, and T Sejnowski. A theoretical framework for back-propagation. In *Proceedings of the 1988 connectionist models summer school*, volume 1, pages 21–28, 1988.
- Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feed-forward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.
- Rich Caruana, Steve Lawrence, and C Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. *Advances in neural information processing systems*, 13, 2000.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- Anders Krogh and John Hertz. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991.

- Ashish Kapoor, Nathan Wiebe, and Krysta Svore. Quantum perceptron models. *Advances in neural information processing systems*, 29, 2016.
- Muhammad AbuGhanem. Ibm quantum computers: Evolution, performance, and future directions. *arXiv preprint arXiv:2410.00916*, 2024.
- Dongxin Gao, Daojin Fan, Chen Zha, Jiahao Bei, Guoqing Cai, Jianbin Cai, Sirui Cao, Xiangdong Zeng, Fusheng Chen, Jiang Chen, et al. Establishing a new benchmark in quantum computational advantage with 105-qubit zuchongzhi 3.0 processor. *arXiv preprint arXiv:2412.11924*, 2024.
- Yunfei Wang and Junyu Liu. A comprehensive review of quantum machine learning: from nisq to fault tolerance. *Reports on Progress in Physics*, 2024.
- Iris Cong, Soonwon Choi, and Mikhail D Lukin. Quantum convolutional neural networks. *Nature Physics*, 15(12):1273–1278, 2019.
- Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J Love, Alán Aspuru-Guzik, and Jeremy L O’brien. A variational eigenvalue solver on a photonic quantum processor. *Nature communications*, 5(1):4213, 2014.
- Gavin E Crooks. Gradients of parameterized quantum gates using the parameter-shift rule and gate decomposition. *arXiv preprint arXiv:1905.13311*, 2019.
- John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.
- Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Seth Lloyd and Christian Weedbrook. Quantum generative adversarial learning. *Physical review letters*, 121(4):040502, 2018.
- Sergey Bravyi, David Gosset, and Robert König. Quantum advantage with shallow circuits. *Science*, 362(6412):308–311, 2018.
- Elton Yechao Zhu, Sonika Johri, Dave Bacon, Mert Esencan, Jungsang Kim, Mark Muir, Nikhil Murgai, Jason Nguyen, Neal Pisenti, Adam Schouela, et al. Generative quantum learning of joint probability distribution functions. *Physical Review Research*, 4(4):043092, 2022.

- He-Liang Huang, Yuxuan Du, Ming Gong, Youwei Zhao, Yulin Wu, Chaoyue Wang, Shaowei Li, Futian Liang, Jin Lin, Yu Xu, et al. Experimental quantum generative adversarial networks for image generation. *Physical Review Applied*, 16(2):024051, 2021b.
- Marcello Benedetti, Delfina Garcia-Pintos, Oscar Perdomo, Vicente Leyton-Ortega, Yunseong Nam, and Alejandro Perdomo-Ortiz. A generative modeling approach for benchmarking and training shallow quantum circuits. *npj Quantum Information*, 5(1):45, 2019a.
- Yuxuan Du, Zhuozhuo Tu, Xiao Yuan, and Dacheng Tao. Efficient measure for the expressivity of variational quantum algorithms. *Physical Review Letters*, 128(8):080506, 2022c.
- Thomas Barthel and Jianfeng Lu. Fundamental limitations for measurements in quantum many-body systems. *Physical Review Letters*, 121(8):080406, 2018.
- Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Advances in neural information processing systems*, 21, 2008.
- Richard M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.
- ABAED Haussler and M Warmuth. Occam’s razor. *Information Processing Letters*, 24:377–380, 1987.
- Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J Coles, and Marco Cerezo. Theory of overparametrization in quantum neural networks. *Nature Computational Science*, 3(6):542–551, 2023.
- Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature communications*, 9(1):4812, 2018.
- Christoph Dankert, Richard Cleve, Joseph Emerson, and Etera Livine. Exact and approximate unitary 2-designs and their application to fidelity estimation. *Physical Review A—Atomic, Molecular, and Optical Physics*, 80(1):012304, 2009.

- Aram W Harrow and Richard A Low. Random quantum circuits are approximate 2-designs. *Communications in Mathematical Physics*, 291:257–302, 2009.
- Jonas Haferkamp. Random quantum circuits are approximate unitary  $t$ -designs in depth  $o(nt^{5+o(1)})$ . *Quantum*, 6:795, 2022.
- Marco Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J Coles. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature communications*, 12(1):1791, 2021b.
- Arthur Pesah, Marco Cerezo, Samson Wang, Tyler Volkoff, Andrew T Sornborger, and Patrick J Coles. Absence of barren plateaus in quantum convolutional neural networks. *Physical Review X*, 11(4):041011, 2021.
- Kaining Zhang, Min-Hsiu Hsieh, Liu Liu, and Dacheng Tao. Toward trainability of deep quantum neural networks. *arXiv preprint arXiv:2112.15002*, 2021b.
- Edward Grant, Leonard Wossnig, Mateusz Ostaszewski, and Marcello Benedetti. An initialization strategy for addressing barren plateaus in parametrized quantum circuits. *Quantum*, 3:214, 2019.
- Kaining Zhang, Liu Liu, Min-Hsiu Hsieh, and Dacheng Tao. Escaping from the barren plateau via gaussian initializations in deep variational quantum circuits. *Advances in Neural Information Processing Systems*, 35:18612–18627, 2022b.
- Andrea Skolik, Jarrod R McClean, Masoud Mohseni, Patrick Van Der Smagt, and Martin Leib. Layerwise learning for quantum neural networks. *Quantum Machine Intelligence*, 3:1–11, 2021.
- Tobias Haug and MS Kim. Optimal training of variational quantum algorithms without barren plateaus. *arXiv preprint arXiv:2104.14543*, 2021.
- Farid Ablayev, Marat Ablayev, Joshua Zhexue Huang, Kamil Khadiev, Nailya Salikhova, and Dingming Wu. On quantum methods for machine learning problems part ii: Quantum classification algorithms. *Big Data Mining and Analytics*, 3(1):56–67, 2019.
- Fabio Valerio Massoli, Lucia Vadico, Giuseppe Amato, and Fabrizio Falchi. A leap among quantum computing and quantum neural networks: A survey. *ACM Computing Surveys*, 55(5):1–37, 2022.

- Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. Quantum circuit learning. *Physical Review A*, 98(3):032309, 2018.
- Davis Arthur et al. A hybrid quantum-classical neural network architecture for binary classification. *arXiv preprint arXiv:2201.01820*, 2022.
- Johannes Bausch. Recurrent quantum neural networks. *Advances in neural information processing systems*, 33:1368–1379, 2020.
- Jinjing Shi, Ren-Xin Zhao, Wenxuan Wang, Shichao Zhang, and Xuelong Li. Qsan: A near-term achievable quantum self-attention network. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.
- Liangliang Fan and Haozhen Situ. Compact data encoding for data re-uploading quantum classifier. *Quantum Information Processing*, 21(3):87, 2022.
- Shi-Xin Zhang, Chang-Yu Hsieh, Shengyu Zhang, and Hong Yao. Differentiable quantum architecture search. *Quantum Science and Technology*, 7(4):045023, 2022c.
- Kehuan Linghu, Yang Qian, Ruixia Wang, Meng-Jun Hu, Zhiyuan Li, Xuegang Li, Huikai Xu, Jingning Zhang, Teng Ma, Peng Zhao, et al. Quantum circuit architecture search on a superconducting processor. *Entropy*, 26(12):1025, 2024.
- Xinbiao Wang, Junyu Liu, Tongliang Liu, Yong Luo, Yuxuan Du, and Dacheng Tao. Symmetric pruning in quantum neural networks. *arXiv preprint arXiv:2208.14057*, 2022b.
- Mahabubul Alam, Satwik Kundu, and Swaroop Ghosh. Knowledge distillation in quantum neural network using approximate synthesis. In *Proceedings of the 28th Asia and South Pacific Design Automation Conference*, pages 639–644, 2023.
- Mingze Li, Lei Fan, Aaron Cummings, Xinyue Zhang, Miao Pan, and Zhu Han. Hybrid quantum classical machine learning with knowledge distillation. In *ICC 2024-IEEE International Conference on Communications*, pages 1139–1144. IEEE, 2024.
- Zhan Yu, Hongshun Yao, Mujin Li, and Xin Wang. Power and limitations of single-qubit native quantum neural networks. *Advances in Neural Information Processing Systems*, 35:27810–27823, 2022a.

- Sukin Sim, Peter D Johnson, and Alán Aspuru-Guzik. Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. *Advanced Quantum Technologies*, 2(12):1900070, 2019.
- Kouhei Nakaji and Naoki Yamamoto. Expressibility of the alternating layered ansatz for quantum computation. *Quantum*, 5:434, 2021.
- Zoë Holmes, Kunal Sharma, Marco Cerezo, and Patrick J Coles. Connecting ansatz expressibility to gradient magnitudes and barren plateaus. *PRX Quantum*, 3(1):010313, 2022.
- Zhan Yu, Qiuhan Chen, Yuling Jiao, Yinan Li, Xiliang Lu, Xin Wang, and Jerry Zhijian Yang. Non-asymptotic approximation error bounds of parameterized quantum circuits. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2022b.
- Carlos Ortiz Marrero, Mária Kieferová, and Nathan Wiebe. Entanglement-induced barren plateaus. *PRX Quantum*, 2(4):040316, 2021.
- Samson Wang, Enrico Fontana, Marco Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J Coles. Noise-induced barren plateaus in variational quantum algorithms. *Nature communications*, 12(1):6961, 2021b.
- Martin Larocca, Piotr Czarnik, Kunal Sharma, Gopikrishnan Muraleedharan, Patrick J Coles, and Marco Cerezo. Diagnosing barren plateaus with tools from quantum optimal control. *Quantum*, 6:824, 2022.
- Bobak Toussi Kiani, Seth Lloyd, and Reevu Maity. Learning unitaries by gradient descent. *arXiv preprint arXiv:2001.11897*, 2020.
- Roeland Wiersema, Cunlu Zhou, Yvette de Sereville, Juan Felipe Carrasquilla, Yong Baek Kim, and Henry Yuen. Exploring entanglement and optimization within the hamiltonian variational ansatz. *PRX quantum*, 1(2):020319, 2020.
- Eric R Anschuetz. Critical points in quantum generative models. *arXiv preprint arXiv:2109.06957*, 2021.
- Xuchen You, Shouvanik Chakrabarti, and Xiaodi Wu. A convergence theory for over-parameterized variational quantum eigensolvers. *arXiv preprint arXiv:2205.12481*, 2022.

- Zhiqiang Xu, Xin Cao, and Xin Gao. Convergence analysis of gradient descent for eigenvector computation. International Joint Conferences on Artificial Intelligence, 2018.
- Junyu Liu, Khadijeh Najafi, Kunal Sharma, Francesco Tacchino, Liang Jiang, and Antonio Mezzacapo. Analytic theory for the dynamics of wide quantum neural networks. *Physical Review Letters*, 130(15):150601, 2023.
- Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, Shan You, and Dacheng Tao. Learnability of quantum neural networks. *PRX quantum*, 2(4):040337, 2021b.
- Jun Qi, Chao-Han Huck Yang, Pin-Yu Chen, and Min-Hsiu Hsieh. Theoretical error performance analysis for variational quantum circuit based functional regression. *npj Quantum Information*, 9(1):4, 2023.
- Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner. The power of quantum neural networks. *Nature Computational Science*, 1(6):403–409, 2021.
- Matthias C Caro, Hsin-Yuan Huang, Marco Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J Coles. Generalization in quantum machine learning from few training data. *Nature communications*, 13(1):4919, 2022.
- Kaifeng Bu, Dax Enshan Koh, Lu Li, Qingxian Luo, and Yaobo Zhang. Statistical complexity of quantum circuits. *Physical Review A*, 105(6):062431, 2022.
- Hsin-Yuan Huang, Richard Kueng, and John Preskill. Information-theoretic bounds on quantum advantage in machine learning. *Physical Review Letters*, 126(19):190505, 2021c.
- Kaining Zhang, Junyu Liu, Liu Liu, Liang Jiang, Min-Hsiu Hsieh, and Dacheng Tao. The curse of random quantum data. *arXiv preprint arXiv:2408.09937*, 2024a.
- Yuxuan Du, Yibo Yang, Dacheng Tao, and Min-Hsiu Hsieh. Problem-dependent power of quantum neural networks on multiclass classification. *Physical Review Letters*, 131(14):140601, 2023.
- Kyle Poland, Kerstin Beer, and Tobias J Osborne. No free lunch for quantum machine learning. *arXiv preprint arXiv:2003.14103*, 2020.

- Kunal Sharma, Marco Cerezo, Zoë Holmes, Lukasz Cincio, Andrew Sornborger, and Patrick J Coles. Reformulation of the no-free-lunch theorem for entangled datasets. *Physical Review Letters*, 128(7):070501, 2022.
- Xinbiao Wang, Yuxuan Du, Zhuozhuo Tu, Yong Luo, Xiao Yuan, and Dacheng Tao. Transition role of entangled data in quantum machine learning. *Nature Communications*, 15(1):3716, 2024a.
- Xinbiao Wang, Yuxuan Du, Kecheng Liu, Yong Luo, Bo Du, and Dacheng Tao. Separable power of classical and quantum learning protocols through the lens of no-free-lunch theorem. *arXiv preprint arXiv:2405.07226*, 2024b.
- Eric R Anschuetz, Hong-Ye Hu, Jin-Long Huang, and Xun Gao. Interpretable quantum advantage in neural sequence learning. *PRX Quantum*, 4(2):020338, 2023.
- Eric R Anschuetz and Xun Gao. Arbitrary polynomial separations in trainable quantum machine learning. *arXiv preprint arXiv:2402.08606*, 2024.
- Haimeng Zhao and Dong-Ling Deng. Entanglement-induced provable and robust quantum learning advantages. *arXiv preprint arXiv:2410.03094*, 2024.
- Marco Cerezo, Martin Larocca, Diego García-Martín, Nelson L Diaz, Paolo Braccia, Enrico Fontana, Manuel S Rudolph, Pablo Bermejo, Aroosa Ijaz, Supanut Thanasilp, et al. Does provable absence of barren plateaus imply classical simulability? or, why we need to rethink variational quantum computing. *arXiv preprint arXiv:2312.09121*, 2023.
- Pablo Bermejo, Paolo Braccia, Manuel S Rudolph, Zoë Holmes, Lukasz Cincio, and M Cerezo. Quantum convolutional neural networks are (effectively) classically simulable. *arXiv preprint arXiv:2408.12739*, 2024.
- Armando Angrisani, Alexander Schmidhuber, Manuel S Rudolph, M Cerezo, Zoë Holmes, and Hsin-Yuan Huang. Classically estimating observables of noiseless quantum circuits. *arXiv preprint arXiv:2409.01706*, 2024.
- Sacha Lerch, Ricard Puig, Manuel S Rudolph, Armando Angrisani, Tyson Jones, M Cerezo, Supanut Thanasilp, and Zoë Holmes. Efficient quantum-enhanced classical simulation for patches of quantum landscapes. *arXiv preprint arXiv:2411.19896*, 2024.

- Seongwook Shin, Yong Siah Teo, and Hyunseok Jeong. Dequantizing quantum machine learning models using tensor networks. *Physical Review Research*, 6(2):023218, 2024.
- Jonas Landman, Slimane Thabet, Constantin Dalyac, Hela Mhiri, and Elham Kashefi. Classically approximating variational quantum machine learning with random fourier features. *arXiv preprint arXiv:2210.13200*, 2022.
- Franz J Schreiber, Jens Eisert, and Johannes Jakob Meyer. Classical surrogates for quantum learning models. *Physical Review Letters*, 131(10):100803, 2023.
- Yuxuan Du, Min-Hsiu Hsieh, and Dacheng Tao. Efficient learning for linear properties of bounded-gate quantum circuits. *arXiv preprint arXiv:2408.12199*, 2024.
- Maxwell Henderson, Samriddhi Shakya, Shashindra Pradhan, and Tristan Cook. Quanvolutional neural networks: powering image recognition with quantum circuits. *Quantum Machine Intelligence*, 2(1):2, 2020.
- HTS ALRikabi, Ibtisam A Aljazaery, Jaafar Sadiq Qateef, Abdul Hadi M Alaidi, and M Roa'a. Face patterns analysis and recognition system based on quantum neural network qnn. *Int. J. Interac. Mob. Tech*, 16(8):35–48, 2022.
- Hongni Jin and Kenneth M Merz Jr. Integrating machine learning and quantum circuits for proton affinity predictions. *arXiv preprint arXiv:2411.17856*, 2024.
- Jiwon Roh, Seunghyeon Oh, Donggyun Lee, Chonghyo Joo, Jinwoo Park, Il Moon, Insoo Ro, and Junghwan Kim. Hybrid quantum neural network model with catalyst experimental validation: Application for the dry reforming of methane. *ACS Sustainable Chemistry & Engineering*, 12(10):4121–4131, 2024.
- Yanan Li, Zhimin Wang, Rongbing Han, Shangshang Shi, Jiaxin Li, Ruimin Shang, Haiyong Zheng, Guoqiang Zhong, and Yongjian Gu. Quantum recurrent neural networks for sequential learning. *Neural Networks*, 166:148–161, 2023a.
- Nouhaila Innan, Abhishek Sawaika, Ashim Dhor, Siddhant Dutta, Sairupa Thota, Husayn Gokal, Nandan Patel, Muhammad Al-Zafar Khan, Ioannis Theodonis, and Mohamed Bennai. Financial fraud detection using

- quantum graph neural networks. *Quantum Machine Intelligence*, 6(1):7, 2024.
- Huda Ghazi Enad, Mazin Abed Mohammed, et al. A review on artificial intelligence and quantum machine learning for heart disease diagnosis: Current techniques, challenges and issues, recent developments, and future directions. *Fusion: Pract Appl (FPA)*, 11(1):08–25, 2023.
- Mohammad H Amin, Evgeny Andriyash, Jason Rolfe, Bohdan Kulchytskyy, and Roger Melko. Quantum boltzmann machine. *Physical Review X*, 8(2):021050, 2018.
- Jonathan Romero, Jonathan P Olson, and Alan Aspuru-Guzik. Quantum autoencoders for efficient compression of quantum data. *Quantum Science and Technology*, 2(4):045001, 2017.
- Bingzhi Zhang, Peng Xu, Xiaohui Chen, and Quntao Zhuang. Generative quantum machine learning via denoising diffusion probabilistic models. *Physical Review Letters*, 132(10):100602, 2024b.
- Michael Kölle, Gerhard Stenzel, Jonas Stein, Sebastian Zielinski, Björn Omer, and Claudia Linnhoff-Popien. Quantum denoising diffusion models. *arXiv preprint arXiv:2401.07049*, 2024.
- Manuel S Rudolph, Sacha Lerch, Supanut Thanasilp, Oriel Kiss, Oxana Shaya, Sofia Vallecorsa, Michele Grossi, and Zoë Holmes. Trainability barriers and opportunities in quantum generative modeling. *npj Quantum Information*, 10(1):116, 2024.
- Mohamed Hibat-Allah, Marta Mauri, Juan Carrasquilla, and Alejandro Perdomo-Ortiz. A framework for demonstrating practical quantum advantage: comparing quantum against classical generative models. *Communications Physics*, 7(1):68, 2024.
- Xun Gao, Z-Y Zhang, and L-M Duan. A quantum machine learning algorithm based on generative models. *Science advances*, 4(12):eaat9004, 2018.
- Xun Gao, Eric R Anschuetz, Sheng-Tao Wang, J Ignacio Cirac, and Mikhail D Lukin. Enhancing generative models via quantum correlations. *Physical Review X*, 12(2):021037, 2022.
- Yuxuan Du, Zhuozhuo Tu, Bujiao Wu, Xiao Yuan, and Dacheng Tao. Power of quantum generative learning. *arXiv preprint arXiv:2205.04730*, 2022d.

- Javier Alcazar, Vicente Leyton-Ortega, and Alejandro Perdomo-Ortiz. Classical versus quantum models in machine learning: insights from a finance application. *Machine Learning: Science and Technology*, 1(3):035003, 2020.
- Marcello Benedetti, Edward Grant, Leonard Wossnig, and Simone Severini. Adversarial quantum circuit learning for pure state approximation. *New Journal of Physics*, 21(4):043023, 2019b.
- Junde Li, Rasit O Topaloglu, and Swaroop Ghosh. Quantum generative models for small molecule drug discovery. *IEEE transactions on quantum engineering*, 2:1–8, 2021.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL <https://aclanthology.org/P16-1162>.
- Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Eduardo Blanco and Wei Lu, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL <https://aclanthology.org/D18-2012>.
- Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. Between words and characters: A brief history of open-vocabulary modeling and tokenization in nlp. *arXiv:2112.10508*, 2021.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv:1512.03385*, 2015.
- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL <https://arxiv.org/abs/1607.06450>.

- Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL <http://arxiv.org/abs/1412.6980>.
- Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Benoit Steiner, and Hartwig Rolland. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2704–2713, 2018.
- Song Han, Jeff Pool, John Tran, and William J Dally. Learning both weights and connections for efficient neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1135–1143, 2015.
- Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.
- Joseph McDonald, Baolin Li, Nathan Frey, Devesh Tiwari, Vijay Gadepally, and Siddharth Samsi. Great power, great responsibility: Recommendations for reducing energy for training language models. *Findings of the Association for Computational Linguistics: NAACL 2022*, 2022. doi: 10.18653/v1/2022.findings-naacl.151. URL <http://dx.doi.org/10.18653/v1/2022.findings-naacl.151>.
- Radosvet Desislavov, Fernando Martínez-Plumed, and José Hernández-Orallo. Trends in ai inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. *Sustainable Computing: Informatics and Systems*, 38:100857, 2023. ISSN 2210-5379. doi: <https://doi.org/10.1016/j.suscom.2023.100857>. URL <https://www.sciencedirect.com/science/article/pii/S2210537923000124>.
- Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv:2203.03466*, 2022.
- Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shao-han Huang, Li Dong, Ruiping Wang, Jilong Xue, and Furu Wei. The era

- of 1-bit llms: All large language models are in 1.58 bits. *arXiv:2402.17764*, 2024.
- Guang Hao Low. *Quantum Signal Processing by Single-Qubit Dynamics*. Thesis, Massachusetts Institute of Technology, 2017. URL <https://dspace.mit.edu/handle/1721.1/115025>.
- Iordanis Kerenidis and Anupam Prakash. Quantum Recommendation Systems. *arXiv:1603.08675*, September 2016.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- Guangxi Li, Xuanqiang Zhao, and Xin Wang. Quantum self-attention neural networks for text classification, 2023b. URL <https://arxiv.org/abs/2205.05625>.
- El Amine Cherrat, Iordanis Kerenidis, Natansh Mathur, Jonas Landman, Martin Strahm, and Yun Yvonna Li. Quantum vision transformers. *Quantum*, 8:1265, February 2024. ISSN 2521-327X. doi: 10.22331/q-2024-02-22-1265. URL <http://dx.doi.org/10.22331/q-2024-02-22-1265>.
- Ethan N. Evans, Matthew Cook, Zachary P. Bradshaw, and Margarite L. LaBorde. Learning with sasquatch: a novel variational quantum transformer architecture with kernel-based self-attention. 2024. URL <https://arxiv.org/abs/2403.14753>.
- Dominic Widdows, Willie Aboumrad, Dohun Kim, Sayonee Ray, and Jonathan Mei. Quantum natural language processing. 2024. URL <https://arxiv.org/abs/2403.19758>.
- Yeqi Gao, Zhao Song, Xin Yang, and Ruizhe Zhang. Fast quantum algorithm for attention computation. 2023. URL <https://arxiv.org/abs/2307.08045>.
- Chen-Yu Liu, Chao-Han Huck Yang, Min-Hsiu Hsieh, and Hsi-Sheng Goan. A quantum circuit-based compression perspective for parameter-efficient learning. 2024b. URL <https://arxiv.org/abs/2410.09846>.

- Siyi Yang, Naixu Guo, Miklos Santha, and Patrick Rebentrost. Quantum Alphatron: quantum advantage for learning with kernels and noise. *Quantum*, 7:1174, November 2023. ISSN 2521-327X. doi: 10.22331/q-2023-11-08-1174. URL <https://doi.org/10.22331/q-2023-11-08-1174>.
- Chenyi Zhang and Tongyang Li. Comparisons are all you need for optimizing smooth functions, 2024. URL <https://arxiv.org/abs/2405.11454>.
- Hao Wang, Chenyi Zhang, and Tongyang Li. Near-optimal quantum algorithm for minimizing the maximal loss, 2024c. URL <https://arxiv.org/abs/2402.12745>.
- Patrick Rebentrost, Maria Schuld, Leonard Wossnig, Francesco Petruccione, and Seth Lloyd. Quantum gradient descent and newton’s method for constrained polynomial optimization, 2018b. URL <https://arxiv.org/abs/1612.01789>.
- Yidong Liao and Chris Ferrie. Gpt on a quantum computer. 2024. URL <https://arxiv.org/abs/2403.09418>.
- Nikhil Khatri, Gabriel Matos, Luuk Coopmans, and Stephen Clark. Quixer: A quantum transformer model. 2024. URL <https://arxiv.org/abs/2406.04305>.
- Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.
- Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.
- Masashi Okamoto. Some inequalities relating to the partial sum of binomial probabilities. *Annals of the institute of Statistical Mathematics*, 10:29–35, 1959.
- Alfred Haar. Der massbegriff in der theorie der kontinuierlichen gruppen. *Annals of mathematics*, 34(1):147–169, 1933.
- Gabriel Nagy. On the haar measure of the quantum su (n) group. *Communications in mathematical physics*, 153:217–217, 1993.
- Martin Adam. *Applications of unitary k-designs in quantum information processing*. PhD thesis, Masarykova univerzita, Fakulta informatiky, 2013.

Martin Larocca, Supanut Thanasilp, Samson Wang, Kunal Sharma, Jacob Biamonte, Patrick J Coles, Lukasz Cincio, Jarrod R McClean, Zoë Holmes, and M Cerezo. A review of barren plateaus in variational quantum computing. *arXiv preprint arXiv:2405.00781*, 2024.

Z Puchała and JA Miszczak. Symbolic integration with respect to the haar measure on the unitary groups. *Bulletin of the Polish Academy of Sciences. Technical Sciences*, 65(1):21–27, 2017.

Aidan Roy and Andrew J Scott. Unitary designs and codes. *Designs, codes and cryptography*, 53:13–31, 2009.