Noise Tailoring for Enhancing the Capabilities of Quantum Computers

by

Akel Hashim

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Applied Science & Technology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Irfan Siddiqi, Chair
Professor Birgitta Whaley
Professor Harmut Häffner

Summer 2022

Noise Tailoring for Enhancing the Capabilities of Quantum Computers

Abstract

Noise Tailoring for Enhancing the Capabilities of Quantum Computers

by

Akel Hashim

Doctor of Philosophy in Applied Science & Technology

University of California, Berkeley

Professor Irfan Siddiqi, Chair

The successful implementation of algorithms on quantum processors relies on the accurate control of quantum bits (qubits) to perform logic gate operations. However, qubits in the noisy intermediate-scale quantum (NISQ) era are short-lived and susceptible to a variety of errors and noise due to imperfect control signals and incomplete isolation from the surrounding environment. For example, systematic miscalibrations, unwanted entanglement, and crosstalk in the control of qubits can lead to a coherent form of error which has no classical analog. Coherent errors can severely limit the performance of quantum algorithms in an unpredictable manner on timescales shorter than the coherence times of qubits. In recent years, there has been growing interest in using methods which randomize the physical implementation of quantum gates to mitigate the impact of coherent errors, effectively tailoring them into a form of stochastic noise. In this thesis, we study one such method — randomized compiling — and show how gate errors under randomized compiling are accurately described by a stochastic Pauli noise model *without* coherent errors. We demonstrate significant performance gains under randomized compiling for various different quantum algorithms, such as the quantum Fourier transform. We further show that randomized compiling can improve the predictability of quantum algorithms, and enables unique forms of error mitigation for enhancing the performance of quantum computations in the NISQ era. Finally, we show that randomized compiling can reduce worst-case error rates by orders of magnitude, enabling the accurate characterization of quantum gates for fault tolerance. Our results demonstrate that randomized compiling can be utilized to leverage and predict the capabilities of modern-day, noisy quantum processors, paving the way forward for scalable quantum computing and fault tolerant quantum error correction.

To my wife, Tracey, whose unwavering love and support carried me through this journey.

To my sisters, Sahalie and Sada, who made me who I am today.

To my parents, Rosalie, Bill, and Suzanne, who taught me to follow my dreams without question.

To those restless few who dared to venture into the unknown to discover the quantum.

Contents

List of Figures

List of Tables

Acknowledgments

Obtaining a doctorate has not been so much about research papers and my thesis topic, but rather the path I took to get here, and the knowledge I learned along the way. Like any endeavour which takes years to complete, there are innumerable people whose help and support were critical my success, and many points along the way where, had something slightly different happened, I would not be where I am today. I can only begin to thank those who have most contributed to my journey. However, needless to say, there are many others who have also contributed to my success, although they may not know it; to all of you who remained unnamed on this page, thank you.

To my research advisor and friend, Irfan Siddiqi — thank you for putting your trust in me, for guiding my learning and understanding of the complexities of experimental physics, for giving me the freedom to explore the topics that interested me the most, and for cultivating an environment of success for everyone in your lab.

To Kasra Nowrouzi — thank you for being my first friend in Berkeley, for pushing me outside of my comfort zone, and for enthusiastically supporting my research endeavours, despite any personal reservations I had myself.

To Ravi Naik — thank you for teaching me most of what I know about superconducting qubits, for your never-ending patience in (re)explaining complex topics, and for your continued support and dedication to mentoring during the pandemic.

To the post-docs and members of the Quantum Nanoelectronics Laboratory who came before me, Ravi Naik, Brad Mitchell, Machiel Blok, Vinay Ramasesh, Leigh Martin, Will Livingston, John Mark Kreikebaum, Marie Lu, Noah Stevenson, Gerwin Koolstra, Kasra Nowrouzi, Alexis Morvan, Jean-Loup Ville, and many others — thank you for being the foundation for learning within the lab, for being the stewards of knowledge transfer to younger students, and for driving the success of the research projects within the group.

To Professor Robert Littlejohn — thank you for teaching graduate quantum mechanics with such care, passion, dedication, and detail. I have never learned more in a single class than I did taking your year-long course on quantum mechanics. Your class, and the effort I put into it, encouraged me to change course in graduate school, and in large part shaped my graduate research.

To Joseph Emerson, Joel Wallman, Ian Hincks, and Samuele Ferracin — thank you for teaching me all things related to randomized benchmarking, cycle benchmarking, randomized compiling, and error mitigation. Our collaborations have been a core component of my PhD.

To Kenny Rudinger, Tim Proctor, Stefan Seritan, and the other members of the Quantum Performance Laboratory — thank you for teaching me about gate set tomography and the nuances of various benchmarking methods, and for continually putting up with my probing questions and requests for new capabilities.

To the Applied Science & Technology PhD program — thank you for giving me the opportunity to complete my PhD at UC Berkeley. The freedom to design my own PhD was truly unique.

To the National Defense Science & Engineering Graduate (NDSEG) Fellowship program — I would not be here had I not received the NDSEG fellowship. Five years ago, I was not accepted into any of my schools of choice, so I decided to put off my decision to go to grad school one more year. In the middle of drafting my deferment letters, I received the notification that I had been selected for the NDSEG fellowship. Upon requesting UC Berkeley to reconsider my application due to the fellowship funding, they reversed their decision and granted my acceptance. For this I will be forever grateful.

To my parents, Rosalie, Bill, and Suzanne — thank you for supporting my every decision, for teaching me humility, respect for all others, and to believe in myself no matter the odds. Your unconditional pride in my achievements is an ever-present driving force in my life.

To my sisters, Sahalie and Sada — thank you for always being there for me, for supporting me in difficult moments, and for sharing in the excitement of my successes. You are not only my sisters, but also my best friends, and you have been the biggest influence in my life in shaping who I am today.

To my cats, Niko and Chloe, and my dog, Jasmine — thank you for keeping me company during the pandemic, for video bombing my Zoom meetings, and for being a continued source of distraction throughout my PhD.

To my wife and companion, Tracey — our path these past 11 years has been one of shared difficulty and shared success. We sacrificed much to pursue our dreams – seven years in a long-distance relationship, two of which we lived in different states. Of the many times we felt alone, we were never truly alone, both of us ready and willing to drop everything for the other. I would not be where I am today without your unwavering support. I take comfort in knowing that whatever journey we take, and whatever hurdles we encounter, we'll never walk alone.

Chapter 1

Introduction

1.1 Current State of the Field

We have entered what many are calling the "second quantum revolution" [1, 2, 3]. While the *science* of quantum mechanics was developed 100 years ago, we are now beginning to learn how to harness the *power* of quantum mechanics, but we do not yet understand the limits to which we can control quantum systems. In the last two decades, the fields of quantum computing and quantum information science (QIS) have grown at a rapid pace, fueled in part by the excitement of developing technologies based fundamentally on quantum mechanics, and in part by the fear of being left behind in the race to be the first country to build a quantum computer capable of achieving the potential we know exists. Several different "laws" governing the growth of quantum computers have been claimed, in analogy to Moore's law for transistors. Regarding superconducting quantum bits (qubits), "Girvin's law" states that qubit coherence times (e.g. lifetimes) are growing at an exponential rate [4, 5], increasing by a factor of $\sim 10\times$ every three years. Regarding the computational power of quantum computers relative to classical computers, "Neven's law" states that quantum computers are growing at a doubly exponential rate. Whether these claims hold true in the distant future remains to be seen, but needless to say, the power of quantum computers is growing rapidly, bolstered in large part by the vast amount of money in the field, with tens of billions of dollars in public and private funding having been invested in the past 20 years [6]. Despite this, the central question still exists: are quantum computers capable of solving problems that classical computers cannot solve in any reasonable amount of time? Already, several claims to quantum supremacy or quantum advantage have been made using superconducting and photonics-based hardware platforms [7, 8, 9, 10] for carefully designed problems. However, the long-term implications of these experiments are perhaps not yet known, and demonstrating quantum advantage for "practical" applications [11] is still very much an open question.

Much of the current research in the field is focused on improving or building high-coherence qubits capable of performing longer quantum computations [12]. This is par-

ticularly important for superconducting qubits which, despite the fast speeds at which they can perform quantum logic operations, suffer from shorter coherence times compared to atomic-based systems [13]. However, in this era of noisy intermediate-scale quantum (NISQ) [14] computers, qubit coherence times and operations times are not the only characteristics that matter for quantum computations. NISQ era computations are called "noisy" for a reason, not only because noise from the surrounding environment can couple to qubits and cause decoherence, but also because other novel forms of errors can occur which, as we will see in this thesis, have no classical analogue. If the global impact of all such noise and errors limits the timescale within which one can perform useful computations, and if this timescale is less than the coherence times of the qubits, then coherence times are no longer the limiting factor. While engineering high-coherence qubits is of utmost important in the development of scalable quantum computers needed for fault tolerance, work must also be done to ensure that quantum computations can be reliably performed up to the limits of their coherence times.

This thesis focuses on the aforementioned regime, in which the sum of all errors and noise limits the useable timescale of qubit operations below that of their coherence times. We therefore do not worry about qubit coherence times, but rather what types of errors are most pernicious to quantum computations within the timescale of the coherences, and what can be done to mitigate their effects. While the experimental work presented in this thesis was performed on superconducting circuits, none of the methods we describe are specific to superconducting qubit platforms. In fact, the methods and protocols we introduce are general, and are platform agnostic. Therefore, our results have broad applicability across many different hardware platforms.

The thesis is structured as follows: Chapter 2 covers the basic physics of superconducting circuits, how quantum gates are implemented, and how superconducting qubits satisfy all of the requirements necessary for universal quantum computation. Chapter 3 introduces common noise and errors in quantum computations, and different ways of representing quantum processes. Chapter 4 introduces the main concepts behind "noise tailoring," a central theme in this thesis, which will be necessary in order to understand the methods introduced in the subsequent chapters. Chapter 5 outlines the field of quantum characterization, verification, and validation, which encompasses various different methods for benchmarking the performance of quantum gates and quantum processors; we introduce several different benchmarking protocols, and discuss what they can (and cannot) tell us about the performance of general quantum circuits. Chapter 6 covers the primary focus of this thesis — a protocol called randomized compiling, which can be used to change the structure of noise and errors impacting quantum algorithms. Finally, in Chapter 7 we show how randomized compiling enables different novel error mitigation methods for improving the performance of NISQ applications. Chapter 8 concludes with an outlook on randomized compiling and the future of QIS.

1.2 Recommended Reading

Every grad student entering a new field asks the same basic question, "where do I start?" Diving into the literature of any field of physics is a daunting task, not only because the number of research papers on any given topic is large (and growing rapidly!), but also because one's knowledge base may not yet be mature enough to understand the topics and math presented in contemporary research articles. Most graduate courses only teach the fundamentals and rarely have time to delve into specialized topics. In this section, I aim to provide some recommended material for students just starting out in the field of superconducting quantum computing and QIS. This list is by no means comprehensive, and is bias toward experimental physics and superconducting qubits. In other words, it is the list I wish I had when just starting out in the field.

Quantum Optics & Cavity QED

Many of the experimental techniques employed in superconducting circuits are based on quantum optics and cavity quantum electrodynamics (CQED). Therefore, having a basic understanding of quantum optics is extremely important for understanding the physics of superconducting qubits. For example, the treatment of the Jaynes-Cummings Hamiltonian (i.e. the interaction of light with a two-level system; see Section 2.6) is applicable to many different hardware platforms in quantum computing. Furthermore, topics like coherent and squeezed states, vacuum Rabi oscillations, parametric amplification, quantum nondemolition (QND) measurements, and quantum trajectories all have their roots in quantum optics. Below is a list of textbooks that cover these topics in great detail:

1. ***Quantum Optics*** – **Scully & Zubairy** [15]: This is a classic textbook on quantum optics. While some of the notation might appear a bit dated, the material and diagrams are extremely useful for understanding many of the core topics in quantum optics. Scully & Zubairy uses the Jaynes-Cummings model extensively starting in chapter 6, deriving the Hamiltonian from first principles, even though they never refer it by name. The book has a nice treatment of the EPR Paradox, and finishes with a discussion of more advanced topics such as complementarity, which-path detectors, and the quantum eraser.

2. ***Introductory Quantum Optics*** – **Gerry & Knight** [16]: This is a good introductory textbook to quantum optics, including a thorough introduction to QED. The book covers experimental topics such as CQED and trapped ions, and finishes with an introduction to quantum cryptography.

3. ***Quantum Optics*** – **Walls & Milburn** [17]: The updated version of this textbook offers a modern treatment of quantum optics with clean, concise notation, and full chapter treatments of more advanced topics, such as CQED, trapped ion systems, and quantum information (qubits, entanglement, gates, etc.).

4. ***Exploring the Quantum: Atoms, Cavities, and Photons*** – **Haroche & Raimond** [18]: This is a must read for anyone interested in experimental QIS. While the treatment of CQED is (naturally) based on atomic systems, Haroche & Raimond provide a strong connection between experimental methods and quantum gates in circuit-based quantum computing, which is applicable to many other hardware platforms. Their discussion of quantum measurement at the quantum-classical boundary was particularly illuminating.

Superconducting Quantum Computing

The field of superconducting quantum computing is based on the physics of the Josephson junction. Therefore, knowledge of superconductivity and the Josephson Effect is necessary for understanding how quantum circuits are implemented on Josephson junction-based qubits. While many different textbooks on superconductivity exists, the ***Introduction to Superconductivity*** – **Tinkham** [19] is an excellent reference for those interested in superconducting qubits. Tinkham dedicates two full chapters to the Josephson Effect, while also providing a full treatment of the Ginzburg-Landau and BCS theories of superconductivity.

To go from Josephson junctions, to superconducting qubits, to superconducting quantum computing contains a lot of different physics. The following theses and research papers attempt to bridge these gaps, and provide foundational knowledge for the entire field of superconducting quantum computing:

1. ***Introduction to Quantum Electromagnetic Circuits*** – **Vool & Devoret** [20]: This article is based on a series of summer school lectures that provide a nice treatment of superconducting qubits, starting from classical circuits and ending with artificial atoms based on quantized LC circuits.

2. ***The superconducting circuit companion: an introduction with worked examples*** – **Rasmussen et al.** [21]: This article is a nice companion to Vool & Devoret, but provides a more foundational framework for understanding superconducting qubits in terms of circuit analysis.

3. ***A Quantum Engineer's Guide to Superconducting Qubits*** – **Krantz et al.** [22]: This is an *excellent* review article for all things related to superconducting quantum computing, starting from system Hamiltonians, and ending with readout and amplification of superconducting qubits. Re-reading this article 2 – 3 times would provide all of the basic foundational knowledge necessary for understanding circuit-based quantum computing using superconducting qubits.

4. ***Circuit quantum electrodynamics*** – **Blais et al.** [23]: This article provides an thorough overview of circuit QED (cQED)[1], the circuit equivalent of cavity QED, in

[1]Not to be confused with cavity QED, which we denote CQED.

which microwave resonators are coupling to superconducting qubits in order to realize the atom-light interactions governed by the Jaynes-Cummings Hamiltonian.

5. ***Circuit quantum electrodynamics*** – **Schuster** [24]: This thesis provides an in-depth overview of cQED, for those wanting more information beyond Blais et al.

6. ***Quantum Jumps and Measurement Backaction in a Superconducting Qubit*** – **Slichter** [25]: The first three chapters of this thesis provide an excellent introduction to both superconducting qubits and parametric amplification, and is a standard recommendation for new students joining the Quantum Nanoelectronics Laboratory.

Quantum Computing

While many introductory books exist on the topic quantum computing, there is no substitute for ***Quantum Computation and Quantum Information*** – **Nielson & Chuang** [26], more commonly referred to as "Mike & Ike." There is a reason this book (at the time of this writing) has been cited over 45,000 times. It covers the basics of quantum mechanics, classical computing, complexity theory, quantum circuits, quantum algorithms, representations of quantum processes and error channels, distance measures and error metrics, and error correction. I have heard many senior researchers in the field still refer to Mike and Ike as their go-to reference.

Any subtopic within the field of quantum computing is likely to have its own dedicated review article or book (and if not now, then in the near future). For example, I found ***Variational quantum algorithms*** – **Cerezo et al.** [27] an extremely useful introduction to the entire field of variational quantum algorithms and their contemporary challenges. Beyond introducing a specific topic, these types of review papers can be a great resource for finding the proper references for a given subject matter.

Miscellaneous

Many other great introductory articles exist. An example of one which does not fall into any of the above categories is the ***Introduction to quantum noise, measurement, and amplification*** – **Clerk et al.** [28], which is the *de facto* standard reference for quantum noise and the fundamental limits it places on quantum measurement and amplification. Those looking for a brief introduction to general measurement theory will also find this article useful.

General Advice

When I was starting out in the field, I read a similar list of suggested works that I found extremely useful [29]. While the examples provided above are by no means exhaustive, each student or reader will end up compiling their own list of "most useful introductory material."

I highly recommend sharing this list with as many younger students as possible, as theses are one of the best ways to transfer knowledge from one generation of students to the next.

Beyond basic introductory material, staying up-to-date on current research trends and recent results is extremely important. For those unsure of how to do so or overwhelmed by the vast number of papers published every day, I offer the following advice:

1. Make reading or perusing recent papers part of your daily routine. This need not take up a significant amount of time out of your day.

2. The quickest way to learn is to read about topics that interest you the most. Read papers that pique your interest, and follow the cited works in order to fill in the background material. You will soon become very knowledgeable in fields in which you are not an expert or are not currently researching yourself.

3. Do not worry about reading every paper from start to finish. If you find a paper that sounds interesting, start by reading the abstract, introduction, and conclusion. The basic concepts and results should all be contained within these sections. If you are interested in learning more, try to understand the figures just from reading the figure captions. Still interested? Try reading the bulk of the paper. But if you quickly become lost or confused, it is okay to move on to something else. You can always come back to the paper at a later time.

4. A good way to stay on top of recent results is to filter new posting on arXiv depending on your field of interest. For example, visiting `https://arxiv.org/list/quant-ph/new` lists new submissions under "Quantum Physics" for the current day. If you are interested in a more social media approach, try perusing the top papers on SciRate, where anyone can vote on which recent arXiv papers they find most interesting.

5. Collect and organize papers that you are interested in and/or are applicable to your current research. This will become useful later on when you are trying to remember in which paper you saw a certain result, or if you need to cite various papers in your own manuscript or thesis. One great resource for organizing research papers is Zotero.

6. Offer to present a research paper that interests you to the rest of your group or in a journal club. This will force you to learn the material to the best of your ability in a short time span. Furthermore, presenting and discussing results with other students or researchers in your field is a great way to solidify and expand upon what you already learned.

7. If research papers are difficult to learn from, try finding recorded videos of the topic (e.g. on YouTube). For example, many research colloquia will post recorded videos of invited presenters who have published top research papers in the field. You might even find a talk covering the exact paper you are interested in.

8. There is no one-fits-all approach. Find what works best for you!

Chapter 2

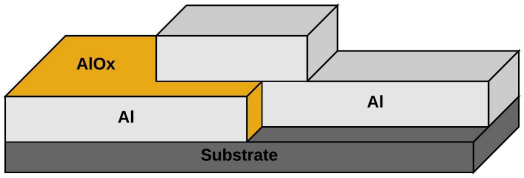
Superconducting Quantum Computing

Superconducting quantum bits (qubits) are at the forefront of the race to build practical quantum computers. The goal of this chapter is to give a brief (yet comprehensive) overview of their foundations — sufficient to build upon and understand the material presented in later chapters — such that future students looking to use this thesis as a reference can understand the basic physics of superconducting qubits from first principles.

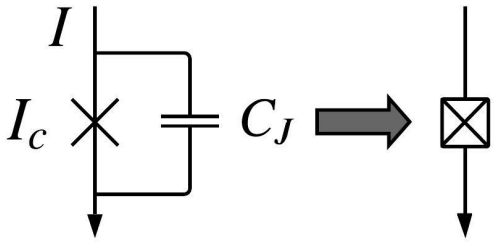
2.1 The Josephson Effect: the primitive building block of superconducting qubits

The Josephson effect is a phenomenon in superconductivity in which Cooper pairs can tunnel between two superconducting electrodes separated by a thin insulating barrier, resulting in a superconducting current (or supercurrent) across the junction, even in the absence of an applied voltage [30]. Prior to the Josephson effect, quantum tunneling of normal (non-superconducting) electrons was a well-known concept. The prediction of quantum tunneling of Cooper pairs by physicist Brian Josephson and the subsequent experimental verification [31, 32] led to the naming of such effect after Josephson, for which he was awarded the 1973 Nobel Prize in Physics.

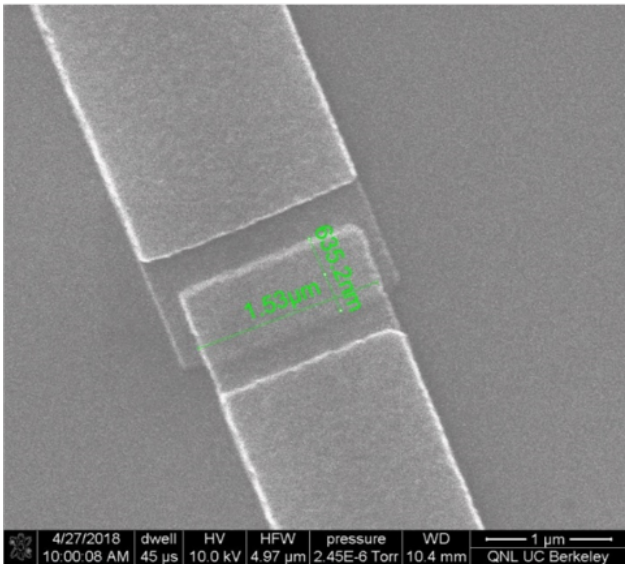
While there are many applications of the Josephson effect, such as Superconducting Quantum Interference Devices (SQUIDs) [33, 34], ultra sensitive magnetometers [35], gradiometers [36], voltmeters [37], amplifiers [38, 39], etc., more recently it has been utilized in the creation of superconducting qubits [40, 41]. In fact, the Josephson junction (JJ) is the fundamental component that makes possible the three basic superconducting qubit archetypes: the flux, charge, and phase qubits. In this chapter, we describe the physics behind the Josephson effect and how it has been utilized to build superconducting qubits. We then discuss the design and behavior of the three primary qubit types, and highlight in detail the most common superconducting qubit archetype — the transmon qubit [42].



(a) Typical layout of an S-I-S junction.



(b) Josephson junction circuit element.



(c) SEM image of an actual S-I-S junction.

Figure 2.1: **Josephson junction.** (a) The layout of a typical S-I-S JJ. Two superconducting electrodes are separated by an insulating (oxide) barrier, forming a junction such as Al-AlOx-Al. The entire junction sits upon some substrate, such as Si. (b) Circuit diagram of a JJ. I is current flowing through the junction, I_c is the junction-dependent critical current, and C_J is the intrinsic capacitance of the junction. (c) An SEM image of an actual junction. Electron-beam lithography allows nanometer resolution for determining the area of overlap between the two electrodes. This image was taken at the Quantum Nanoelectronics Laboratory at UC Berkeley.

Finally, we consider the requirements that a physical system must fulfill in order to build a quantum computer, and discuss how JJ-based qubits satisfy these criteria.

The Josephson Junction

The Josephson effect can be observed in a variety of physical systems, other than simply across an insulating barrier. In fact, the only requirement is that there is a "weak link" connecting two superconducting electrodes. There are three typical types of JJs:

1. Superconductor-Insulator-Superconductor (*S-I-S*): the two electrodes are separated by an insulating barrier.

2. Superconductor-Normal metal-Superconductor (*S-N-S*): the two superconducting electrodes are separated by a gap and sit on top of a substrate made of non-superconducting "normal" metal. The only method of tunneling from one electrode to the other is via

the *proximity effect*, in which the Copper pairs diffuse into the layer of normal metal, creating a thin, weakly superconducting layer between the two electrodes.

3. Superconductor-constriction-Superconductor (S-c-S): the two superconducting electrodes are connected by a narrow, but otherwise continuous, superconducting material. This method provides a "constricted" path for the Cooper pairs to travel from one electrode to the other.

The most common junction is the S-I-S type, with an insulating barrier typically made from an oxide material. For example, it may be constructed from two aluminum electrodes, with an aluminum-oxide layer in between as the insulator (i.e. Al-AlOx-Al). Because a JJ is formed by two electrodes separated by an insulator, there is an intrinsic parallel-plate capacitance C_J associated with the junction. It is also characterized by a critical current I_c, which is the maximum supercurrent that the junction can sustain. In circuit diagrams, a JJ is denoted by a boxed "X". A schematic diagram of a typical junction, its circuit equivalent, and an image taken of a real junction using scanning electron microscopy (SEM) can be seen in Fig. 2.1.

The typical method of fabricating JJs is to use electron-beam lithography to etch away resist sitting on top of a substrate, and then use thin-film deposition of aluminum at multiple angles (called double-angle evaporation), while allowing the first layer to oxidize before depositing the second layer. This method is referred to as the Niemeyer-Dolan technique for fabricating very small overlapping structures. (For a full explanation of this technique, the reader is referred to Ref. [25].)

The Josephson Effect

In his original paper, Josephson postulated the tunnelling current density through an insulating barrier separating two superconducting electrodes to be

$$j = j_0 + \frac{1}{2} j_1 \psi_1^* \psi_2 + \frac{1}{2} j_1^* \psi_2^* \psi_1, \tag{2.1}$$

where j_0 is the DC supercurrent density, $|j_1|$ is related to the matrix element governing the transfer of Cooper pairs across the barrier, and ψ_1 and ψ_2 are quasiparticle wavefunctions on the left- and right-hand sides of the junction, respectively. Josephson made two predictions from the above formula:

1. At zero applied voltage, j_0 will be zero, but a maximum DC supercurrent of magnitude $|j_1|$ can still flow through the junction. This is known as the DC Josephson effect.

2. At a nonzero voltage difference V, a DC current j_0 will be superimposed by an AC current of amplitude $|j_1|$ and frequency $f = 2eV/h$. This is known as the AC Josephson effect.

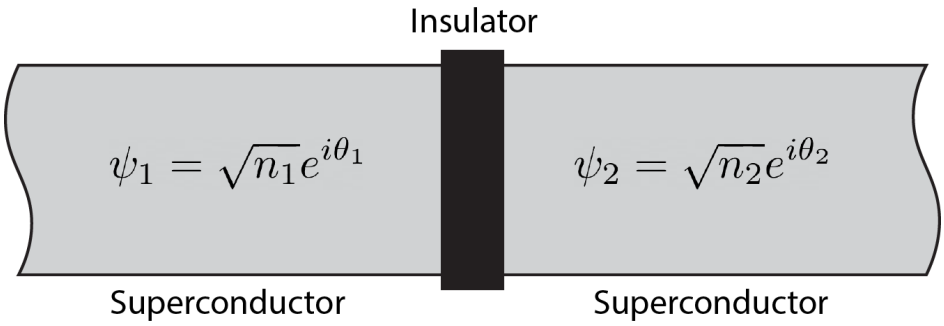


Figure 2.2: **The Josephson Effect.** A JJ is composed of two superconductors separated by a thin insulating barrier. Each superconductor can be described by a macroscopic wavefunction ψ describing the number density n of Cooper pairs and phase θ of the macroscopic superconductor. The phase difference $\Delta = \theta_2 - \theta_1$ across the junction controls the underlying physics of Cooper pairs tunneling across the thin insulating barrier.

From the second conclusion above, we can equate a phase evolution to an applied voltage:

$$\omega = \frac{d\theta}{dt} = \frac{2eV}{\hbar} \tag{2.2}$$

where $2e/h \approx 483.6$ GHz/mV. For example, an applied voltage of 10 μV equates to a characteristic oscillation frequency of approximately 5 GHz for the AC supercurrent flowing across the junction. Given that we are considering the superposition of states in two separated superconducting regions, there is no guarantee that the phases of ψ_1 and ψ_2 are the same. In fact, in order to treat the problem correctly, Josephson required that the current density be a superposition of states with arbitrary phases that are allowed vary over the thickness of the barrier. Thus, the total supercurrent density j will depend directly on the total overlap and phase difference between ψ_1 and ψ_2.

The underlying physics of JJs involves the discrete tunneling of Cooper pairs across the thin insulating barrier. While a proper treatment of this problem can be derived using Bogoliubov operators and the BCS theory of superconductivity [30, 43], because we are dealing with macroscopic states of superconductors, it is sufficient to treat the problem in the Ginzburg-Landau approach. Consider the Ginzburg-Landau wavefunction (or order parameter) describing the (uncoupled) superconductor on either side of an insulating barrier (see Fig. 2.2),

$$\psi_i = \sqrt{n_i}e^{i\theta_i}, i \in \{1, 2\} \tag{2.3}$$

where n_i is the number density of Cooper pairs and θ_i the phase of each superconductor. The uncoupled wavefunctions are normalized such that

$$\int d^3r \psi_i^* \psi_j = \delta_{ij}. \tag{2.4}$$

The Schrödinger equations for the uncoupled regions are

$$i\hbar\frac{\partial\psi_i}{\partial t} = E_i\psi_i, \tag{2.5}$$

where E_i is the on-site energy of each electrode. If we apply a voltage V across the junction, the two superconductors will be separated by an energy difference of $2eV$ and the on-site energy is the potential energy of a Cooper pair of electrons on either side of the junction: $E_1 = (2e)(\frac{1}{2}V) = eV$ and $E_2 = -eV$, where we have defined the potential energy to be zero at the midpoint for convenience (i.e. a floating potential).

If we now consider the Schrödinger equation for the full system,

$$i\hbar\frac{\partial\Psi}{\partial t} = H\Psi, \tag{2.6}$$

the total Hamiltonian H must also contain a term which couples the two regions, and can be decomposed into a 2×2 matrix with elements

$$H_{ij} = \int d^3r\psi_i^* H\psi_j, \tag{2.7}$$

where diagonal elements of H_{ij} are the on-site potential energies of the Cooper pairs $H_{11} = eV$ and $H_{22} = -eV$, and the off-diagonal matrix elements are related to the hopping of Cooper pairs across the barrier, and are identical. Thus, we can define $H_{12} = H_{21} = -T$, where T is related to the tunneling amplitude. We can consider these terms to represent the energy it takes for a tunneling event to occur. The full Schrödinger equation can be now written as

$$i\hbar\frac{\partial}{\partial t}\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix} = H\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix} = \begin{pmatrix}eV & -T\\ -T & -eV\end{pmatrix}\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix}. \tag{2.8}$$

However, this matrix equation can be reduced a set of coupled differential equations:

$$\begin{aligned} i\hbar\frac{\partial\psi_1}{\partial t} &= eV\psi_1 - T\psi_2,\\ i\hbar\frac{\partial\psi_2}{\partial t} &= -T\psi_1 - eV\psi_2. \end{aligned} \tag{2.9}$$

Recalling our definitions $\psi_1(t) = \sqrt{n_1(t)}e^{i\theta_1(t)}$ and $\psi_2(t) = \sqrt{n_2(t)}e^{i\theta_2(t)}$, where we have highlighted the specific time dependence, we can simultaneously solve the equations:

$$\begin{aligned} i\hbar\frac{\partial}{\partial t}\Big(\sqrt{n_1(t)}e^{i\theta_1(t)}\Big) &= eV\Big(\sqrt{n_1}e^{i\theta_1}\Big) - T\Big(\sqrt{n_2}e^{i\theta_2}\Big),\\ i\hbar\frac{\partial}{\partial t}\Big(\sqrt{n_2(t)}e^{i\theta_2(t)}\Big) &= -T\Big(\sqrt{n_1}e^{i\theta_1}\Big) - eV\Big(\sqrt{n_2}e^{i\theta_2}\Big). \end{aligned}$$

Computing the derivatives,

$$\frac{i\hbar}{2\sqrt{n_1}}\frac{dn_1}{dt} - \hbar\sqrt{n_1}\frac{d\theta_1}{dt} = eV\sqrt{n_1} - T\sqrt{n_2}e^{i(\theta_2-\theta_1)},$$
$$\frac{i\hbar}{2\sqrt{n_2}}\frac{dn_2}{dt} - \hbar\sqrt{n_2}\frac{d\theta_2}{dt} = -eV\sqrt{n_2} - T\sqrt{n_1}e^{-i(\theta_2-\theta_1)},$$

we can see that these equations can be separated into real and imaginary solutions,

$$\begin{aligned}\hbar n_1\frac{d\theta_1}{dt} &= -eVn_1 + T\sqrt{n_1n_2}\cos(\theta_2-\theta_1)\\ \hbar n_2\frac{d\theta_2}{dt} &= eVn_2 + T\sqrt{n_1n_2}\cos(\theta_2-\theta_1)\end{aligned} \tag{2.10}$$

and

$$\begin{aligned}\hbar\frac{dn_1}{dt} &= -2T\sqrt{n_1n_2}\sin(\theta_2-\theta_1)\\ \hbar\frac{dn_2}{dt} &= 2T\sqrt{n_1n_2}\sin(\theta_2-\theta_1),\end{aligned} \tag{2.11}$$

respectively.

From the set of imaginary solutions (Eq. 2.11), we can see that current (or charge) conservation across the junction is guaranteed by $\dot{n_1} = -\dot{n_2}$. Furthermore, if we multiply the second equation by the charge $2e$ of a single Cooper pair, we can derive the supercurrent flowing from left to right across the junction:

$$I = 2e\frac{dn_2}{dt} = \frac{4eT\sqrt{n_1n_2}}{\hbar}\sin(\theta_2-\theta_1) \tag{2.12}$$

By defining $I_c \equiv 4eT\sqrt{n_1n_2}/\hbar$ to be the maximum supercurrent (i.e. critical current) that the junction can sustain, with T the characteristic number of a particular junction, and $\delta = \theta_2 - \theta_1$ the phase difference across the junction[1], we arrive at Josephson's current-phase relationship:

$$\boxed{I = I_c\sin\delta.} \tag{2.13}$$

This is the DC Josephson effect, and is equivalent to the observations made from Eq. 2.1

To understand the real solutions (Eq. 2.10), we first note that only the relative phase difference δ is important, not the overall phase. Thus, without loss of generality we can define

[1]This is not rigorously correct. For a phase difference $\Delta\theta = \theta_2 - \theta_1$ across the junction, δ is actually the *gauge-invariant* phase difference. Since the current density in a superconductor (an observable) must be gauge invariant in the presence of a magnetic field, but depends on the vector potential $\mathbf{A}$, it can be shown that the relationship between $\Delta\theta$ and δ is given by $\delta = \Delta\theta - \frac{2\pi}{\Phi_0}\int_1^2 \mathbf{A}\cdot d\mathbf{s}$, where $\Phi_0 = h/2e$ is the flux quantum. See Ref. [19] for a full discussion.

$\theta_2 = \delta/2$ and $\theta_1 = -\delta/2$. Inputting these definitions and subtracting the first equation from the second, we find

$$\frac{\hbar}{2}\frac{d\delta}{dt} = \frac{\hbar}{2}\omega = eV. \tag{2.14}$$

Thus, we have arrived at the AC Josephson effect mentioned in the beginning of the section:

$$\boxed{\omega = \frac{2eV}{\hbar}.} \tag{2.15}$$

Equations 2.13 and 2.15 are Josephson's relations governing the physics of JJs, and from these equations a plethora of interesting physics can be derived. We will not go into much more detail about the physics of JJs, but it is worth mentioning that the observations made by Josephson listed in the beginning of this section can be understood directly from these relations. Equation 2.13 states that even in the absence of any externally applied voltage, a nonzero supercurrent $-I_c \leq I \leq I_c$ can flow across the junction as long as the phase different between the two electrodes is not zero. If, on the other hand, we apply a voltage V across the junction, then by Eq. 2.15 the phase difference increases linearly in time, but the supercurrent will oscillate rapidly due to the sine term, $\sin(\delta) = \sin\left(\delta_0 + \frac{2e}{\hbar}Vt\right)$, and will thus average to zero over time. Therefore, the phases of macroscopic superconductors have real physical consequences! These effects were observed experimentally not long after they were predicted by Josephson [31].

Interestingly, because the AC Josephson effect is defined in terms of fundamental constants, it has been used to define the voltage standard [44]. By driving JJs using precise radio frequency instruments such that the AC frequency ω_{AC} is known to good precision, the resulting DC voltage can be measured at integer n steps of

$$V_{\text{DC}} = \left(\frac{\hbar\omega_{\text{AC}}}{2e}\right) n. \tag{2.16}$$

The idea then is to place many thousands JJs in series and measure the precise voltage intervals for a given applied frequency. For example, apply a frequency of 1 GHz would result in DC voltages at integral multiples of approximately 2 μV. This is known as the inverse AC Josephson effect, and was first observed by Shapiro in 1963 [45]. Thus, JJs can act as perfect frequency-to-voltage converters.

Josephson Inductance & Energy

The phase difference across a JJ leads to a non-linear inductance, which itself is associated with a coupling energy between the electrodes. The inductance L_J associated with a JJ is not a geometric quantity; rather, it is related to the inertia of Cooper pairs tunneling across the barrier. To calculate the Josephson inductance, we can use the fact that voltage and inductance are related by $V = L\frac{dI}{dt}$. Using Eq. 2.15, we can write the voltage as

$$V = \frac{\hbar}{2e}\frac{d\delta}{dt} = \frac{\Phi_0}{2\pi}\frac{d\delta}{dI}\frac{dI}{dt} = \frac{\Phi_0}{2\pi}\frac{1}{I_c\cos\delta}\frac{dI}{dt}, \tag{2.17}$$

where in the second equality we have used the definition of the *flux quantum* $\Phi_0 \equiv h/2e$ and the chain rule, and in the third equality we have taken the derivative of Eq. 2.13. Thus, we can identify the Josephson inductance as

$$\boxed{L_J(\delta) = \frac{L_{J0}}{\cos\delta}} \tag{2.18}$$

where $L_{J0} \equiv \frac{\Phi_0}{2\pi I_c}$ is known as the Josephson effective inductance. It is interesting to note that the Josephson inductance can be negative or become infinite, depending on the phase difference δ.

Because the Josephson inductance depends on δ, there must be a connection between the magnetic flux and the phase across a junction. To see how the two are related, we introduce the *branch flux* of an electrical element

$$\Phi(t) = \int_{-\infty}^{t} V(t')dt', \tag{2.19}$$

where $V(t)$ is the time-dependent voltage across the circuit element. This integral assumes that the voltage is zero at $t = -\infty$. $\Phi(t)$ can be defined for any electrical component with two leads, and from this we can derive the *branch flux-phase* relationship. If we assume that $\delta = 0$ at time $t = -\infty$, and that the junction accumulates a phase difference δ at time t, then the branch flux at time t is

$$\Phi(t) = \int_{-\infty}^{t} \left(\frac{\hbar}{2e}\frac{d\delta'}{dt'}\right)dt' = \frac{\hbar}{2e}\int_0^{\delta} d\delta' = \frac{\Phi_0}{2\pi}\delta. \tag{2.20}$$

The branch flux-phase relationship is often written as $\Phi = \phi_0\delta$, where $\phi_0 \equiv \Phi_0/2\pi$ is called the *reduced flux quantum*.

The inertia of Cooper pairs tunneling across the barrier results in the buildup of magnetic flux density at the junction. For a more intuitive understanding, we can re-write the current-phase relation (Eq. 2.13) using the branch flux-phase relationship:

$$I(t) = I_c \sin\left(\frac{2\pi\Phi(t)}{\Phi_0}\right). \tag{2.21}$$

While Eqs. 2.13 and 2.15 are often referred to as Josephson's relations, Eq. 2.21 is sometimes called the constitutive relation [46], as it contains the same information as Josephson's relations in only one equation. As we can see, for small branch flux Φ, the junction behaves like a standard inductor, with $I(t) \approx \Phi(t)/L_J$ to first order.

When supercurrent flows across the barrier, magnetic flux accumulates at the junction due to its non-linear inductance. This results in the development of a potential energy across the junction. Using Eqs. 2.13 and 2.15, we can derive the potential energy by integrating the power delivered to the junction due the applied voltage V and the current flowing through the junction I:

$$U = \int_{-\infty}^{t} IV\,dt' = \int_{-\infty}^{t} (I_c\sin\delta')\left(\frac{\hbar}{2e}\frac{d\delta'}{dt'}\right)dt' = \frac{\hbar I_c}{2e}\int_0^{\delta}\sin\delta' d\delta' = E_J(1-\cos\delta), \tag{2.22}$$

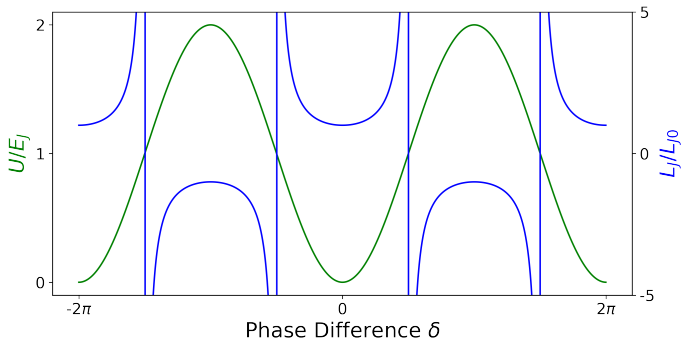


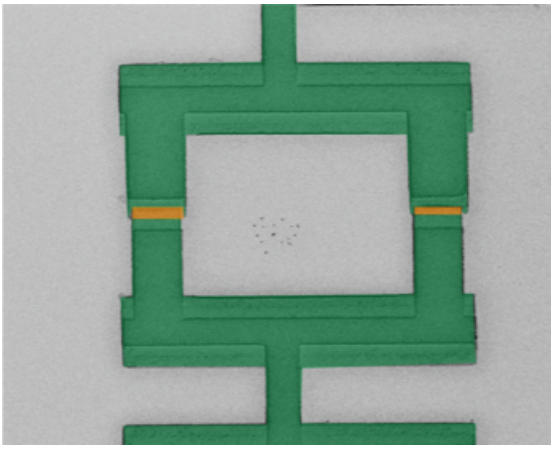
Figure 2.3: **Josephson energy and inductance.** The Josephson energy $U/E_J = 1 - \cos\delta$ is plotted above in green and the Josephson inductance $L_J/L_J = 1/\cos\delta$ is plotted above in blue versus the phase difference δ. The potential energy is a minimum when $\delta = 2\pi n$, where n is any integer, and is often described as having the shape of a "cosine washboard." The inductance is highly non-linear and is finite at any phase $\delta \neq \frac{\pi}{2}m$, where m is any odd integer. As we can see from the above graph, the Josephson inductance can become negative or even infinite.

where $E_J \equiv \frac{\hbar I_c}{2e}$ is the characteristic energy scale of the junction, and can be written in terms of the flux quantum Φ_0: $E_J = \frac{\Phi_0 I_c}{2\pi}$. This energy is often written without the constant term:

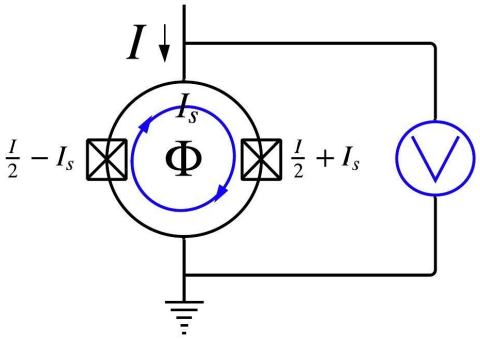
$$\boxed{U(\delta) = -E_J \cos\delta}. \tag{2.23}$$

$U(\delta)$ is the Josephson (potential) energy of the junction, and is clearly a minimum when $\delta = 0$ or any integer of 2π; see Fig. 2.3. In other words, the energy is minimized when the phases are equal, or rather, in the absence of phase gradients across the barrier. Additionally, we can see that the strength of the coupling energy depends on the magnitude of the critical current, which itself is a function of the type of material used for the barrier, its thickness and cross-sectional area, and the type of weak link employed. As we will see in the next section, the sinusoidal form of the Josephson energy allows us to create an anharmonic oscillator out of a standard LC circuit, which is the basis of the superconducting qubit.

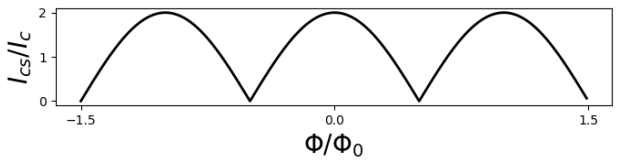
There is one more important JJ-based circuit element that must be covered before we can understand how they are the fundamental component of superconducting qubits. When JJs are placed around a superconducting loop, this is known as a Superconducting QUantum Interference Device (SQUID) [33, 34] (see Fig. 2.4a). There are two main types of SQUIDs: the DC SQUID, which contains parallel JJs, and the RF SQUID, which is a loop interrupted by only a single JJ. In a DC SQUID, the total input current I is split equally between the two branches. When an external magnetic field is applied to the loop, the induced back-EMF around the loop establishes a circulating "screening" current I_s to cancel the applied magnetic field, reducing the total current in one branch and increasing the total current in the other: $I_1 = I/2 - I_s$ and $I_2 = I/2 + I_s$. When I_1 or I_2 exceeds the critical current I_c



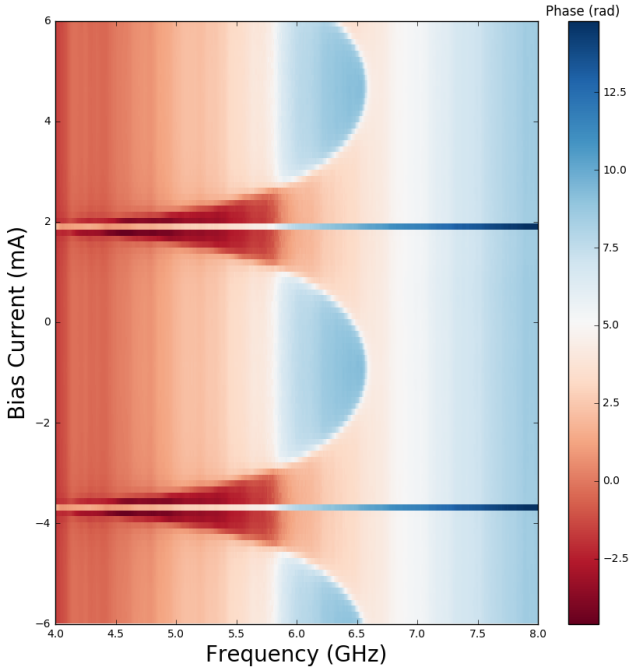
(a) SEM image of a SQUID



(b) Circuit schematic of a SQUID



(c) Critical Current vs. Applied Flux for SQUID



(d) Flux-tuning data from a DC SQUID

Figure 2.4: **Superconducting QUantum Interference Device.** (a) A false-colored SEM image of a DC SQUID, which is composed of two JJs in parallel around a loop. (b) The electrical circuit element of a DC SQUID is denoted by a loop with a JJ on both sides. An external flux Φ applied through the SQUID establishes a screening current I_s around the loop to cancel the field. When the bias current through one of the junctions plus the screening current, $I/2+I_s$, exceeds the critical current I_c of the junction, a voltage V can be measured. (c) Critical current vs. Applied flux for a DC SQUID, given by Eq. 2.24. (d) Experimental flux-tuning data for a DC SQUID in a Josephson parametric amplifier (JPA). The applied flux is recorded as the bias current of an external DC power source. The x-axis is the output frequency of a Vector Network Analyzer (VNA). The pixel intensity at each location is the unwrapped phase difference across the SQUID measured by the VNA. A sharp phase shift indicates the location of the resonant frequency of the SQUID. As the external flux is tuned through one flux quantum Φ/Φ_0, the resonant frequency shifts from low to high, and then back to low again; this is equivalent to changing the critical current I_{cs} of the SQUID as a function of applied flux Φ (Eq. 2.24). The offset of the maximum resonant frequency from zero bias current is most likely due to trapped magnetic flux in the system, as SQUIDs are extremely susceptible to the magnetic fields from any parasitic circulating currents.

of its respective junction, a voltage difference will appear across the SQUID (see Fig. 2.4b); therefore, SQUIDs can act as a flux-to-voltage converter. The energy in the SQUID loop is minimized when the flux enclosed is an integer multiple of the flux quantum Φ_0. As the external flux is tuned upward through half a flux quantum, $\Phi_0/2$, the screening current will reverse direction in order to encourage a total enclosed flux Φ_0 through the loop. The current will then reverse direction again when the applied flux is tuned through Φ_0, and will continue to do so periodically every half-integer multiple of a flux quantum. In the limit in which the geometric loop inductance is much smaller than the Josephson inductances of the individual junctions, we can write the critical current I_{cs} of the SQUID as a function of the applied flux Φ [34]:

$$I_{cs}(\Phi) = 2I_c\left|\cos\left(\frac{\pi\Phi}{\Phi_0}\right)\right|, \tag{2.24}$$

where I_c is the critical current of each Josephson junction, which we take to be identical. As we can see, the DC SQUID behaves as a single Josephson junction, but with a flux-tunable critical current I_{cs}.

2.2 Quantization of LC Circuits

The behavior of a quantum mechanical circuit is determined by its Hamiltonian. Consider a standard LC circuit, as shown on the left in Fig. 2.5a. The mechanical energy stored in a capacitor and inductor is $E_C = \frac{1}{2}CV^2$ and $E_L = \frac{1}{2}LI^2$, respectively. We are justified in ignoring any resistance in this circuit, as the components are made of superconductors. For the LC circuit in Fig. 2.5a, we choose the voltages across both elements to be equal: $V_C = V_L = V$, where $Q = CV_C$ and $V_L = L\frac{dI}{dt} = L\dot{I}$. By defining the voltages to be equal and taking the positive current to be in the clockwise direction through the inductor, we must therefore adopt the convention that the current through the capacitor is given by $I = -\frac{dQ}{dt} = -\dot{Q}$.

Using the classical equations that define capacitance $Q = CV$ and inductance $\Phi = LI$, we can write the Hamiltonian of the system as

$$H = \frac{1}{2}CV^2 + \frac{1}{2}LI^2 = \frac{Q^2}{2C} + \frac{\Phi^2}{2L}. \tag{2.25}$$

In order to map classical observables into quantum mechanical operators, it is necessary to find the canonically conjugate pair of coordinates, q_i and p_i, defined by Hamilton's equations:

$$\frac{\partial H}{\partial p_i} = \dot{q}_i, \quad \frac{\partial H}{\partial q_i} = -\dot{p}_i. \tag{2.26}$$

If we consider the partial derivative of our Hamiltonian with respect to Q and Φ, we find

$$\frac{\partial H}{\partial Q} = \frac{Q}{C} = V = L\dot{I} = \dot{\Phi}, \tag{2.27}$$

$$\frac{\partial H}{\partial \Phi} = \frac{\Phi}{L} = I = -\dot{Q}. \tag{2.28}$$

Thus, we have shown that Q and Φ are canonically conjugate variables for our Hamiltonian, which map to the classical coordinates p and q, respectively. By equating the capacitance with the mechanical analog of a mass m, and noting that the resonant frequency of an LC circuit is $\omega = 1/\sqrt{LC}$, we can see that our Hamiltonian is simply the classical Hamiltonian for a harmonic oscillator,

$$H = \frac{Q^2}{2C} + \frac{\Phi^2}{2L} = \frac{p^2}{2m} + \frac{1}{2}m\omega^2 x^2, \tag{2.29}$$

where x is the spatial coordinate and p the momentum. Therefore, an LC circuit is just a simple harmonic oscillator with resonant frequency $\omega = 1/\sqrt{LC}$.

In order to quantize our Hamiltonian, we map our classical observables into quantum operators, $Q \to \hat{Q}$ and $\Phi \to \hat{\Phi}$, which obey the canonical commutation relation, $[\hat{\Phi}, \hat{Q}] = i\hbar$. Next, we define $\hat{Q}$ and $\hat{\Phi}$ in terms of creation and annihilation operators,

$$\hat{Q} = -iQ_{\text{ZPF}}\left(a - a^\dagger\right), \tag{2.30}$$

$$\hat{\Phi} = \Phi_{\text{ZPF}}\left(a + a^\dagger\right), \tag{2.31}$$

where

$$Q_{\text{ZPF}} = \sqrt{\frac{\hbar}{2Z}}, \tag{2.32}$$

$$\Phi_{\text{ZPF}} = \sqrt{\frac{\hbar Z}{2}} \tag{2.33}$$

are the zero-point fluctuations of the charge and phase variables, respectively, and $Z = \sqrt{L/C}$ is the complex impedance of the LC circuit. Using these definitions, we can define creation and annihilation operators

$$\hat{a}^\dagger \equiv \frac{1}{\sqrt{2\hbar Z}}\left(\hat{\Phi} - iZ\hat{Q}\right), \tag{2.34}$$

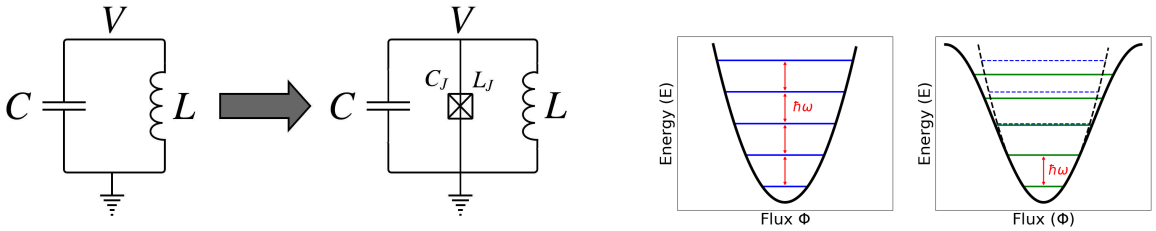
$$\hat{a} \equiv \frac{1}{\sqrt{2\hbar Z}}\left(\hat{\Phi} + iZ\hat{Q}\right). \tag{2.35}$$

Thus, we can write the Hamiltonian as a quantum harmonic oscillator (QHO),

$$\hat{H} = \frac{\hat{Q}^2}{2C} + \frac{\hat{\Phi}^2}{2L} = \hbar\omega\left(\hat{a}^\dagger\hat{a} + \frac{1}{2}\right) = \hbar\omega\left(\hat{n} + \frac{1}{2}\right), \tag{2.36}$$

where $\hat{n}$ is the number operator acting on a Fock state $|n\rangle$, which defines the number of photons n at frequency ω. As such, the energy of a quantized LC circuit is simply the number of photons at the resonant frequency, plus the zero-point contribution $\frac{1}{2}\hbar\omega$.

A quantized LC circuit is nothing more than a QHO. At low temperatures and high resonant frequencies, such that $k_B T \ll \hbar\omega$, the circuit will primarily be in its ground state



(a) Josephson junction in an LC circuit

(b) Harmonic vs. Anharmonic potential

Figure 2.5: **Harmonic vs. Anharmonic potential.** (a) (Left) A standard quantized LC circuit is not sufficient for creating a qubit. (Right) In order to create an isolated two-state quantum system out of an LC circuit, we must add a JJ, which is a lossless nonlinear circuit element with an intrinsic capacitance and phase-dependent inductance, C_J and $L_J(\delta)$. (b) (Left) The energy spacing of a quantum harmonic oscillator is degenerate for all pairs of energy levels. Thus, a photon of energy $\hbar\omega$ is equally likely to drive a transition upward as it is to drive a transition downward. This makes a standard harmonic oscillator unsuitable for creating qubits. (Right) By adding a JJ to the LC circuit, we can change the shape of the potential well due to the sinusoidal dependence on the phase (or flux, since $\Phi = \frac{\Phi_0}{2\pi}\delta$). The anharmonicity of the sinusoidal potential shifts the excited energy levels down with respect to the parabolic potential. Thus, the energy spacings are no longer degenerate, allowing us to drive a transition between the ground and first excited states without exciting the higher levels.

$|0\rangle$. Statistically, there will be a small but non-zero chance of measuring it in an excited state, but this is exponentially suppressed by a Boltzmann factor $\exp(-h\omega/k_{\mathrm{B}}T)$. By injecting a photon of energy $\hbar\omega$ into the circuit, we can excite the QHO into its first excited state $|1\rangle$. If we could resonantly drive the circuit between its ground state and first excited state, this would constitute a quantum bit. However, upon injecting a second photon into the circuit, due to the equal spacing of the energy levels in a parabolic potential, the QHO will have an equal probability of de-excitation into the ground state as being promoted into the second excited state $|2\rangle$ (see Fig. 2.5b). Thus, a simple QHO does not allow for control over a two-state quantum system due to its harmonic potential, because photons of energy $\hbar\omega$ can drive transitions between any two of its energy levels. For an isolated two-state quantum system, we need to add a Josephson junction.

Recall that the JJ is a lossless nonlinear circuit element with a potential energy given by $U = -E_J \cos\delta$. This potential well has a sinusoidal dependence on the phase δ, but for small δ it approximates a parabola: $U = E_J\frac{\delta^2}{2} - \mathcal{O}(\delta^4)$. Thus, at the bottom of the well U is approximately parabolic, but as you move higher up the walls begin to broaden into a sinusoidal potential. This broadening causes the higher energy levels to shift downward, creating an anharmonic oscillator (AHO) in which the spacing between any set of two energy levels is different (see Fig. 2.5b). Therefore, by adding a JJ in parallel with the existing

LC circuit, we can change the energy levels such that the spacing between the ground and excited states is non-degenerate with the rest of the spectrum. In this way, we can drive an isolated transition between the two quantum states using microwave photons, thus creating a quantum bit. By including a JJ, we have effectively created an artificial atom.

After adding the JJ to our LC circuit, we must consider the Hamiltonian for the entire system. By defining the charge operator on the capacitor in terms of the number operator of Cooper pairs, $\hat{Q} = (2e)\hat{n}$, we can define the characteristic energy scale for a capacitor C: $E_C = (e)^2/2C$.[2] Similarly, by equating the energy stored in the inductor with the branch flux through the element (Eq. 2.20), we can define the characteristic energy scale of an inductor L: $E_L = (\Phi_0/2\pi L)^2$. Thus, the Hamiltonian for the circuit on the right in Fig. 2.5a is

$$\hat{H} = \frac{\hat{Q}^2}{2C} + \frac{\hat{\Phi}^2}{2L} + U = 4E_C\hat{n}^2 + \frac{1}{2}E_L\hat{\delta}^2 - E_J\cos\hat{\delta}. \tag{2.37}$$

As is written above, $\hat{\delta}$ and $\hat{n}$ now constitute the canonically conjugate variables of our Hamiltonian, with a commutation relation given by $[\hat{\delta}, \hat{n}] = i$.

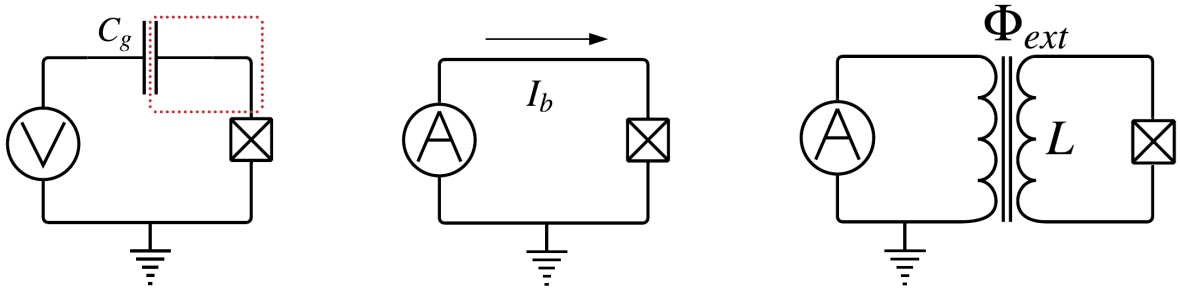
2.3 Superconducting Qubits

The basis of superconducting qubits is the quantization of electrical LC circuits at low temperatures. However, in order for LC circuits to behave quantum mechanically and maintain quantum coherence, there must be little to no dissipative energy loss in the system. Naturally, superconductors are ideal candidates for for this purpose. In addition to the zero resistance requirement, we must be able to control the state of the qubit (i.e. drive it resonantly between two quantum states, say $|0\rangle$ and $|1\rangle$, and make a measurement). As we saw in the previous section, the Josephson junction allows us to create an artificial atom out of a resonant LC circuit, in which we can drive the qubit between its ground state and its first excited state using microwave photons. Given that JJs can be fabricated in large numbers using standard integrated-circuit microfabrication techniques, this allows for tunability in the properties of a qubit. As we will see below, this tunability enables the creation of several different "flavors" of superconducting qubits.

The Charge Qubit

The charge qubit [47, 48, 49, 50] is based off of the circuit diagram in Fig. 2.6a, in which a JJ is biased by a voltage source V through a series "gate" capacitance C_g. The portion of the circuit between C_g and the JJ is called a *Cooper pair box*: a small superconducting island capacitively isolated from the rest of the circuit in which the number n of Cooper pairs on the island is discrete. Both the charge island and the charge reservoir are assumed

[2] E_C does not contain the factor of 4 due to the charge carrier being a Cooper pair (not an electron) for historical reasons.



(a) Charge qubit (b) Phase qubit (c) Flux qubit

Figure 2.6: **Superconducting qubits.** (a) Circuit diagram of a charge qubit. The JJ is voltage-biased through a gate capacitor C_g. The portion highlighted in red is called the Cooper pair box. (b) Circuit diagram of a phase qubit. The JJ is current-biased using current source. (c) Circuit diagram of a flux qubit. A loop containing a single JJ is flux biased using a current source.

to be good BCS superconductors, such that the energy gap Δ exceeds the $k_B T$ thermal noise of the system. The Hamiltonian for a charge qubit is

$$\hat{H} = 4E_C(\hat{n} - n_g)^2 - E_J \cos\hat{\delta}, \tag{2.38}$$

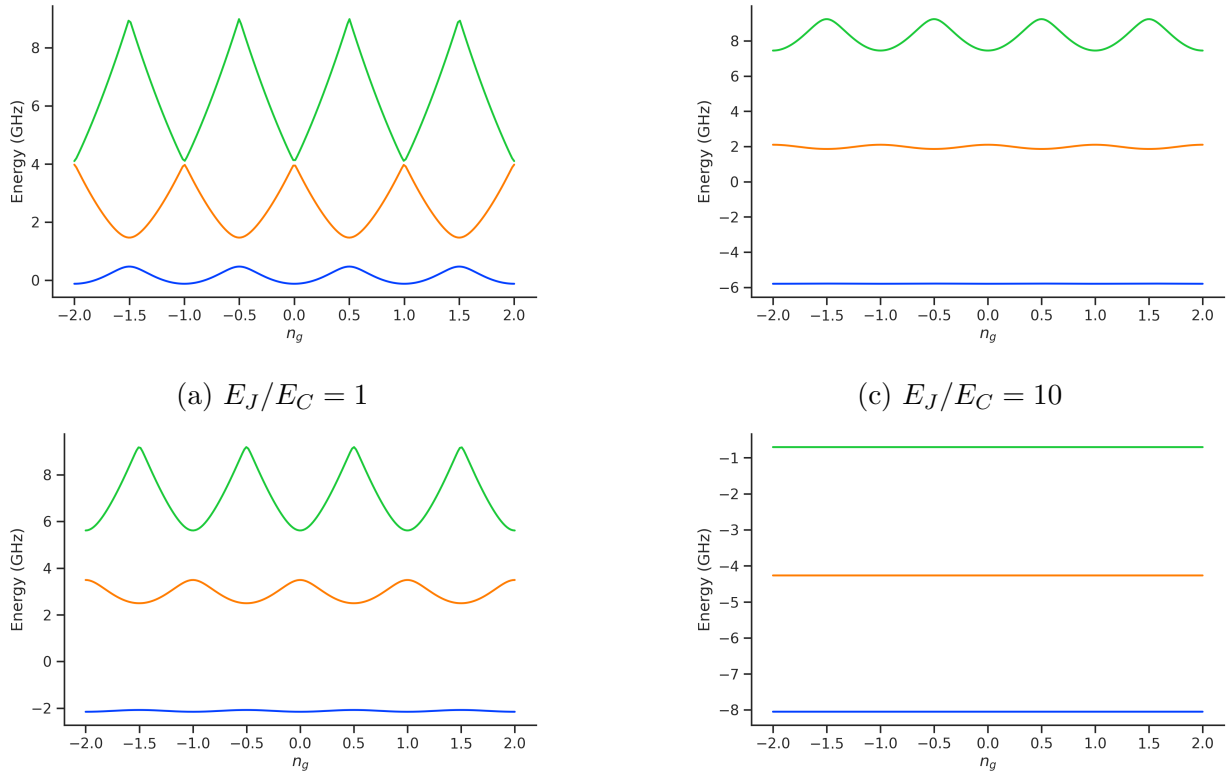
where we have defined the charging energy with respect to the voltage-induced charge on the gate capacitor, $V = (2e)n_g/C_g$, and $E_C = e^2/2(C_g + C_J)$. While $\hat{n}$ is constrained to be an integer, n_g is a continuous variable determined by the voltage and gate capacitance. In fact, it is this discreteness in $\hat{n}$ (i.e. the integer number of Cooper pairs tunneling through the JJ) which causes the periodic dependence on $\hat{\delta}$ in the Josephson energy term.

To find the energy eigenvalues of this Hamiltonian, we must first solve the following Schrödinger equation:

$$\hat{H}\Psi = \left[4E_C\left(-i\frac{\partial}{\partial\hat{\delta}} - n_g\right)^2 - E_J\cos\hat{\delta}\right]\Psi = E\Psi, \tag{2.39}$$

where we have defined $\hat{n} = -i\frac{\partial}{\partial\hat{\delta}}$ in the phase-basis representation. While this equation can be solved exactly, doing so requires special functions, which is beyond the scope of this thesis. Nevertheless, it is useful to consider the energy spectrum with respect to n_g, as shown in Fig. 2.7.

Because $\hat{n}$ must be an integer, we can choose adjacent energy eigenstates in the Cooper pair number basis, $|n\rangle$ and $|n+1\rangle$, to represent our two-state quantum system. Thus, we can write our qubit as an linear combination of these two states with arbitrary complex coefficients α and β: $|\psi\rangle = \alpha\,|n\rangle + \beta\,|n+1\rangle$. On the other hand, because n_g is continuous,



(a) $E_J/E_C = 1$

(c) $E_J/E_C = 10$

(b) $E_J/E_C = 5$

(d) $E_J/E_C = 50$

Figure 2.7: **Charge qubit energy spectrum and charge dispersion.** Energy spectrum as a function of n_g for different ratios of E_J/E_C for the first three eigenenergies of the charge qubit (Eq. 2.38). (a) $E_J/h = 1$ GHz, $E_C/h = 1$ GHz. (b) $E_J/h = 5$ GHz, $E_C/h = 1$ GHz. (c) $E_J/h = 10$ GHz, $E_C/h = 1$ GHz. (d) $E_J/h = 10$ GHz, $E_C/h = 0.2$ GHz. We can see that for $n_g = 1/2$, the first two energy levels in (a) are nearly degenerate. In fact, in the absence of the Josephson potential, the energy spectrum would simply be quadratic in n and the two points would touch. However, these degeneracy points are lifted by the Josephson energy term, with the difference given by $\sim E_J$. As we increase the ratio E_J/E_C, the energy bands become flatter with respect to n_g, which reduces the Cooper pair box's sensitivity to charge noise at the expense of reduced anharmonicity. For large E_J/E_C, as in (d), the energy bands are essentially flat; this is known as the transmon regime. (The charge qubit eigenenergies were calculated using the `scQubits` Python package [51].)

we can use it as a parameter to tune how close together our basis states are (see Fig. 2.7). Typically, a voltage V is chosen such that $n_g = n + 1/2$. This is known as the charge degeneracy point (or the charge "sweet spot"), corresponding to half a Cooper pair of charge for both basis states. In other words, we can use n_g to change where the qubit operates within the energy spectrum. $|n\rangle$ and $|n+1\rangle$ would be fully degenerate at $n_g = 1/2$ if it were not for the Josephson coupling term $U(\hat{\delta})$. It is called a coupling term because, in the Cooper pair number basis, we can write the Hamiltonian for $U(\hat{\delta})$ as [49]

$$H_J = -\frac{E_J}{2}\sum_n (|n\rangle\langle n+1| + |n+1\rangle\langle n|). \tag{2.40}$$

The JJ forces Cooper pairs to tunnel to and from the charge island one by one, connecting the energy eigenstates and creating a separation between the two energy levels given by E_J. For small E_J/E_C, the separation between the energy levels is extremely dependent on the value of n_g. Small variations in n_g can cause large variations in the energy difference between $|n\rangle$ and $|n+1\rangle$; this is known as charge noise, as it depends on the charge induced by the voltage source. However, by tuning the ratio of E_J/E_C, it is possible to make the qubit less sensitive to variations in n_g near the charge degeneracy point. As E_J/E_C is increased, the bands become flatter, reducing the qubit's sensitivity to charge noise at the expense of reduced anharmonicity of the potential energy (i.e. the difference between subsequent pairs of energy levels). A popular variant of the charge qubit, known as the *transmon* [42], takes advantage of this phenomenon by fabricating the qubit such that $E_J \gg E_C$; we discuss this more below.

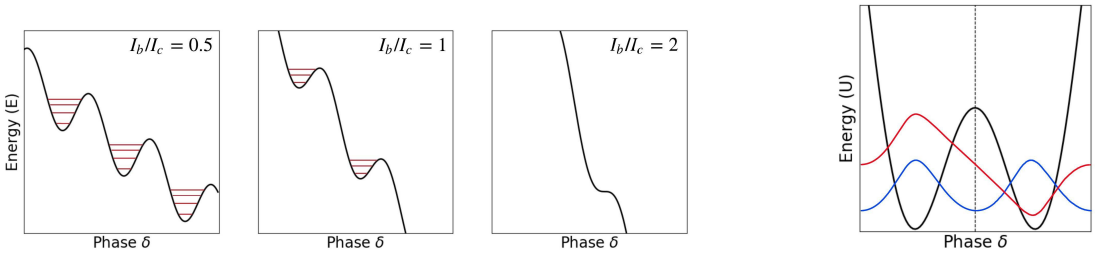
It is also necessary to read out the state of the qubit. With the charge qubit, this can be done with extremely sensitive electrometers, such as a single-electron transistor[52], with the requirement that the accuracy of the instrument be much greater than $2e$. For increased tunability of the charge qubit, the single JJ can be replaced by a DC SQUID, which allows the experimenter to modify E_J by changing the critical current of the SQUID with a DC flux bias.

The Phase Qubit

The phase qubit [53] is composed of a single current-biased Josephson junction (Fig. 2.6b). The energy associated with the bias current I_b can be calculated by integrating out the power $P = I_b V = I_b\left(\frac{\hbar}{2e}\frac{d\delta}{dt}\right)$, as we did for the Josephson energy in Eq. 2.22. The full Hamiltonian is given by

$$\hat{H} = 4E_{C_J}\hat{n}^2 - E_J\cos\hat{\delta} - \frac{\Phi_0}{2\pi}I_b\hat{\delta}, \tag{2.41}$$

where E_{C_J} is the charging energy associated with the intrinsic Josephson capacitance. The second and third terms in the Hamiltonian represent the potential energy of the system. Using the relation $E_J = \frac{\Phi_0 I_c}{2\pi}$ derived in Section 2.1, we can re-write the potential energy in



(a) Phase qubit potential

(b) Flux qubit potential

Figure 2.8: **Phase and flux qubit potentials.** (a) Energy spectrum of the phase qubit as a function of δ for different ratios of I_b/I_c. This is known as the tilted washboard potential. (Note: the energy levels depicted in the figures above are not exact; they are qualitatively right, but are used for illustrative purposes only.) (b) Energy spectrum of the flux qubit as a function of δ with the external flux through the RF SQUID loop tuned to half a flux quantum, $\Phi_{ext} = \Phi_0/2$, and $E_J/E_L = 10$. The ground (blue) and excited (red) states are taken to be the symmetric and anti-symmetric wavefunctions of the double-well potential, respectively. Each wavefunction is a superposition of magnetic flux pointing up $|\uparrow\rangle$ and down $|\downarrow\rangle$ through the loop, denoted by the bound states at the bottom of each well.

terms of the Josephson energy and the junction critical current as a function of the phase $\hat{\delta}$:

$$U(\hat{\delta}) = -E_J \cos\hat{\delta} - E_J \frac{I_b}{I_c}\hat{\delta}. \tag{2.42}$$

Thus, we can see that the potential energy in the Hamiltonian has the usual periodic Josephson term, modulated by a term that is linear in $\hat{\delta}$. Together, these two terms form what is known as a "tilted washboard" potential, whose shape depends critically on the ratio I_b/I_c; see Fig. 2.8a. As I_b/I_c increases, the height of the potential well decreases, and as a consequence, so do the energy spacings in the well. As the potential becomes shallower, the probability that a Cooper pair in an excited state $|n\rangle$ can tunnel through the wall into the next well increases exponentially as n increases. In fact, the shape of each potential well can be accurately approximated by a cubic potential. However, when I_b becomes larger than I_c, the shape of the potential approaches that of a straight line, at which point there are no longer any bound states.

A qubit is created out of a single current-biased JJ by driving transitions between the ground state $|0\rangle$ and first excited state $|1\rangle$. One advantage of the phase qubit is that the readout-circuitry is already built into the architecture of the qubit. By biasing the junction such that the second excited state is nearly at the top of the potential well, any excitation into $|2\rangle$ is exponentially likely to tunnel out of the potential barrier and down the washboard into the next well. A measurement of the state of a phase qubit is carried out by applying

a microwave pulse at frequency $\omega_{12} = (E_2 - E_1)/\hbar$ to drive a transition from $|1\rangle$ to $|2\rangle$. If the qubit is in the ground state $|0\rangle$, it will not be excited to $|2\rangle$ and nothing will happen. If, however, the qubit is in the first excited state $|1\rangle$, the pulse will excite a transition to state $|2\rangle$, causing a tunneling event to occur through the JJ, which will result in the appearance of a finite voltage across the junction. Thus, measuring no voltage after applying the probe pulse implies that the qubit was in state $|0\rangle$, whereas the appearance of a nonzero voltage indicates that the qubit was in state $|1\rangle$.

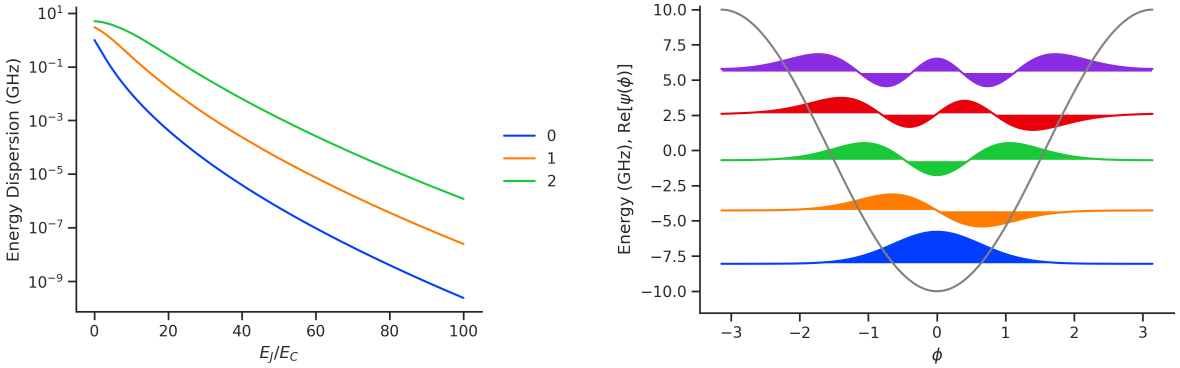
The Flux Qubit

The basis of a flux qubit [54] is an RF SQUID — a superconducting loop interrupted by a single Josephson junction. The Josephson loop is inductively coupled to a current source, such that an external flux Φ_{ext} can bias the state of the SQUID. The Hamiltonian is given by

$$\hat{H} = 4E_{C_J}\hat{n}^2 + \frac{1}{2}E_L\hat{\delta}^2 - E_J\cos\left(\hat{\delta} - \frac{2\pi}{\Phi_0}\Phi_{ext}\right), \tag{2.43}$$

where the second term is the energy associated with the magnetic flux stored in the inductive loop L. We can understand the third term by recalling that the flux across a circuit element (the branch flux) is simply associated with a phase shift for a JJ (see Eq. 2.20), so the Josephson energy will acquire a phase shift that depends on the external flux. Since the last two terms constitute the potential energy of the system, we can see that it is composed of a term quadratic in phase, modulated by the usual periodic Josephson term, whose phase now depends on the external flux. As discussed in Section 2.1, a SQUID energetically prefers an enclosed flux that is an integer number of flux quanta. Thus, by tuning the external magnetic flux to half a flux quantum, the SQUID will acquire counter-circulating currents to encourage a total enclosed flux of 0 or Φ_0. Therefore, for $\Phi_{ext} = \Phi_0/2$ the SQUID will have flux pointing both upward and downward through the loop. At this degeneracy point, the qubit itself is a superposition of supercurrents circulating clockwise and anticlockwise around the loop, or, equivalently, a superposition of magnetic flux pointing up and pointing down: $|\psi\rangle = \alpha\,|\uparrow\rangle \pm \beta\,|\downarrow\rangle$. Plotted versus the phase in Fig. 2.8b, we can see that $U(\delta)$ is a quadratic potential with a cosine corrugation in the middle, creating a double-well potential. The ground and excited states are simply the symmetric and anti-symmetric superpositions of the two basis states: $|0\rangle = \frac{1}{\sqrt{2}}(|\uparrow\rangle + |\downarrow\rangle)$ and $|1\rangle = \frac{1}{\sqrt{2}}(|\uparrow\rangle - |\downarrow\rangle)$.

By sending current pulses to the RF SQUID through an inductive coupler, we can determine whether the qubit is in the $|0\rangle$ or $|1\rangle$ state. If the current pulse results in no change in voltage across the JJ, the qubit is in the $|0\rangle$ state. However, if the current pulse results in a non-zero voltage across the junction, the qubit has made a transition to the excited state. A measurement of the voltage state of a flux qubit is typically made with a DC SQUID aligned co-axially with the RF SQUID. By also biasing the DC SQUID with a flux $\Phi_0/2$, the critical current I_{cs} of the SQUID will be zero (see Eq. 2.24). Thus, by monitoring the critical current of the DC SQUID, any change in flux produced by the nearby qubit can be



(a) Energy dispersion as a function of E_J/E_C (b) Wavefunctions in cosine potential

Figure 2.9: **Transmon qubit energy dispersion and wavefunctions.** (a) The dispersion in the energy levels decreases exponentially with E_J/E_C, flattening the energy bands (as seen in Fig. 2.7d). (b) The wavefunctions for the first five energy levels of the transmon qubit in a cosine potential in the flux-basis representation ($E_J/h = 10$ GHz, $E_C/h = 0.2$ GHz). The separation between adjacent energy levels is distinct for each pair of eigenenergies; however, the reduced anharmonicity of the transmon means that a transmon qubit is more likely to be excited to higher energy levels (i.e. "leak" out of the first excited state).

measured with high accuracy. Due to the high sensitivities of SQUIDs to magnetic flux, flux qubits are extremely susceptible to any stray magnetic fields.

The Transmon Qubit

The charge qubit, phase qubit, and flux qubit represent the three prototypical types of superconducting qubits. However, modern superconducting quantum processors utilize qubits which have been developed from these three main archetypes to be much more intrinsically robust to noise. For example, the charge qubit's sensitivity to charge noise (i.e. fluctuations in n_g) lead to the development of the transmon qubit [42]. The transmon is a charge qubit that is shunted by a large capacitance C_s. The Hamiltonian of the transmon is identical to the charge qubit (Eq. 2.38), however, the charging energy $E_C = (e)^2/2(C_g + C_J + C_s)$ is made much smaller by the addition of the large shunting capacitor. The transmon regime is define to be the regime in which the Josephson energy is much larger than the charging energy ($E_J >> E_C$) such that the ratio $E_J/E_C \gtrsim \mathcal{O}(10)$ (the transmon is typically operated around $E_J/E_C \approx 50$). In fact, the dispersion in the energy levels (i.e. the variation in the energy levels with respect to n_g) decreases exponentially in E_J/E_C [42] (see Fig. 2.7).

At large E_J/E_C the transmon qubit is insensitive to charge noise and can be operated at any gate voltage (not just at the charge degeneracy point). This, however, comes at the

cost of reducing the anharmonicity α of the qubit:

$$\alpha \equiv E_{12} - E_{01}. \tag{2.44}$$

The anharmonicity defines how the transition energy between the ground $|0\rangle$ and the first $|1\rangle$ excited states (E_{01}) differs from the transition energy between the first $|1\rangle$ and second $|2\rangle$ excited states (E_{12}). For the charge qubit at the sweet spot, the anharmonicity is positive and large (see Fig. 2.7a), meaning that the energy levels defining the qubit are well separated from the higher energy levels. However, the transmon has an anharmonicity that is small and negative, meaning that excitations from $|0\rangle \rightarrow |1\rangle$ are much more likely to drive a transition to $|2\rangle$; this is often referred to as "leakage," since it represents information leaking out of the basis states, which we define to be $|0\rangle$ and $|1\rangle$. In fact, the anharmonicity decreases with a slow power law in E_J/E_C [42], much slower than the exponential reduction in the charge dispersion, and therefore the benefits generally outweigh the costs. However, a smaller anharmonicity severely limits the speed with which one can reasonable drive transitions from $|0\rangle \rightarrow |1\rangle$ (e.g. via Rabi oscillations), as faster control pulses will naturally include higher harmonics, which can drive transitions from $|1\rangle \rightarrow |2\rangle$.

For transmon qubits, the $|0\rangle \rightarrow |1\rangle$ frequency is given as

$$\omega_{01} = \left(\sqrt{8E_J E_C} - E_C\right)/\hbar, \tag{2.45}$$

and the anharmonicity is $\alpha = -E_C$. Typically, ω_{01} is in the range of $4 - 6$ GHz, and α is generally designed to be around $100 - 300$ MHz.

The Qubit Hamiltonian

The insensitivity of the transmon to charge noise means that the gate charge n_g can be omitted from the qubit Hamiltonian:

$$\boxed{\hat{H}_q = 4E_C\hat{n}^2 - E_J\cos\hat{\delta}}. \tag{2.46}$$

However, it is often more convenient to represent this Hamiltonian using second quantization, as the energy diagram is most intuitively understood in terms of excitations involving quanta of discrete energy $\hbar\omega$, as opposed to thinking about the problem in terms of discrete tunneling of Cooper pairs.

As a reminder, the introduction of the JJ in the qubit Hamiltonian transforms the quadratic potential of the QHO into an AHO with a cosine potential (see Fig. 2.5b). The effect is to shift the energy levels such that they are no longer degenerate with the $|0\rangle \rightarrow |1\rangle$ transition frequency ω_{01}. We can gain some intuition by expanding the cosine potential in terms of a power series,

$$-E_J\cos\hat{\delta} = -1 + \frac{1}{2}E_J\hat{\delta}^2 - \frac{1}{24}E_J\hat{\delta}^4 + \mathcal{O}(\hat{\delta}^6), \tag{2.47}$$

where the second-order term represents the typical quadratic potential of the QHO, and the fourth (and higher) order terms represent modifications to the quadratic potential for higher-energy eigenstates. Thus, when representing the qubit Hamiltonian in terms of the second quantization approach, the AHO acquires higher order terms which define the deviation of the higher energy levels from the first excited state:

$$\hat{H} = \hbar\omega_{01}\hat{a}^\dagger\hat{a} + \frac{\hbar\alpha}{2}\hat{a}^\dagger\hat{a}^\dagger\hat{a}\hat{a}, \tag{2.48}$$

where the first term represents a typical harmonic oscillator for the $|0\rangle \rightarrow |1\rangle$ transition, and the second term is often called a fourth-order Kerr-nonlinearity, with α the qubit anharmonicity. For simplicity, we omit the zero-point energy and truncate our expansion at fourth order. Eq. 2.48 resembles that of a classical Duffing oscillator.

If we take $|\alpha|$ to be large enough, or assume that leakage to higher energy states can be sufficiently suppressed, we may effectively treat the qubit as a two-level system. In this case, we can write the Hamiltonian as

$$\boxed{\hat{H}_q = \frac{1}{2}\hbar\omega_q\sigma_z}, \tag{2.49}$$

where $\omega_q = \omega_{01}$ is the $|0\rangle \rightarrow |1\rangle$ transition frequency and σ_z the Pauli-z operator. Therefore, we see that in the limit that a qubit can be treated as an effective two-level system, the qubit Hamiltonian is exactly the same as the Hamiltonian for a spin-$\frac{1}{2}$ particle in quantum mechanics, with $|0\rangle$ and $|1\rangle$ representing the spin up and spin down states of the particle. Therefore, two-level systems are isomorphic to spin-$\frac{1}{2}$ systems, greatly simplifying our treatment of qubits.

2.4 Single-qubit Gates

In the previous section, we have shown that the non-linearity of the Josephson junction enables one to drive transitions between two distinct energy levels in an anharmonic potential, creating the basis for a class of different superconducting qubits. However, developing quantum gates for superconducting qubits requires more control than simply driving excitations from $|0\rangle \rightarrow |1\rangle$. In this section, we describe how microwave pulses can be used to perform single-qubit quantum gates in transmon qubits. We restrict ourselves to fixed-frequency transmon qubits (i.e. those that do not contain a SQUID for tunability) and fix coupling to neighboring qubits (which can also be made tunable using a SQUID). However, it only requires a small modification of the methods described below to adapt them to other classes of superconducting qubits. For a good overview of this topic, the reader is referred to Ref. [22].

The Bloch Sphere

To understand how to control the state of qubit, one must be familiar with the Bloch sphere representation of two level systems (see Fig. 2.10). The sphere itself is a unit sphere defined

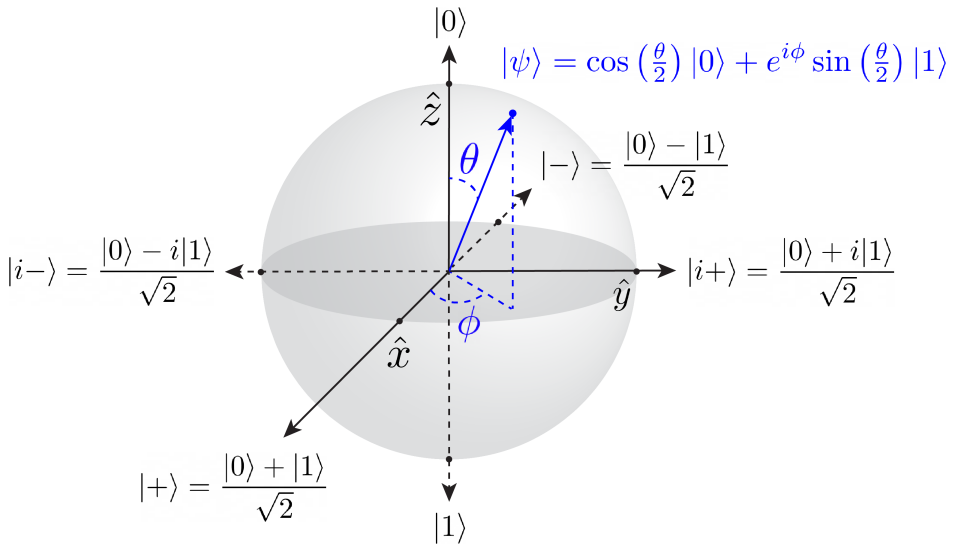


Figure 2.10: **Bloch sphere representation of a two-level system.** The north and south poles represent the ground $|0\rangle$ and first excited $|1\rangle$ states, respectively. The $|+\rangle$ and $|-\rangle$ states represent superposition states along the x-axis, and the $|i+\rangle$ and $|i-\rangle$ states represent superposition states along the y-axis. Any arbitrary state can be represented by $|\psi\rangle = \cos\left(\frac{\theta}{2}\right)|0\rangle + e^{i\phi}\sin\left(\frac{\theta}{2}\right)|1\rangle$ (blue), where θ and ϕ are polar and azimuthal angles, respectively.

by standard spherical coordinates. A state vector $|\psi\rangle$ can point in any direction along the surface of the Bloch sphere. We define the north and south poles of the Bloch sphere to be the ground state $|0\rangle$ and the first excited state $|1\rangle$, respectively. These two state vectors form the standard basis of the two-state quantum system. In matrix representation, these are given as

$$|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad |1\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{2.50}$$

Thus, any arbitrary state vector $|\psi\rangle$ can be expanded as a linear combination of $|0\rangle$ and $|1\rangle$ with complex coefficients α and β:

$$\boxed{|\psi\rangle = \alpha\,|0\rangle + \beta\,|1\rangle}. \tag{2.51}$$

Consequently, unlike a classical bit, which must be in a definite state of 0 or 1, a qubit can be in a superposition of $|0\rangle$ and $|1\rangle$, with α and β representing the relative weight of the two states. Using the normalization condition $\langle\psi|\psi\rangle = |\alpha|^2 + |\beta|^2 = 1$, we can write out α and β in terms of the spherical coordinates θ and ϕ:

$$\boxed{|\psi\rangle = \cos\left(\frac{\theta}{2}\right)|0\rangle + e^{i\phi}\sin\left(\frac{\theta}{2}\right)|1\rangle}. \tag{2.52}$$

Measurement of the qubit in the $\{|0\rangle, |1\rangle\}$ "computational" basis will result in 0 or 1 with probabilities $|\alpha|^2$ and $|\beta|^2$, respectively.

It is also possible to measure the state of a qubit in a different basis. For example, if we were to measure the qubit along the x-axis, we can define the $\{|+\rangle, |-\rangle\}$ basis to be

$$|+\rangle \equiv \frac{|0\rangle + |1\rangle}{\sqrt{2}}, \ |-\rangle \equiv \frac{|0\rangle - |1\rangle}{\sqrt{2}}. \tag{2.53}$$

Solving for $|0\rangle$ and $|1\rangle$ in terms of $|+\rangle$ and $|-\rangle$ yields

$$|0\rangle = \frac{|+\rangle + |-\rangle}{\sqrt{2}}, \ |1\rangle = \frac{|+\rangle - |-\rangle}{\sqrt{2}}. \tag{2.54}$$

Thus, in this basis, we can write an arbitrary state $|\psi\rangle$ as

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \tag{2.55}$$

$$= \alpha\frac{|+\rangle + |-\rangle}{\sqrt{2}} + \beta\frac{|+\rangle - |-\rangle}{\sqrt{2}} \tag{2.56}$$

$$= \frac{\alpha+\beta}{\sqrt{2}}|+\rangle + \frac{\alpha-\beta}{\sqrt{2}}|-\rangle \tag{2.57}$$

A similar expansion can be done for basis states $\{|i+\rangle, |i-\rangle\}$ defined along the y-axis as well.

Unitary Rotations for Qubit Control

Unitary rotation operators are necessary for qubit control in quantum computing, as quantum algorithms rely on qubit rotations for implementing logical gate operations. A unitary rotation operator in quantum mechanics takes the form

$$U(\hat{\mathbf{n}}, \theta) = e^{-\frac{i}{\hbar}\theta\hat{\mathbf{n}}\cdot\mathbf{J}}, \tag{2.58}$$

where θ is the rotation angle, $\hat{\mathbf{n}}$ the axis of rotation, and $\mathbf{J}$ a vector of Hermitian angular momentum operators (also called the generators of rotations); this is known as the *axis-angle parametrization* of unitary rotations[3]. For infinitesimal rotations, we can take the Taylor series of $U(\hat{\mathbf{n}}, \theta)$:

$$U(\hat{\mathbf{n}}, \theta) = 1 - \frac{i}{\hbar}\theta\hat{\mathbf{n}}\cdot\mathbf{J} + ... \tag{2.59}$$

Since finite rotations can be built out of the product of small-angle rotations, the infinitesimal (or near-identity) form of $U(\hat{\mathbf{n}}, \theta)$ will be useful for constructing a matrix representation of the unitary rotation operators.

[3]In the quantum computing literature, the axis-angle form is often written as $R_{\hat{\mathbf{n}}}(\theta)$; however, this notation does not differentiate from classical rotations, from which unitary rotations are derived.

Qubits are two-state quantum systems which are isomorphic to spin-$\frac{1}{2}$ systems, so the generators of rotations are simply the Pauli matrices:

$$\mathbf{J} = \mathbf{S} = \frac{\hbar}{2}\boldsymbol{\sigma}, \tag{2.60}$$

where

$$\boldsymbol{\sigma} = \sigma_x\hat{\mathbf{x}} + \sigma_y\hat{\mathbf{y}} + \sigma_z\hat{\mathbf{z}}, \tag{2.61}$$

and

$$\begin{aligned}
\sigma_x = X &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \\
\sigma_y = Y &= \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \\
\sigma_z = Z &= \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.
\end{aligned} \tag{2.62}$$

Thus, the unitary rotation operators for qubits can be written as

$$\boxed{U(\hat{\mathbf{n}}, \theta) = e^{-\frac{i}{2}\theta\hat{\mathbf{n}}\cdot\boldsymbol{\sigma}} = \cos\left(\tfrac{\theta}{2}\right) - i(\hat{\mathbf{n}}\cdot\boldsymbol{\sigma})\sin\left(\tfrac{\theta}{2}\right)}, \tag{2.63}$$

where we have used the standard properties of the Pauli matrices to simplify the Taylor expansion of $U(\hat{\mathbf{n}}, \theta)$ in terms of trigonometric functions. We can use the last part of Eq. 2.63 to write out the matrix form of the unitary rotation operators about each axis of the Bloch sphere:

$$\begin{aligned}
U(\hat{\mathbf{x}}, \theta) &= \cos\left(\tfrac{\theta}{2}\right) - i\sigma_x \sin\left(\tfrac{\theta}{2}\right) = \begin{pmatrix} \cos\frac{\theta}{2} & -i\sin\frac{\theta}{2} \\ -i\sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{pmatrix}, \\
U(\hat{\mathbf{y}}, \theta) &= \cos\left(\tfrac{\theta}{2}\right) - i\sigma_y \sin\left(\tfrac{\theta}{2}\right) = \begin{pmatrix} \cos\frac{\theta}{2} & -\sin\frac{\theta}{2} \\ \sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{pmatrix}, \\
U(\hat{\mathbf{z}}, \theta) &= \cos\left(\tfrac{\theta}{2}\right) - i\sigma_z \sin\left(\tfrac{\theta}{2}\right) = \begin{pmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{i\theta/2} \end{pmatrix}.
\end{aligned} \tag{2.64}$$

So, for example, performing a 90-degree rotation about the y-axis of a qubit on the $|0\rangle$ state results in

$$U\left(\hat{\mathbf{y}}, \tfrac{\pi}{2}\right)|0\rangle = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{|0\rangle + |1\rangle}{\sqrt{2}} = |+\rangle.$$

As we can see, in accordance with the Bloch sphere, a $\pi/2$ rotation about the y-axis results in the $|+\rangle$ superposition state.

To have complete control over the state of a qubit, one must be able to perform rotations on the qubit about all three axes. However, depending on the experimental setup and hardware platform, it might not be possible to have control over all three rotation axes. Nevertheless, by using the Euler angle parameterization of spinor rotations, it is possible to rotate about all three axes with control over only two of the three. Suppose, for example, it is only possible to control rotations about the x-axis and z-axis. In this case, any arbitrary rotation about the Bloch sphere is parametrized by three Euler angles,

$$U(\theta, \phi, \gamma) = U(\hat{\mathbf{z}}, \theta)U(\hat{\mathbf{x}}, \phi)U(\hat{\mathbf{z}}, \gamma), \tag{2.65}$$

which can range from

$$\begin{aligned} 0 \leq \theta \leq 2\pi, \\ 0 \leq \phi \leq \pi, \\ 0 \leq \gamma \leq 4\pi, \end{aligned} \tag{2.66}$$

where the 4π upper bound for γ is due to the double-valued representation for spin-$\frac{1}{2}$ systems (i.e. a particle rotated by 2π does not return to the original value, but rather acquires a phase of -1; see Eq. 2.63). For example, suppose $\theta = \frac{\pi}{2}$, $\gamma = -\frac{\pi}{2}$, and ϕ remains arbitrary. In this case, Eq. 2.65 gives

$$\begin{aligned} U\left(\tfrac{\pi}{2}, \phi, -\tfrac{\pi}{2}\right) &= U\left(\hat{\mathbf{z}}, \tfrac{\pi}{2}\right) U\left(\hat{\mathbf{x}}, \phi\right) U\left(\hat{\mathbf{z}}, -\tfrac{\pi}{2}\right) \\ &= \begin{pmatrix} e^{-i\pi/4} & 0 \\ 0 & e^{i\pi/4} \end{pmatrix} \begin{pmatrix} \cos\frac{\phi}{2} & -i\sin\frac{\phi}{2} \\ -i\sin\frac{\phi}{2} & \cos\frac{\phi}{2} \end{pmatrix} \begin{pmatrix} e^{i\pi/4} & 0 \\ 0 & e^{-i\pi/4} \end{pmatrix} \\ &= \begin{pmatrix} \cos\frac{\phi}{2} & \sin\frac{\phi}{2} \\ -\sin\frac{\phi}{2} & \cos\frac{\phi}{2} \end{pmatrix} \\ &= U(\hat{\mathbf{y}}, \phi) \end{aligned}$$

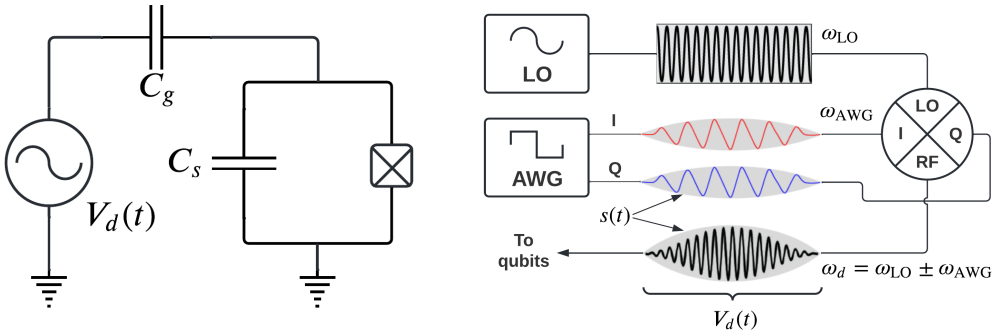
Thus, we can obtain universal control over the qubit with control over only two of the three rotation axes.

Rabi-driven Single-qubit Gates

In superconducting qubits, single-qubit gates are typically calibrated by Rabi driving a qubit. For fixed-frequency transmons, the control signal comes from microwave tones propagating down a control line which is capacitively-coupled to the qubit using a time-dependent voltage $V_d(t)$ (see Fig. 2.11a). The Hamiltonian for this system can be written as[4]

$$H(t) = \frac{Q(t)^2}{2C_T} + \frac{\Phi^2}{2L} + \frac{C_g}{C_T} Q V_d(t), \tag{2.67}$$

[4]This section follows the notation and derivation given in Ref. [22].



(a) Circuit diagram of qubit driving (b) IQ mixing for waveform generation

Figure 2.11: **Qubit driving.** (a) Circuit diagram of driving a qubit with a time dependent voltage $V_d(t)$. (b) Schematic of pulse generation for generating single-qubit gates. A local oscillator (LO) microwave source generates a carrier frequency at ω_{LO}. An arbitrary waveform generator (AWG) generates a pulse envelope $s(t)$ for in-phase I and quadrature Q baseband pulses at a frequency of ω_{AWG}. An IQ mixer combines these signals into a single pulse that is sent to the qubit.

where $C_T = C_g + C_s$ is the total capacitance to ground, and the charge variable $Q(t) = (C_g + C_s)\dot{\Phi} - C_g V_d(t)$ now takes into account the time-dependent voltage. Promoting the charge and flux variables to quantum operators, $Q \to \hat{Q}$ and $\Phi \to \hat{\Phi}$, we arrive at the familiar form for a quantum harmonic oscillator (Eq. 2.36) with an additional driving term:

$$\hat{H} = \hat{H}_{\mathrm{QHO}} + \frac{C_g}{C_T}\hat{Q}V_d(t). \tag{2.68}$$

By expressing the charge operator in terms of raising and lower operators,

$$\hat{Q} = -i\sqrt{\frac{\hbar}{2Z}}(a - a^\dagger) = -iQ_{\mathrm{ZPF}}(a - a^\dagger), \tag{2.69}$$

where Z is the impedance to ground and Q_{ZPF} the zero-point charge fluctuations, we may write the above Hamiltonian as

$$\hat{H} = \hbar\omega\left(a^\dagger a + \frac{1}{2}\right) - i\frac{C_g}{C_T}Q_{\mathrm{ZPF}}V_d(t)(a - a^\dagger). \tag{2.70}$$

By taking the two-level system approximation, we can replace $a \to \sigma^-$ and $a^\dagger \to \sigma^+$, and write the final Hamiltonian as

$$\hat{H} = \frac{1}{2}\hbar\omega\sigma_z + \hbar\Omega V_d(t)\sigma_y = \hat{H}_q + \hat{H}_d(t), \tag{2.71}$$

where $\hat{H}_q = \frac{1}{2}\hbar\omega\sigma_z$ is the qubit Hamiltonian and $\Omega = \frac{C_g}{C_T}Q_{\text{ZPF}}$ is related to the Rabi driving frequency between the ground state and first excited state.

To understand the impact of the driving term $\hat{H}_d(t)$ in the Hamiltonian, we must move into the rotating frame of the qubit. The propagator $\hat{U}_0$ for the qubit Hamiltonian is given as

$$\hat{U}_0(t) = e^{-i\hat{H}_q t/\hbar}. \tag{2.72}$$

Therefore, to move into the rotating frame of the qubit, we define a propagator $\hat{U}(t) = \hat{U}_0^\dagger(t)$ and apply this propagator to the state of the qubit in the lab frame $|\psi_0\rangle$, $|\psi(t)\rangle = \hat{U}(t)\,|\psi_0\rangle$. $|\psi(t)\rangle$ represents the state of the qubit in a frame that rotates at the same frequency as the qubit transition frequency $\omega_q = (E_1 - E_0)/\hbar$ in the lab frame. Writing the time-dependent Schrödinger equation for $|\psi(t)\rangle$, we find

$$\begin{aligned}
i\hbar\frac{\partial}{\partial t}|\psi(t)\rangle &= i\hbar\frac{\partial}{\partial t}\hat{U}(t)\,|\psi_0\rangle \\
&= i\hbar\frac{\partial\hat{U}(t)}{\partial t}|\psi_0\rangle + i\hbar\hat{U}(t)\frac{\partial\,|\psi_0\rangle}{\partial t} \\
&= i\hbar\dot{U}(t)\,|\psi_0\rangle + \hat{U}(t)H_q\,|\psi_0\rangle \\
&= i\hbar\dot{U}(t)\hat{U}^\dagger(t)\,|\psi(t)\rangle + \hat{U}(t)H_q\hat{U}^\dagger(t)\,|\psi(t)\rangle \\
&= \tilde{H}_q(t)\,|\psi(t)\rangle\,,
\end{aligned}$$

where $\tilde{H}_q = i\hbar\dot{U}\hat{U}^\dagger + \hat{U}H_q\hat{U}^\dagger$ is the qubit Hamiltonian in the rotating frame.

In the rotating frame the qubit state is static on the Bloch sphere, except under the influence of an external driving term, such as $H_d(t)$. To understand the influence of $H_d(t)$ on the state of the qubit in the rotating frame, we must also transform $H_d(t)$ to the rotating frame. Doing so, we find the transformed Hamiltonian to be

$$\tilde{H}_d(t) = \hbar\Omega V_d(t)\left[\cos(\omega_q t)\sigma_y - \sin(\omega_q t)\sigma_x\right]. \tag{2.73}$$

The time-dependent driving voltage $V_d(t)$ is composed of three different components: an amplitude V_0, a phase $\sin(\omega_d t + \phi)$, and a time-dependent pulse envelope $s(t)$ (see Fig. 2.11b). Therefore, we can write $V_d(t)$ as

$$\begin{aligned}
V_d(t) &= V_0 s(t)\sin(\omega_d t + \phi) &(2.74)\\
&= V_0 s(t)\left[\sin(\omega_d t)\cos(\phi) + \cos(\omega_d t)\sin(\phi)\right], &(2.75)
\end{aligned}$$

where in the second line we have used the difference formula for sines. By defining $I \equiv \cos(\phi)$ as the in-phase component of the driving signal, and $Q \equiv \sin(\phi)$ as the out-of-phase (or "quadrature") component of the driving signal, we may write the time-dependent voltage as

$$V_d(t) = V_0 s(t)\left[I\sin(\omega_d t) + Q\cos(\omega_d t)\right]. \tag{2.76}$$

Therefore, our driving Hamiltonian becomes

$$\tilde{H}_d(t) = \hbar\Omega V_0 s(t)\left[I\sin(\omega_d t) + Q\cos(\omega_d t)\right]\left[\cos(\omega_q t)\sigma_y - \sin(\omega_q t)\sigma_x\right]. \tag{2.77}$$

To simplify the above equation, we multiply out the various terms involving products of sines and cosines, and then write them out using product-to-sum trigonometric formulas. When we do so, we will see terms involving the sum $(\omega_q + \omega_d)$ and terms involving the difference $(\omega_q - \omega_d)$. Next, we invoke the rotating wave approximation (RWA) and drop the fast rotating terms which contain $(\omega_q + \omega_d)$, which will average to zero on the timescale associated with the terms containing $(\omega_q - \omega_d)$. Simplifying the resulting terms and defining $\delta = \omega_q - \omega_d$, we can write the driving Hamiltonian as

$$\tilde{H}_d(t) = -\frac{\Omega}{2} V_0 s(t) \left\{ I \left[\cos(\delta t)\sigma_x - \sin(\delta t)\sigma_y \right] + Q \left[\cos(\delta t)\sigma_y - \sin(\delta t)\sigma_x \right] \right\}. \tag{2.78}$$

If we assume that the detuning $\delta = 0$ (i.e. driving our qubit on-resonance), then the above equation simplifies to

$$\tilde{H}_d(t) = -\frac{\hbar\Omega}{2} V_0 s(t)(I\sigma_x + Q\sigma_y). \tag{2.79}$$

Here, we can see that the in-phase pulse corresponds to a rotating about the x-axis, whereas an out-of-phase pulse corresponds to a rotation about the y-axis. However, if we allow off-resonant drives, then general form of $\tilde{H}_d(t)$ can be written as

$$\boxed{\tilde{H}_d(t) = -\frac{\hbar\Omega}{2} V_0 s(t) \begin{pmatrix} 0 & e^{i(\delta t+\phi)} \\ e^{-i(\delta t+\phi)} & 0 \end{pmatrix}}. \tag{2.80}$$

The choice of which phase component — I or Q — defines an X gate and which defines a Y gate is arbitrary, representing a gauge degree of freedom in the system. However, as defined in Eq. 2.79, $\phi = 0$ corresponds to an X gate and $\phi = \pi/2$ corresponds to a Y gate. To understand how Rabi-driven single-qubit gates are implemented via unitary operators, consider the case in which $\delta = 0$ (on resonance) and $\phi = 0$ (i.e. $I = 1$ and $Q = 0$), where we may write the following unitary operator, corresponding to the Bloch state evolution under $\tilde{H}_d(t)$:

$$U(t) = \exp\left(-\frac{i}{\hbar} \int_0^t \tilde{H}_d(t')dt' \right) \tag{2.81}$$

$$= \exp\left(\frac{i}{2} \left[\Omega V_0 \int_0^t s(t')dt' \right] \sigma_x \right) \tag{2.82}$$

$$= \exp\left(-\frac{i}{2} \Theta(t)\sigma_x \right), \tag{2.83}$$

where $\Theta(t) = -\Omega V_0 \int_0^t s(t')dt$. Eq. 2.81 governs the dynamics of Rabi-driven single-qubit gates. In the following equations, we see that we recover that exact same functional form of unitary gates in SU(2) (Eq. 2.63) for a rotation about the x-axis, where the rotation angle $\Theta(t)$ depends on the amplitude of the driving signal V_0 and the time-integrated envelope

function $s(t)$. Therefore, in order to "calibrate" a quantum gate (e.g. $\Theta = \pi/2$), one can either fix V_0 and vary the shape and/or time of the signal envelope $s(t)$, or fix $s(t)$ and sweep across values of V_0 until the desired rotation angle is found.

In superconducting systems, arbitrary single-qubit gates containing both in-phase I and out-of-phase Q signal components are generated using an arbitrary waveform generator (AWG), which also allows the experimenter to define the shape of the pulse envelopes $s(t)$ (e.g. Gaussian, cosine, etc.; see Fig. 2.11b) and choose whether to perform single-qubit gates about the x- or y-axis. However, AWGs typically cannot generate pulses at the frequencies necessary to drive superconducting qubits ($\sim 5 - 6$ GHz). Therefore, to generate a pulse at a frequency ω_d (which is assumed to be close or equal to ω_q), it is necessary to mix the $I(Q)$ signals with the in-phase (out-of-phase) components of a local oscillator (LO), termed "IQ mixing." The detuning between the AWG baseband frequency ω_{AWG} and the LO frequency ω_{LO} is such that the drive frequency is given as $\omega_d = \omega_{\text{LO}} \pm \omega_{\text{AWG}}$. This means that no matter whether the two frequencies add or subtract to give ω_d, there exists an unwanted side band signal at either $\omega_{\text{LO}} + \omega_{\text{AWG}}$ (if $\omega_d = \omega_{\text{LO}} - \omega_{\text{AWG}}$) or $\omega_{\text{LO}} - \omega_{\text{AWG}}$ (if $\omega_d = \omega_{\text{LO}} + \omega_{\text{AWG}}$) which must be "nulled" using filters in order to ensure a hygienic EM environment inside the dilution fridge. Poorly nulled LO signals can lead to "environmental noise" on the qubit processor, leading to dephasing (since it is off-resonance) and thus low coherence times. Once quantum gates have been well calibrated, any circuit $\mathcal{C}$ of single-qubit gates can be written in terms of Rabi-driven unitary rotations:

$$\mathcal{C} = U_N U_{N-1} ... U_2 U_1 = \prod_{n=0}^{N} e^{-\frac{i}{2}\Theta_n(t)(I_n\sigma_x + Q_n\sigma_y)}. \tag{2.84}$$

Universal Qubit Control via Virtual Z Gates

One of the primary bottlenecks in real-world quantum computers is the calibration of quantum gates. If one desires to perform arbitrary SU(2) rotations, it is not feasible to calibrate V_0 and/or $s(t)$ for all possible values of $\Theta(t)$, and for all rotation axes, since the space of all possible SU(2) rotations is infinite. We have previously shown that we can perform rotations about the x- or y-axis by changing the phase ϕ of the drive signal, generating polar rotations about the Bloch sphere (see the Θ parameter in Fig. 2.10). However, in order to generate phase rotations about the Bloch sphere (i.e. changing the ϕ parameter in Fig. 2.10, not to be confused with the phase ϕ of the drive signal), we must be able to control rotations about the z-axis.

In general, any arbitrary SU(2) gate can be decomposed as a unitary operator containing three different rotation angles:

$$\boxed{U(\phi, \theta, \lambda) = \begin{pmatrix} \cos\left(\frac{\theta}{2}\right) & -e^{i\lambda}\sin\left(\frac{\theta}{2}\right) \\ e^{i\phi}\sin\left(\frac{\theta}{2}\right) & e^{i(\phi+\lambda)}\cos\left(\frac{\theta}{2}\right) \end{pmatrix}}. \tag{2.85}$$

Gate	θ	ϕ	λ	Type
I	0	0	0	Pauli
X_π	π	0	0	Pauli
Y_π	π	$\pi/2$	-$\pi/2$	Pauli
Z_π	0	$\pi/2$	$\pi/2$	Pauli
$X_{\pi/2}$	$\pi/2$	0	0	Clifford
$Y_{\pi/2}$	$\pi/2$	$\pi/2$	-$\pi/2$	Clifford
S	0	$\pi/4$	$\pi/4$	Clifford
H	$\pi/2$	$\pi/2$	$\pi/2$	Clifford
$X_{\pi/4}$	$\pi/4$	0	0	Non-Clifford
T	0	$\pi/8$	$\pi/8$	Non-Clifford

Table 2.1: **Angles for common SU(2) gates performed with the $ZXZXZ$-decomposition.** (Table adapted from Ref. [55].)

This is equivalent to the following decomposition (up to a global phase),

$$U(\phi, \theta, \lambda) = Z_\phi X_\theta Z_\lambda. \tag{2.86}$$

To see this[5], imagine we implement two $X_\theta = e^{-i\frac{\theta}{2}\sigma_x}$ pulses, but add some phase *offset* γ between the pulses. This may be written as a circuit

$$\mathcal{C} = e^{-i\frac{\theta}{2}[\cos(\gamma)\sigma_x + \sin(\gamma)\sigma_y]} e^{-i\frac{\theta}{2}\sigma_x}. \tag{2.87}$$

The leftmost term in the above equation is equivalent to the unitary evolution under the Hamiltonian in Eq. 2.79. If the phase $\gamma = 0$, then $\mathcal{C} = X_\theta X_\theta$. However, for general γ, we can expand the above equation as

$$\begin{aligned} \mathcal{C} &= e^{i\frac{\gamma}{2}\sigma_z} e^{-i\frac{\theta}{2}\sigma_x} e^{-i\frac{\gamma}{2}\sigma_z} e^{-i\frac{\theta}{2}\sigma_x} && (2.88) \\ &= Z_{-\gamma} X_\theta Z_\gamma X_\theta. && (2.89) \end{aligned}$$

We see that what differentiates the first X_θ from the second X_θ is the phase offset Z_γ in between the two. For example, if $\gamma = \pi/2$, then $Z_{-\pi/2} X_\theta Z_{\pi/2} = Y_\theta$. Note that this is equivalent to Eq. 2.65, in which we showed that $U\left(\frac{\pi}{2}, \theta, -\frac{\pi}{2}\right) = U\left(\hat{\mathbf{z}}, \frac{\pi}{2}\right) U\left(\hat{\mathbf{x}}, \theta\right) U\left(\hat{\mathbf{z}}, -\frac{\pi}{2}\right) =$

[5] This section follows Ref. [55].

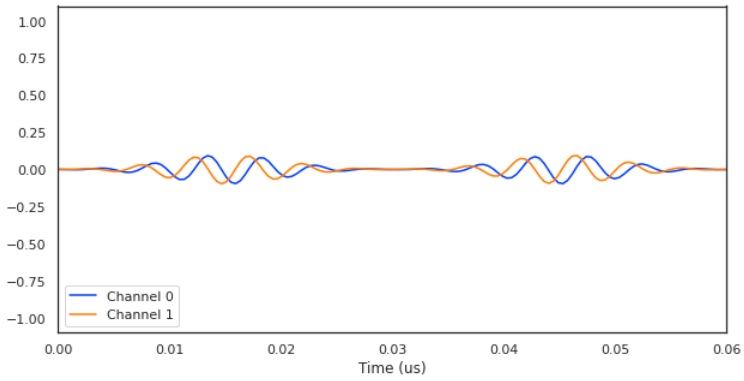


Figure 2.12: **Physical pulse structure of an X_π gate in the $ZXZXZ$-decomposition.** Channel 0 (1) denotes the in-phase I (quadrature Q) component of the pulse, plotted in arbitrary amplitude units. The two pulse envelopes represent the physical $X_{\pi/2}$ gates. All virtual Z gates are implemented as phase shifts in the sequence between $X_{\pi/2}$ gates, and therefore do not appear as physical pulses.

$U(\hat{\mathbf{y}}, \theta)$ via Euler-angle decomposition. Therefore, we can achieve arbitrary Z phase gates by simply adding a phase *offset* between physical pulses. Because these phase offsets occur in software (they are not physical pulses implemented on the qubits, but rather phase shifts between physical pulses), they are often referred to as *virtual Z gates* [55]. The trailing $Z_{-\gamma}$ in the above equation is due to the rotation being in the reference frame of the qubit and has no impact on the measurement outcomes of the circuit in the computational basis, but it is necessary if we were to add another gate at the end of the sequence.

To minimize the calibration overhead for generating arbitrary single-qubit gates, we fix our angle $\theta = \pi/2$ and calibrate only a single gate $X_{\pi/2}$ for each qubit. Using the identity

$$X_\theta = Z_{-\phi/2} X_{\pi/2} Z_{\pi-\theta} X_{\pi/2} Z_{-\phi/2}, \tag{2.90}$$

we can write any SU(2) as

$$\boxed{U(\phi, \theta, \gamma) = Z_{\phi-\pi/2} X_{\pi/2} Z_{\pi-\theta} X_{\pi/2} Z_{\gamma-\pi/2}}. \tag{2.91}$$

This is termed the $ZXZXZ$-decomposition of single-qubit gates. Therefore, calibrating a single physical $X_{\pi/2}$ gate (via resonant Rabi-driven pulses), combined with virtual Z gates, is sufficient to perform arbitrary single-qubit rotations. In Table 2.1, we list the phases for Eq. 2.91 for several common SU(2) gates. In Fig. 2.12, we plot an example pulse sequence for an X_π gate in the $ZXZXZ$-decomposition.

2.5 Two-qubit Gates

Methods for generating entanglement between two or more qubits are unique to different quantum computing platforms. Even within the field of superconducting circuits, the mech-

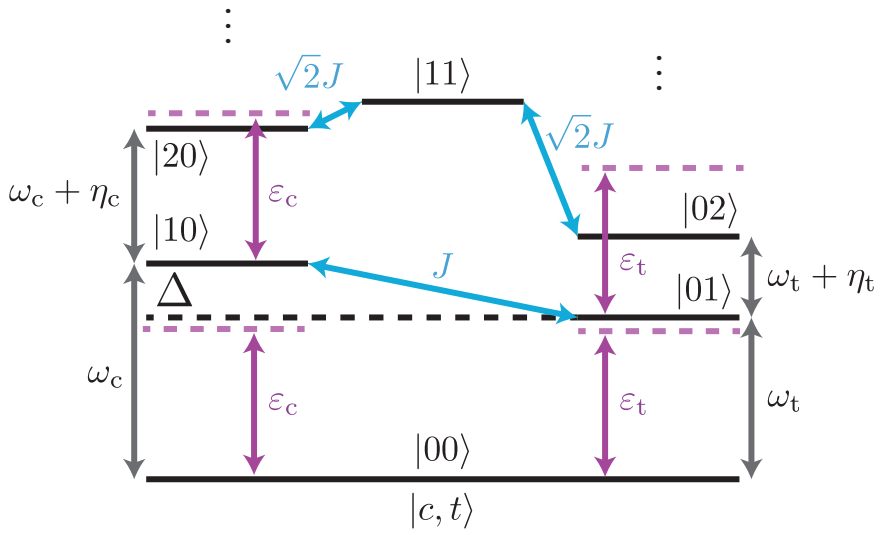


Figure 2.13: **Energy level diagram for Stark-induced ZZ interaction.** The control and target qubits are characterized by transition frequencies ω_c and ω_t (solid black lines), and anharmonicities η_c and η_t (purple dashed lines), respectively. $\Delta = \omega_c - \omega_t$ denotes the detuning between qubits (black dashed line), with J the qubit-to-qubit coupling strength. ε_c and ε_t denote the drive amplitudes on the control and target qubits, respectively. The qubits are driven simultaneous at a detuned frequency between ω_t and $\omega_c + \eta_c$, resulting in a conditional Stark shift on both qubits. (Figure reprinted with permission from Ref. [60].)

anism by which two qubits are entangled depends not only on the flavor of qubit, but also the type of coupler employed, which can be inductive, capacitive, or parametric, in which the coupling element can be tuned to generate different types of entangling interactions. A popular two-qubit gate on superconducting hardware platforms with fixed-frequency transmon qubits is the so-called "cross-resonance" gate [56, 57, 58, 59], in which the target qubit is driven on-resonance through the control line of the control qubit, such that the amplitude of the driving signal at the location of the target qubit depends on the state of the control qubit. By calibrating the duration of the cross-resonance gate such that the target qubit is unchanged (flipped) when the control qubit is in $|0\rangle$ ($|1\rangle$), one can create a two-qubit interaction that is locally-equivalent to a controlled-X gate. While the cross-resonance gate is employed in some of the experiments described in this work, here we instead describe a novel scheme for calibrating a tuneable interaction via off-resonant drives between fixed-frequency, fix-coupling transmon qubits. This section is based on the work presented in Ref. [60].

Figure 2.13 shows an energy level diagram for two qubits of frequency ω_c and ω_t, and anharmonicities η_c and η_t, coupled with a coupling rate J, where c (t) denotes the control (target) qubit. For a qubit drive frequency ω_d, the system Hamiltonian in the Duffing

approximation [42] is given as

$$H = \sum_{i=\mathrm{c,t}} \left[\left(\omega_i - \omega_\mathrm{d}\right) a_i^\dagger a_i + \frac{\eta_i}{2} a_i^\dagger a_i^\dagger a_i a_i + \left(\varepsilon_i a_i + \varepsilon_i^* a_i^\dagger \right) \right] + J \left(a_\mathrm{c}^\dagger a_\mathrm{t} + a_\mathrm{c} a_\mathrm{t}^\dagger \right), \tag{2.92}$$

where a_i ($a_i^\dagger$) is the bosonic annihilation (creation) operator, ε_i the complex drive amplitude, and $\hbar = 1$. A Stark-induced ZZ interaction can be achieved by driving the control qubit at the target qubit frequency ω_t with an amplitude ε_c. Similar to the cross-resonance interaction described above, the target qubit experiences a drive amplitude $\tilde{\varepsilon}_n$, where n denotes the state of the control qubit ($|n\rangle \in \{|0\rangle, |1\rangle)\}$). In the limit that the drive amplitude ε_c is much smaller than the detuning between the drive frequency and target qubit frequency $\Delta_\mathrm{t} = \omega_\mathrm{t} - \omega_\mathrm{d}$ (i.e. $\varepsilon_\mathrm{c}/\Delta_\mathrm{t} \ll 1$), the drive signal results in a conditional Stark shift of the target qubit frequency

$$\tilde{\delta}_n = \frac{\tilde{\varepsilon}_n^2}{\Delta_\mathrm{t}}. \tag{2.93}$$

This conditional Stark shift manifests as a drive-induced ZZ interaction $\zeta = \tilde{\delta}_0 - \tilde{\delta}_1$, which can be written as

$$\zeta = 2\mu \left(\tilde{\varepsilon}_0 + \tilde{\varepsilon}_1\right) / \Delta_\mathrm{t}, \tag{2.94}$$

where $\mu = \left(\tilde{\varepsilon}_0 - \tilde{\varepsilon}_1\right)/2$ is cross-resonance coupling rate. In general, the conditional stark shift is much smaller than the cross-resonance rate $\zeta \ll \mu$. However, when a driving term is applied directly to the target qubit at amplitude ε_t (in addition to the driving term driven through the control qubit), the total drive amplitude becomes $\tilde{\varepsilon}_n + \varepsilon_\mathrm{t}$. Replacing $\tilde{\varepsilon}_n \to \tilde{\varepsilon}_n + \varepsilon_\mathrm{t}$ in the above equation, we see that, to first order, ζ scales linearly with ε_t:

$$\zeta = \frac{2\mu}{\Delta_\mathrm{t}} \left(\tilde{\varepsilon}_0 + \tilde{\varepsilon}_1 + 2\varepsilon_\mathrm{t}\right) + O(|\varepsilon_\mathrm{t}|^2). \tag{2.95}$$

Therefore, by driving both transmons simultaneously off-resonant, we can generate a ZZ coupling rate that is comparable to the cross-resonance rate.

Using this tunable ZZ interaction, we calibrate a CZ gate, given by the following unitary:

$$CZ = \exp\left(-\frac{i}{2}\frac{\pi}{2} \left(-ZI - IZ + ZZ\right) \right), \tag{2.96}$$

where IZ, ZI, and ZZ represent the relevant terms in the Hamiltonian. In matrix form, the CZ is given by

$$CZ = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \tag{2.97}$$

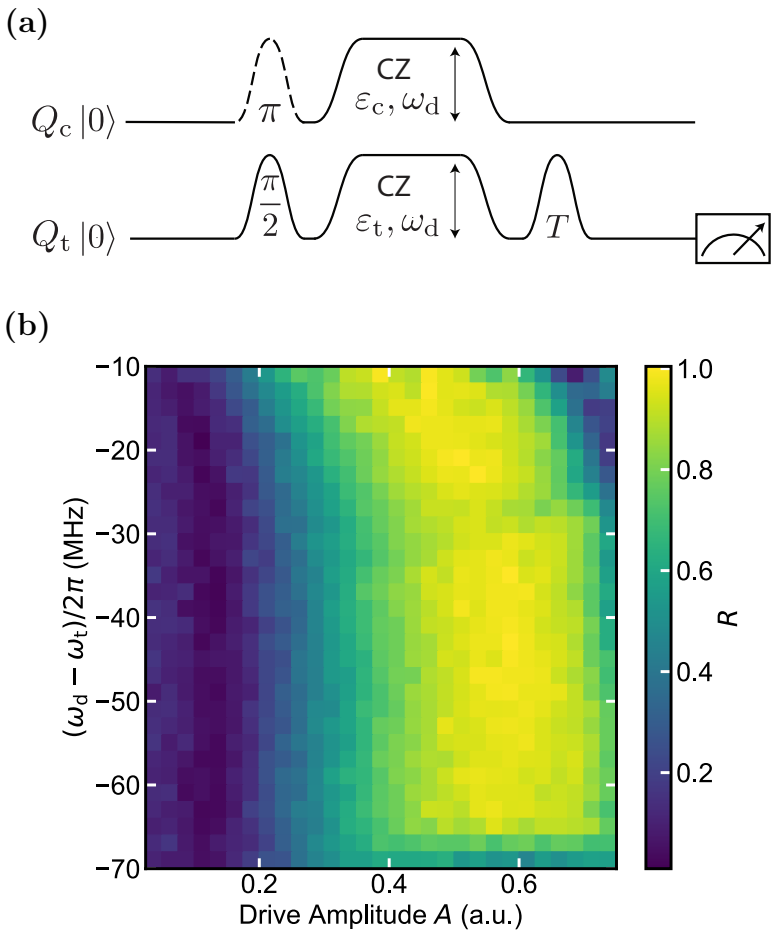


Figure 2.14: CZ **Gate Calibration.** (a) Pulse sequence for calibrating the amplitude and frequency of the CZ pulse. The control qubit is prepared in $|0\rangle$ (or $|1\rangle$, by application of a π gate), and the target qubit is prepared in a superposition state by application of a $\pi/2$ gate. The CZ gate is then applied for a duration of time, with the pulse amplitudes ε and pulse frequency ω_d variable parameters. At the end of the CZ gate, partial tomography is applied to the target qubit to measure the target Bloch vector $\mathbf{r}_i$ for each control qubit state $|i\rangle \in \{0, 1\}$. The global pulse amplitude and frequency are calibrated by selecting parameters that maximize the entanglement measure R (Eq. 2.98). (b) 2D plot of R as a function of the CZ gate amplitude A and drive detuning from the target $(\omega_d - \omega_t)$. There exists regions where R is maximal (yellow), which is used to realize the CZ gate. There also exists regions where R is minimal (blue), corresponding to ZZ cancellation. The global drive amplitude A is related to the individual drive line amplitudes (see Ref. [60] for more details). (Figure reprinted with permission from Ref. [60].)

which is symmetric between both qubits; therefore, even though we have been using the terms "control" and "target" to describe the two qubits involved in the CZ gate, these labels are interchangeable.

To calibrate the ZZ term, we prepare the target qubit in a superposition state, prepare the control qubit in $|0\rangle$ ($|1\rangle$), apply the CZ pulse sequence for some user-specified duration, and then perform partial tomography on the final state of the target qubit by measuring in X and Y in order to measure the length of the block vector $\mathbf{r}_0$ ($\mathbf{r}_1$); see Fig. 2.14a. For a maximally-entangling gate like the CZ, the difference between the Bloch vectors $\mathbf{r}_0$ and $\mathbf{r}_1$ (corresponding to when the control qubit is in $|0\rangle$ and $|1\rangle$, respectively) should be maximized. Therefore, we search for a parameter regime which maximizes the following quantity, termed the "conditionality,"

$$R = \frac{1}{2}||\mathbf{r}_0 - \mathbf{r}_1||^2, \tag{2.98}$$

which measures the normalized vector distance between target Bloch vectors conditioned on the control qubit state [59]. The drive amplitude, drive frequency, and pulse time are determined by finding regions in which the conditionality $R = 1$ for a single CZ pulse (see Ref. [60] for more details). In Fig. 2.14b, we observe a ~ 40 MHz bandwidth region with maximal conditionality. The drive amplitudes and frequency can be further fine-tuned via pulse amplification (i.e. performing the sequences in Fig. 2.14a with an odd integer number of CZ gates). We also observe regions in which the conditionality $R = 0$, indicating parameter regions in which the static ZZ coupling is entirely nulled. Once the ZZ term has been calibrated, to map this to a CZ gate we calibrate the IZ and ZI terms in the Hamiltonian by measuring the individual qubit Pauli Z errors using Ramsey-type experiments (see Ref. [60] for more details), and correcting the errors on each qubit using virtual Z gates [55].

We assess the performance of the CZ gate using various benchmarking procedures, which we discuss in Chapter 5. As previously mentioned, some of the data presented in Chapters 6 and 7 utilize the cross-resonance CX gate, but most of the data utilizes the CZ gate outlined above. While CX and CZ gates are locally-equivalent, the Stark-induced ZZ interaction is preferred to the cross-resonance interaction, as it makes use of the native ZZ coupling between transmon qubits, and has a much larger operating regime for maximal entanglement than the cross-resonance gate (see Fig. 2.14), lowering the calibration overhead for an entangling gate. We note that a similar scheme has been used to generate entanglement between two qutrits [61].

2.6 Qubit Readout

The final component needed to perform computations on superconducting quantum processors is the ability to read out the state of the qubits at the end of a quantum circuit. The measurement of transmon qubits can be understood in the circuit QED framework, which, like cavity QED, can be represented by the Jaynes-Cummings Hamiltonian [62, 63, 16] for

a two-level system coupled to a cavity resonator:

$$H_{\text{JC}} = H_q + H_r + H_{\text{int}} \tag{2.99}$$

$$= \frac{1}{2}\hbar\omega_q\sigma_z + \hbar\omega_r\left(a^\dagger a + \frac{1}{2}\right) + \hbar g(a + a^\dagger)(\sigma^- + \sigma^+), \tag{2.100}$$

where H_q is the qubit Hamiltonian (Eq. 2.49), H_r is the resonator Hamiltonian (i.e. QHO, Eq. 2.36), and H_{int} is the interaction Hamiltonian which describes the dipole coupling between the qubit and cavity, characterized by a coupling rate g; here, $a^\dagger$ (a) denotes the bosonic creation and annihilation, and σ^+ (σ^-) the qubit raising (lowering) operator. If we multiply out the interaction Hamiltonian, we obtain

$$H_{\text{int}} = \hbar g(a\sigma^- + a\sigma^+ + a^\dagger\sigma^- + a^\dagger\sigma^+). \tag{2.101}$$

We see that H_{int} contains terms like $a\sigma^+$ and $a^\dagger\sigma^-$, which represent the coherent swapping of quanta between the qubit and cavity resonator. We additionally see terms like $a\sigma^-$ and $a^\dagger\sigma^+$ which appear to remove or add two quanta of energy, violating energy conservation. The typical approach here is to invoke the rotating wave approximation (RWA), in which the terms in H_{int} which do not conserve energy are simply discarded. However, the RWA is only valid on certain timescales. To see this, we express the interaction Hamiltonian in the interaction picture by performing the following transformation:

$$H_{\text{int,I}} = e^{iH_0 t/\hbar} H_{\text{int}} e^{-iH_0 t/\hbar},$$

where $H_0 = H_q + H_r$. In the interaction picture, H_{int} becomes

$$H_{\text{int,I}} = \hbar g\left(a\sigma^- e^{-i(\omega_q+\omega_r)t} + a\sigma^+ e^{-i(\omega_q-\omega_r)t} + a^\dagger\sigma^- e^{i(\omega_q-\omega_r)t} + a^\dagger\sigma^+ e^{i(\omega_q+\omega_r)t}\right), \tag{2.102}$$

where we see that the swapping terms oscillate at a frequency $(\omega_q - \omega_r)$, whereas the double excitation/de-excitation terms oscillate at a frequency $(\omega_q + \omega_r)$. Therefore, it is only valid to discard $a\sigma^-$ and $a^\dagger\sigma^+$ on timescales much longer than $t \sim 1/(\omega_q + \omega_r)$, in which these fast oscillating terms average to zero. Superconducting transmon qubits are typically measured in the dispersive coupling limit, in which the qubit-resonator detuning $\Delta = \omega_q - \omega_r$ is much larger than the coupling rate g (i.e. $\Delta >> g$). In this regime, we are only concerned with timescales on the order of $t \sim 1/(\omega_q - \omega_r)$, and we are therefore justified in dropping the terms that oscillate at $(\omega_q + \omega_r)$. However, we note that describing the terms $a\sigma^-$ and $a^\dagger\sigma^+$ as non-energy conserving is somewhat misleading, as the full Jaynes-Cummings Hamiltonian governs the true behavior of the coupled qubit-resonator system. Rather, we see that $a\sigma^-$ and $a^\dagger\sigma^+$ are allowed to occur, even though they appear to violate the conservation of energy, but only for a short amount of time. On longer timescales, these terms do not contribute to the system dynamics, and therefore after invoking the RWA and transforming back to the Schrödinger picture, the Jaynes-Cummings Hamiltonian becomes

$$\boxed{H_{\text{JC}} = \frac{1}{2}\hbar\omega_q\sigma_z + \hbar\omega_r\left(a^\dagger a + \frac{1}{2}\right) + \hbar g(a\sigma^+ + a^\dagger\sigma^-)}. \tag{2.103}$$

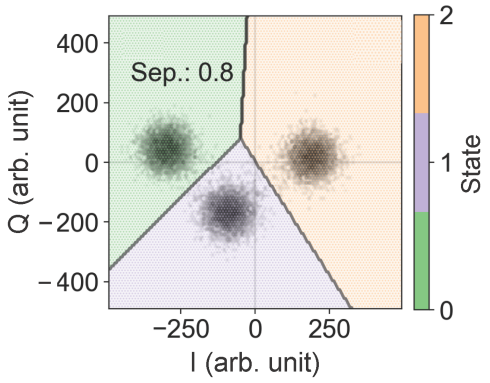


Figure 2.15: **Qutrit classification.** Ensemble measurements of a qutrit prepared in $|0\rangle$, $|1\rangle$, and $|2\rangle$ are plotted in phase space. We discriminate between the three states by classifying the results using a Gaussian mixture model fit to the in-phase (I) and quadrature (Q) heterodyne voltage signals. (Figure courtesy of Long Nguyen.)

In the dispersive limit, in which the qubit and resonator are far detuned such that $\Delta >> g$, there is no longer any direct exchange of energy between the qubit and resonator. Because g/Δ is small, one can derive the dispersive approximation to H_{JC} by performing a second-order perturbative expansion in terms of g/Δ [64, 65], yielding

$$H_{\text{JC}} = \frac{1}{2}\hbar(\omega_q + \chi)\sigma_z + \hbar\omega_r\left(a^\dagger a + \frac{1}{2}\right) + \hbar\chi a^\dagger a\sigma_z, \tag{2.104}$$

where $\chi = \frac{g^2}{\Delta}$ is the dispersive frequency shift. Interestingly, we see that the qubit frequency is shifted by χ; this is due to the qubit interacting with vacuum fluctuations in the cavity, known as the Lamb shift. It is more informative to rearrange the last term in the dispersive approximation of H_{JC} in one of two ways:

$$H_{\text{JC}} = \begin{cases} \frac{1}{2}\hbar(\omega_q + \chi + 2\chi a^\dagger a)\sigma_z + \hbar\omega_r\left(a^\dagger a + \frac{1}{2}\right) \\ \frac{1}{2}\hbar(\omega_q + \chi)\sigma_z + \hbar\left(\omega_r + \chi\sigma_z\right)\left(a^\dagger a + \frac{1}{2}\right) \end{cases} \tag{2.105}$$

In the first case, we see that the presence of photons in the resonator shifts the qubit frequency by an amount given by $2\chi n$, where $n = a^\dagger a$ is the number of photons in the resonator; this is known as the AC Stark shift, and has the unintended consequence of causing qubit dephasing if there are fluctuations in n due to thermal noise or spurious photon occupation in the resonator. In the second case, we see that the presence of the qubit coupled to the cavity causes a state-dependent frequency shift of χ in the resonator frequency, but the sign of the frequency shift depends on whether the qubit is in $|0\rangle$ or $|1\rangle$. By probing the resonant frequency of the resonator, and performing heterodyne detection of the reflected or

transmitted signal to measure the resonator amplitude and phase[6], we are able to measure the qubit state and determine whether it was in $|0\rangle$ or $|1\rangle$. While our treatment of this topic was limited to a two-level system in order to derive the dispersive approximation to the Jaynes-Cummings Hamiltonian, one can extend this derivation to the higher energy levels in superconducting qubits. In Fig. 2.15, we plot the classification of quantum trit (qutrit) states in phase space for our system. We see the presence of three distinct pointer states, allowing us to determine whether our qutrit was in $|0\rangle$, $|1\rangle$, or $|2\rangle$ before measurement. Even if computations are limited to qubit gates, qutrit classification allows one to monitor leakage outside of the computational basis states.

2.7 DiVincenzo's Criteria

There is a specific set of criteria that a physical system needs to fulfill in order to be considered a viable candidate for building a quantum computer. These requirements are known as DiVincenzo's criteria [66]:

1. It must be a scaleable physical system with well-characterized qubits.
2. It must be possible to initialize the qubits into a known state (i.e. $|0\rangle$ or $|1\rangle$).
3. The decoherence time of the qubits must be much longer than the operation time.
4. A "universal" set of quantum gates must exist for the qubits.
5. Measurement of the state of a qubit must be attainable.

In Fig. 2.16, we show a micrograph of the superconducting quantum processing unit (QPU) used in this work, and highlight the various components of the QPU that enable the implementation of arbitrary quantum computations. We use this figure to guide our discussion of how each of DiVincenzo's criteria is fulfilled by Josephson junction-based qubits:

1. The JJ is the critical component of all three superconducting qubit archetypes. Since these junctions can be readily fabricated using standard industrial microfabrication techniques, superconducting qubits are very scaleable on the surface of 2D chips. Work is also underway to integrate superconducting qubits into 3D chips, to set the stage for larger-scale processors [67]. Because of the tunability of JJs and the surrounding circuitry, superconducting qubits are easily characterized. Single-qubit manipulations (i.e. read and write) through dedicated microwave control lines allow for the unique characterization (i.e. decoherence time, dephasing time, gate error rate, etc.) of each qubit on a chip.

[6] We do not cover this topic here. However, the interested reader is referred to Ref. [22] for a brief introduction to this topic.

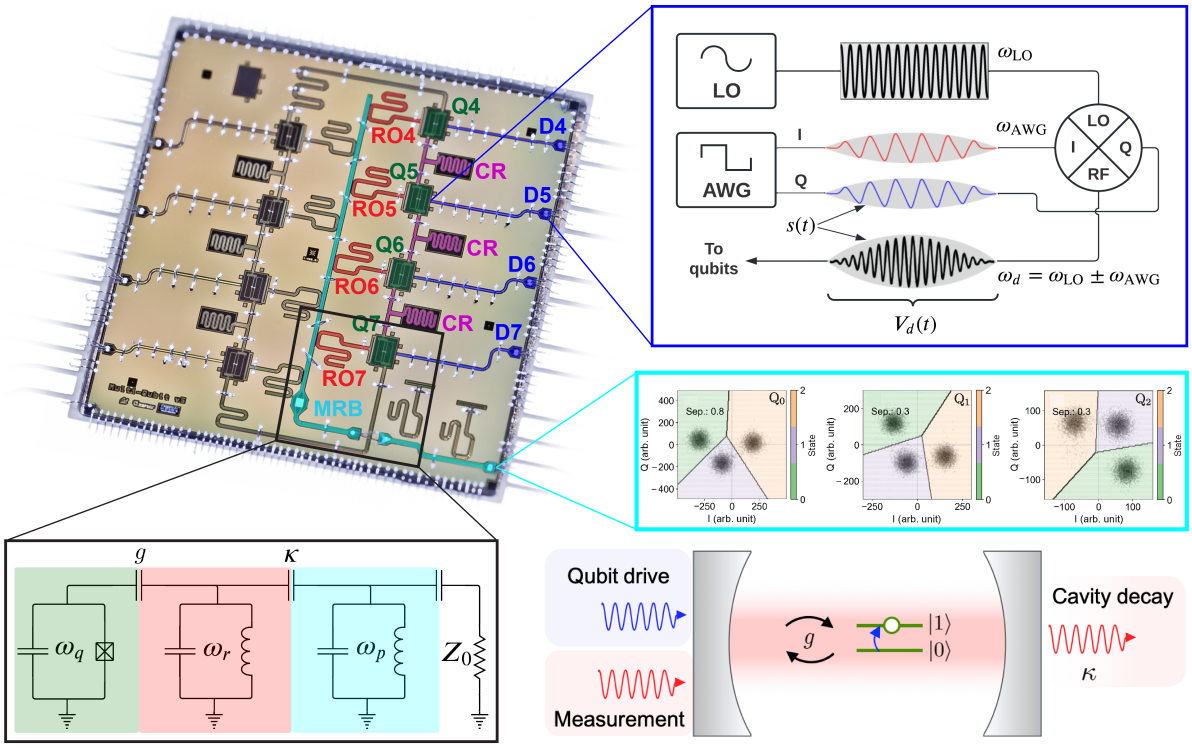


Figure 2.16: **Superconducting quantum processor.** (Top left) False-colored micrograph of our eight-qubit superconducting quantum processor. In this work, we used four transmon qubits (green) with independent microwave control lines (blue). Two-qubit cross-resonance or controlled-Z gates are mediated by coupling resonators (CR, purple) between nearest-neighbors. The qubits are simultaneously measured via dispersive coupling to independent readout resonators (RO, red) coupled to a multiplexed readout bus (MRB, cyan). (Top right) Single- and two-qubit gates are implemented down the microwave control lines (see Fig. 2.11b). (Bottom left) The circuit QED readout circuitry is analogous to cavity QED (bottom right), in which the qubit and resonator cavity are coupled at a rate g. The cavity decays at a rate κ, which can be modified by means of a Purcell filter at frequency ω_p. (Middle right) Multiplexed readout of all qubits (or qutrits) is performed at the end of a sequence of gates in order to classify the measured state of the qubits. (Cavity QED figure courtesy of Noah Stevenson.)

2. The non-linearity of the JJ results in an anharmonic potential term in the qubit Hamiltonian, breaking the degeneracy of the energy spacings of the harmonic potential (see 2.5b). As such, initializing the state of the qubit is trivial, either by waiting long enough for the qubit to relax into its ground state, or by driving a transition into its excited state using microwave photons at the $|0\rangle \rightarrow |1\rangle$ transition frequency.

3. Quantum coherence is necessary in order to perform quantum computations. The de-

coherence time of a superconducting qubit is affected by numerous factors, including isolation from the surrounding environment, the quality of the JJ, etc. Typical coherence times of superconducting transmon qubits are of order $\mathcal{O}(100)$ μs, which is much longer than the time is takes to perform an operation on the qubit, which is typically of order $\mathcal{O}(10)$ – $\mathcal{O}(100)$ ns.

4. The implementation of single-qubit gates is achieved by sending microwave pulses down individual qubit control lines. We can obtain universal control over a qubit with control over only two of the three rotation axes in an experiment using the $ZXZXZ$-decomposition of single-qubit gates. In addition, it is also necessary to perform a two-qubit operation (e.g. controlled-NOT gate), in which the state of one qubit depends on the state of another qubit. This is achieved by the quantum entanglement of two superconducting qubits, which is easily incorporated onto qubit chips by fabricating coupling resonators between two or more qubits. Various methods exist for generating entanglement between two transmons qubits, including the cross-resonance effect, as well as the hardware-efficient controlled-Z gate described in this work. Together, arbitrary single-qubit unitary operations and an entangling gate constitute a universal set of quantum gates.

5. Transmon qubits can be measured via dispersive coupling to a capacitively-coupled readout resonator. Other superconducting qubit archetypes have their own unique readout circuitry.

JJ-based qubits satisfy all of the above criteria for building a quantum computer. Given that JJs are a relatively simple device to fabricate in a laboratory, superconducting qubits are the leading candidate for scaling the size of a quantum computers. This is only made possible by the Josephson effect.

There are two remaining criteria of DiVincenzo that have to do with building a network for quantum communication. These involve being able to convert stationary qubits to transmittable qubits, and being about to faithfully transmit quantum information between specified locations. While research is currently underway for building quantum networks, distributed entanglement, modular computing, microwave to optical transducers, etc., it remains an open question as to whether these criteria can be satisfied by superconducting systems.

Chapter 3

Errors and Noise in Quantum Computations

Qubits in the NISQ era are susceptible to a variety of different error and noise processes. These processes can be broadly placed into two different categories: coherent errors and incoherent noise. Here, the notion of quantum "coherence" relates to how well a process preserves quantum information. Furthermore, we generally distinguish *errors* from *noise*, as errors tend to be systematic processes, whereas noise is – by definition – random or stochastic. To understand how errors and noise impact quantum states (and thus quantum computation), we must first familiarize ourselves with the density matrix formalism of quantum mechanics, as this is necessary for the treatment of open (i.e. real) quantum systems.

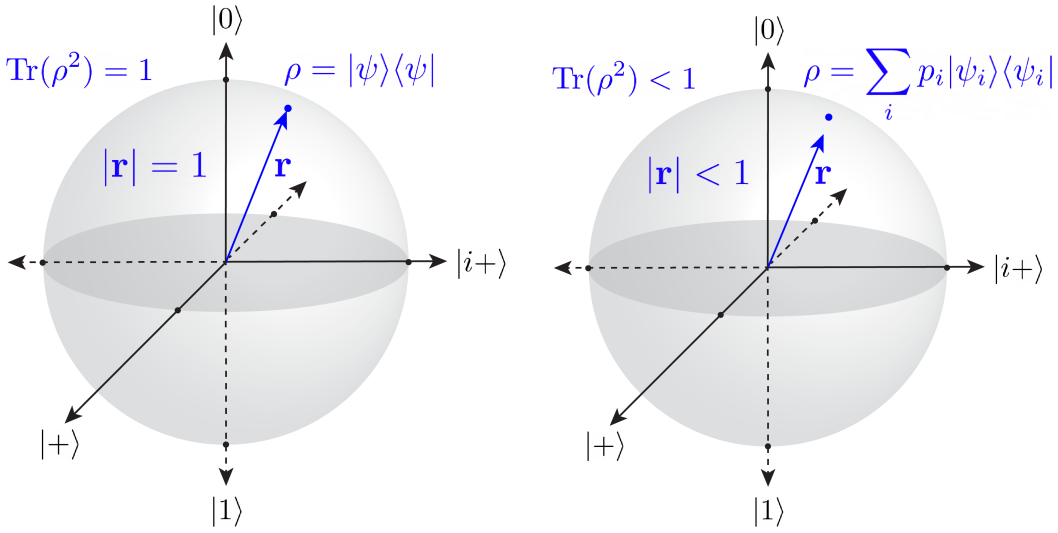
3.1 Density Matrix Formalism

Quantum mechanics is typically taught using the wavefunction formalism. While this is valid and useful in a theoretical context in which one can consider quantum states in isolation, this formalism is typically inadequate for actual quantum systems (i.e. those measured in a lab), which are always to some extent "open" to the rest of the world. To treat "open quantum systems," it is necessary to introduce the density matrix formalism of quantum mechanics, which will facilitate our discussion of errors and noise in quantum computers in this chapter.

The density matrix ρ is a generalized linear operator that defines the complete state of a physical quantum system. If the quantum state is "pure" (see Fig. 3.1a), then the density matrix can be written as the outer product of a state vector $|\psi\rangle$ with itself:

$$\rho = |\psi\rangle\langle\psi| \,. \tag{3.1}$$

However, more generally, for a "mixed" quantum state (see Fig. 3.1b) the density matrix is



(a) Pure state (b) Mixed state

Figure 3.1: **Density matrix formalism.** (a) A pure state is one whose Bloch vector has unity length, for which $\mathrm{Tr}(\rho^2) = 1$. (b) A mixed state is one whose Bloch vector has length less than unity, for which $\mathrm{Tr}(\rho^2) < 1$.

written as a sum over pure states,

$$\boxed{\rho = \sum_i p_i \, |\psi_i\rangle\langle\psi_i|}, \tag{3.2}$$

where p_i is the probability of measuring ρ in the state $|\psi_i\rangle$. The decomposition of ρ into pure states and their respective weights is not unique, nor is there any requirement that the set of states $\{|\psi_i\rangle\}$ be orthonormal.

There are several important properties of density matrices:

1. Hermiticity: $\rho = \rho^\dagger$
2. Positivity: $\rho \geq 0$
3. Normalization: $\mathrm{Tr}(\rho) = 1$
4. Purity: $\mathrm{Tr}(\rho^2) \leq 1$
5. Idempotency: $\rho = \rho^2$, i.f.f. $\mathrm{Tr}(\rho^2) = 1$

Property (1) follows immediately from the definition of the density operator (Eq. 3.2). Property (2) states that ρ is non-negative definite (or positive semi-definite) and follows from

computing the expectation value for ρ with respect to an arbitrary state $|\phi\rangle$:

$$\langle\phi|\,\rho\,|\phi\rangle = \sum_i p_i \,\langle\phi|\psi_i\rangle\,\langle\psi_i|\phi\rangle = \sum_i p_i |\,\langle\phi|\psi_i\rangle\,|^2 \geq 0. \tag{3.3}$$

Property (3) can be derived by requiring that the basis states $|\psi_i\rangle$ of ρ are normalized, and that all probabilities sum to 1: $\sum_i p_i = 1$. As a reminder, the trace of a matrix is the sum of its diagonal components:

$$\mathrm{Tr}(\rho) = \sum_i \rho_{ii}. \tag{3.4}$$

Property (4) is a statement regarding the purity γ of ρ (see Fig. 3.1), which is defined to be

$$\boxed{\gamma \equiv \mathrm{Tr}\big(\rho^2\big)}. \tag{3.5}$$

If ρ is a pure state (Eq. 3.1), then $\gamma = 1$; if ρ is a mixed state (Eq. 3.2), then $\gamma < 1$. For a *maximally mixed state* ρ_{mix} in d dimensions, the purity is

$$\mathrm{Tr}\big(\rho_{\mathrm{mix}}^2\big) = \frac{1}{d} > 0. \tag{3.6}$$

For example, $d = 2$ for a qubit, therefore $\gamma_{\mathrm{mix}} = \frac{1}{2}$. Property (5) follows from property (4) for pure states; in other words, ρ is idempotent if and only if (i.f.f.) it is pure. This can also be understood by stating that if ρ is a projection operator, then it must represent a pure state.

If the complete density matrix of a system is known, the expectation value of an operator A can easily be calculated:

$$\boxed{\langle A\rangle = \mathrm{Tr}(\rho A)}. \tag{3.7}$$

This follows directly from expectation value of A with respect to a pure state $|\psi\rangle$ and Eq. 3.1:

$$\begin{aligned}
\langle A\rangle &= \langle\psi|\,A\,|\psi\rangle \\
&= \sum_n \langle\psi|\,A\,|n\rangle\,\langle n|\psi\rangle \\
&= \sum_n \langle n|\psi\rangle\,\langle\psi|\,A\,|n\rangle \\
&= \mathrm{Tr}(|\psi\rangle\langle\psi|\,A) \\
&= \mathrm{Tr}(\rho A),
\end{aligned}$$

where in line 2 we insert the resolution of the identity (or completeness relation) $\sum_n |n\rangle\,\langle n| = \mathbb{I}$ of an orthonormal basis set $\{|n\rangle\}$, and in line 3 we make use of the commutativity property of traces of inner products of matrices: $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ for matrices A and B; this is true even if $AB \neq BA$.

The Bloch sphere representation of two-level systems is useful for understanding the properties of density matrices (Fig. 2.10). For an arbitrary pure state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, we can write the density matrix as

$$\rho = \begin{pmatrix} |\alpha|^2 & \alpha\beta^* \\ \alpha^*\beta & |\beta|^2 \end{pmatrix}. \tag{3.8}$$

Using of the spherical parametrization of arbitrary single-qubit quantum states (Eq. 2.52), we can write out the explicit form of the density matrix on the Bloch sphere:

$$\rho = \begin{pmatrix} \cos^2(\frac{\theta}{2}) & e^{-i\phi}\cos(\frac{\theta}{2})\sin(\frac{\theta}{2}) \\ e^{i\phi}\cos(\frac{\theta}{2})\sin(\frac{\theta}{2}) & \sin^2(\frac{\theta}{2}) \end{pmatrix} \tag{3.9}$$

Using trigonometric half-angle formulas and simplifying the above expression, we can write this as

$$\rho = \frac{1}{2}\left(\mathbb{I} + \sin\theta\cos\phi\sigma_x + \sin\theta\sin\phi\sigma_y + \cos\theta\sigma_z\right). \tag{3.10}$$

By noting that $\mathbf{r} = r_x\hat{\mathbf{x}} + r_y\hat{\mathbf{y}} + r_z\hat{\mathbf{z}} = \sin\theta\cos\phi\hat{\mathbf{x}} + \sin\theta\sin\phi\hat{\mathbf{y}} + \cos\theta\hat{\mathbf{z}}$ is a unit vector in spherical coordinates and $\boldsymbol{\sigma} = \sigma_x\hat{\mathbf{x}} + \sigma_y\hat{\mathbf{y}} + \sigma_z\hat{\mathbf{z}}$ is the Pauli vector, we can simplify the above express and write an arbitrary quantum state on the Bloch sphere as

$$\boxed{\rho = \frac{1}{2}\left(\mathbb{I} + \mathbf{r}\cdot\boldsymbol{\sigma}\right)}. \tag{3.11}$$

Computing the expectation value of $\boldsymbol{\sigma}$, we find that

$$\langle\boldsymbol{\sigma}\rangle = \mathrm{Tr}(\rho\boldsymbol{\sigma}) = \mathbf{r}. \tag{3.12}$$

Therefore, the Bloch vector $\mathbf{r}$ contains all of the information about ρ, including the purity of the state. To see this, we note that $\mathrm{Tr}(\rho^2) = 1$ for a pure state implies that $|\mathbf{r}| = 1$, and $\mathrm{Tr}(\rho^2) < 1$ for a mixed state implies that $|\mathbf{r}| < 1$.

3.2 Coherent Errors

The notion of a "coherent" quantum state is one in which quantum information is perfectly preserved. In the Bloch sphere representation, a perfectly coherent state is one in which the Bloch vector has unit length (i.e. lies on the surface of the Bloch sphere). Unitary rotation operators in the form of Eq. 2.63 rotate a quantum state $|\psi_0\rangle$ into a new quantum state $|\psi\rangle$,

$$|\psi\rangle = U(\hat{\mathbf{n}},\theta)|\psi_0\rangle = e^{-\frac{i}{2}\theta\hat{\mathbf{n}}\cdot\boldsymbol{\sigma}}|\psi_0\rangle, \tag{3.13}$$

which changes where the vector is located on the Bloch sphere, but has no effect on the length of the Bloch vector and therefore maintains the coherence of the quantum state.

Unitary (or coherent) *errors* manifest as imperfect or unwanted unitary rotations acting on qubits. This can be modeled as an ideal operator $U(\hat{\mathbf{n}}, \theta)$ followed by an erroneous operator $U(\hat{\mathbf{m}}, \epsilon)$, such that the actual final state is given according to

$$\begin{aligned} |\psi\rangle &= U(\hat{\mathbf{m}}, \epsilon) U(\hat{\mathbf{n}}, \theta) |\psi_0\rangle \\ &= e^{-\frac{i}{2}\epsilon \hat{\mathbf{m}}\cdot\boldsymbol{\sigma}} e^{-\frac{i}{2}\theta \hat{\mathbf{n}}\cdot\boldsymbol{\sigma}} |\psi_0\rangle , \end{aligned} \tag{3.14}$$

where $\hat{\mathbf{m}}$ can be arbitrary relative to $\hat{\mathbf{n}}$ (i.e. there is no requirement that the error acts in the same direction as the intended rotation); see Fig. 3.3a. For example, an error associated with a rotation about the z-axis will have an evolution operator

$$\begin{aligned} U(\hat{\mathbf{z}}, \epsilon) &= \exp\left(-i\frac{\epsilon}{2}\sigma_z\right) \\ &= \cos\left(\tfrac{\epsilon}{2}\right) - i\sigma_z \sin\left(\tfrac{\epsilon}{2}\right) \\ &= \begin{pmatrix} e^{-i\epsilon/2} & 0 \\ 0 & e^{i\epsilon/2} \end{pmatrix} . \end{aligned} \tag{3.15}$$

Because unitary operators maintain quantum coherence, coherent errors map pure states to pure states,

$$\rho_0 = |\psi_0\rangle \langle\psi_0| \;\mapsto\; \rho = U(\hat{\mathbf{m}}, \epsilon) |\psi_0\rangle \langle\psi_0| U(\hat{\mathbf{m}}, \epsilon)^\dagger,$$

and therefore do not result in decoherence. We call this a *purity-preserving* process. Coherent errors typically result from systematic imperfections in qubit control (e.g. detuning and calibration errors), drift in previously-calibration parameters [68], and stray electromagnetic (EM) signals or unwanted couplings on multi-qubit processors, which is broadly termed *crosstalk* [69, 70, 71, 72, 73]. Because unitary operators are reversible, coherent errors are unique to quantum computations and have no classical analog. While coherent errors may at first appear benign, given that they are in theory reversible and do not cause decoherence, in later chapters we will see that they are in fact extremely pernicious to multi-qubit quantum computations, and pose a serious barrier to achieving reliable fault-tolerant quantum error correction.

Crosstalk

The term "crosstalk" is borrowed from electrical engineering to describe the phenomenon in classical electronics in which stray EM signals cause "unwanted coupling between signal paths" [74]. Crosstalk errors describe a wide variety of physical error processes that occur across many different quantum computing systems involving more than one qubit, often resulting in non-local and correlated errors. Ref. [69] gives a rigorous definition of what constitutes a crosstalk error, which we outline below.

Crosstalk errors are characterized by violating two key assumption: spatial locality and independence of operations. In an ideal world, each qubit is perfectly isolated from the

surrounding environment and evolves only according to the intended operations applied to it. But, as we previously discussed at the beginning of the chapter, real quantum systems are always to some extent open to the rest of the world, which includes other surrounding qubits and the operations performed on them. When operations are applied to one qubit, this can impact others qubits nearby. Below, we give more precise definitions of the two key assumptions:

1. Locality: an operation is considered "local" if the physical implementation of the operation does create any correlation between the intended qubit(s) and any other qubit(s).

2. Independence: an operation is considered "independent" if the physical implementation of the operation at any given time does not depend on the operation(s) acting on any other qubit(s) at the same time.

(1) comes from the assumption that gates applied to one or more qubits should act locally on only the qubits involved in the process. If locality is not violated, then it is necessary to consider the contextual independence of an operation (2), which assumes the operation of gates should not depend on the operations being applied to other qubits (unless they are explicitly involved in an entangling operation). If locality is violated, then the independence of an operation is not well defined. If both assumptions are respected, then a quantum processor can be considered "crosstalk-free." The above requirements for a quantum processor to be considered crosstalk-free are platform agnostic.

The most basic form a crosstalk in superconducting quantum processors is stray (or "parasitic") classical EM signals which cause unwanted operations on other qubits. We typically find microwave control line crosstalk between qubits when performing single- and two-qubit gates. The dominant crosstalk is between nearest-neighbor qubits, but non-trivial crosstalk between more distant qubit exists as well. The effects of this type of crosstalk are varied, depending on the couplings between qubits and their relative transitions frequencies. Some of the errors that result from crosstalk are products of local operations (termed "classical crosstalk"), while others involve multi-qubit, entangling operations (termed "quantum crosstalk").

A simple model of crosstalk between two nearest-neighbor coupled qubits (Q1 and Q2, see Fig. 3.2) can be understood as follows: consider the case where we are driving Q1 at Q1's frequency, f_1. In the presence of crosstalk, a drive on either line will produce a field at both qubits. Driving down Q1's line at f_1 produces the following Hamiltonian term:

$$H_1 = (a_1 X + b_1 Y) \otimes I, \tag{3.16}$$

where $a_1 = \cos\phi$, $b_1 = \sin\phi$, and ϕ the phase of the Rabi drive. However, this will also produce an effective driving term down Q2's line at f_1 due to crosstalk:

$$H_2 = (c_1 X + d_1 Y) \otimes (a_2 Z + b_2 I), \tag{3.17}$$

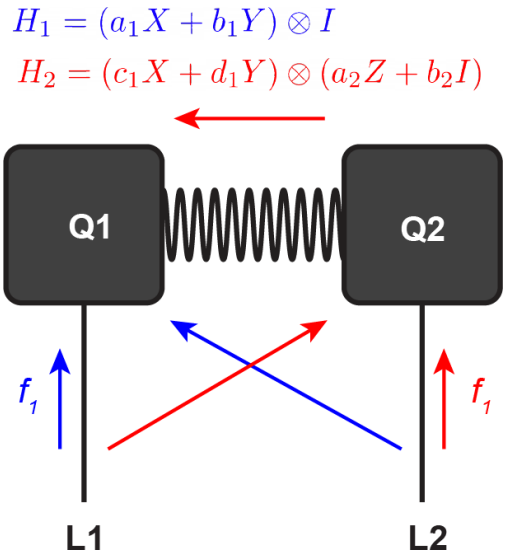


Figure 3.2: **Simple model of crosstalk for two nearest-neighbor coupled qubits.** A drive down Q1's line (blue) at frequency f_1 produces a field at Q2 (red) due to crosstalk. Similarly, a drive down Q2's line (red) at frequency f_1 produces a field at Q1 (blue). The explicit Q1 driving term produces a Hamiltonian $H_1 = (a_1X + b_1Y) \otimes I$. The crosstalk term at Q2 will also pass through the coupling resonator connecting Q2 to Q1 (top red arrow), producing a Hamiltonian of the form $H_2 = (c_1X + d_1Y) \otimes (a_2Z + b_2I)$.

where $c_1 = \cos\phi$ and $d_1 = \sin\phi$. We assume that f_1 is sufficiently far from f_2 such that a pulse down Q2's line at f_1 results in an AC Stark shift on Q2 due to off-resonant driving, and thus a Hamiltonian term of the form $(Z + I)$. The full Hamiltonian is

$$H = H_1 + H_2 = \alpha[(a_1X + b_1Y) \otimes I] + \beta[(c_1X + d_1Y) \otimes (a_2Z + b_2I)], \tag{3.18}$$

where the ratio β/α represents the degree of crosstalk between the two qubits. The Hamiltonian term H_2 can result from microwave line crosstalk on Q2's line when explicitly driving down Q1's line, as stated above, or from directly driving down Q2's line at f_1; the form of the Hamiltonian is the same.

It is common to attempt to remove the conditional Rabi drive term $(X + Y) \otimes Z$ that is produced through Q2 when we drive down Q1's line. We do so by adding an additional single-qubit compensation pulse (with equal amplitude but opposite phase) down Q2's line that cancels this term, thus nulling the field at Q2. This method can be done pairwise for any nearest neighbors, and operates on the assumption that the crosstalk is sufficiently local to few qubits.

Crosstalk can also occur between a pair of qubits involved in a two-qubit gate and idling "spectator" qubits nearby which are not explicitly involved in the gate [75]. Assuming the frequency of the drive of the two-qubit gate is sufficiently far-detuned from the spectator

qubits' frequencies, this can cause unconditional phase errors on the spectator qubits, as well as conditional entangling errors[1]. Both errors can be mitigation to some extent. Unconditional errors can be corrected with virtual phase gates, and conditional errors can be corrected with slow $R_x(2\pi)$ pulses on the spectator qubits to echo away the accumulated phase error, or rotary echos on the qubits involved in the two-qubit gate [76].

3.3 Incoherent Noise

Incoherent processes are typically random or stochastic, as they are not systematic in nature, but rather result from uncontrolled degrees of freedom within (or outside of) the defined system that interact with qubits. This can include external EM fields which cause bit- or phase-flips, $1/f$-type noise from defects in substrates (e.g. charge or flux noise), shot noise due to fluctuations in photon number, noise in qubit parameters (e.g. fluctuations in transition frequencies), thermal "Johnson" noise in the control lines, etc. In contrast to coherent errors, which do not result in decoherence, incoherent noise leads to decoherence and is thus characteristic of being a *purity-decreasing* process.

An important way of characterizing a time-dependent stationary noise source $\hat{\lambda}$ is via its power spectral density (PSD),

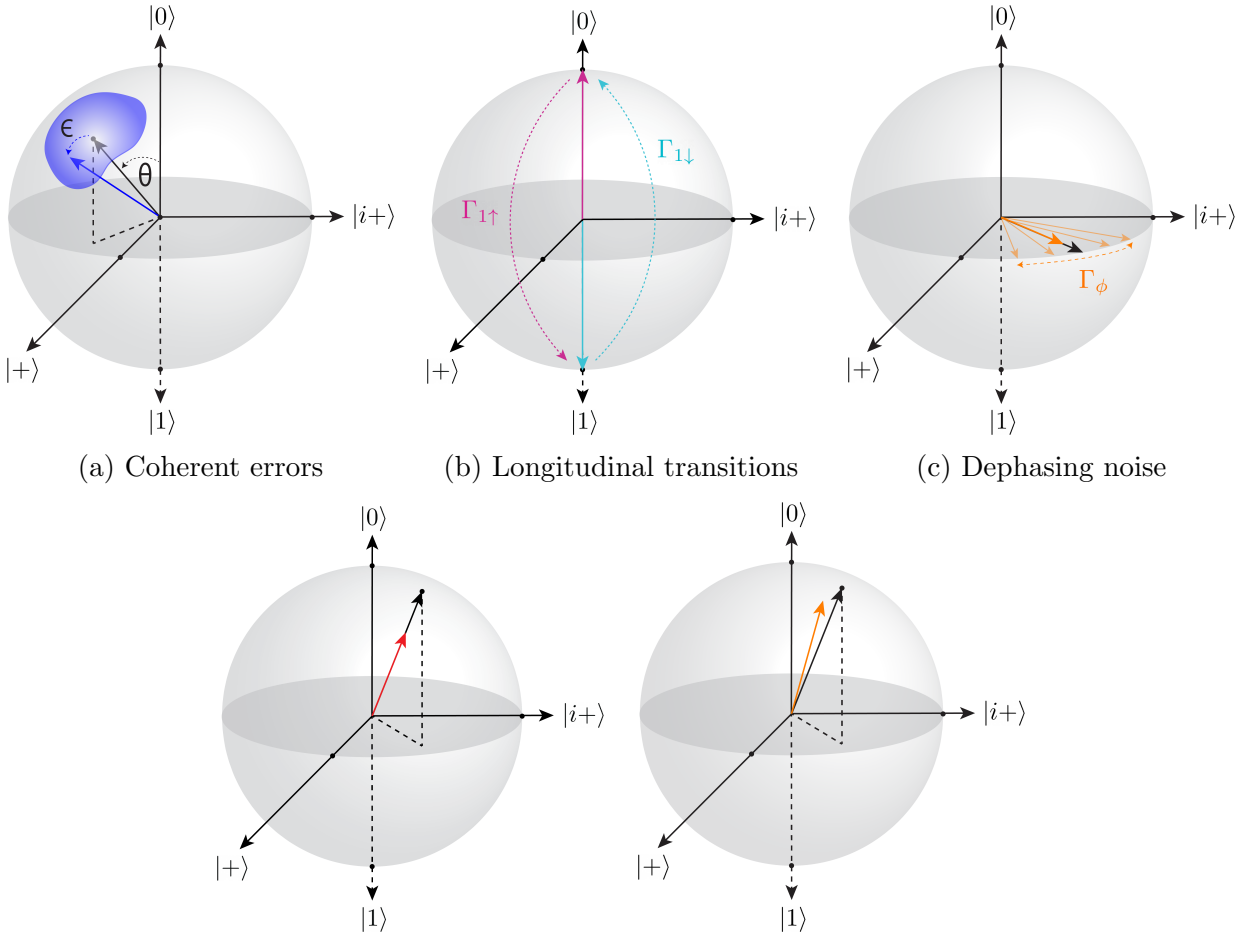
$$S_{\hat{\lambda}\hat{\lambda}}(\omega) = \int_{-\infty}^{\infty} dt' \langle \hat{\lambda}(t')\hat{\lambda}(0)\rangle e^{-i\omega t'}, \tag{3.19}$$

which is the Fourier transform of the autocorrelation function of the noise $\langle \hat{\lambda}(t')\hat{\lambda}(0)\rangle$. This quantity describes the intensity of a noise source at a given frequency ω. For example, $S_{\hat{\lambda}\hat{\lambda}}(\omega_q)$ quantifies the intensity of noise at the qubit frequency ω_q, and allows us to compute the susceptibility of qubit to the noise source $\hat{\lambda}$. We do not go into more detail about this here, but rather focus on ways to visualize, characterize, and model noise in quantum computations. (The interested reader is referred to Ref. [28], which goes into exhaustive detail about the subject of quantum noise and PSDs.) A qubit's susceptibility to noise can be reduced in many ways, such as via materials science and device fabrication for improved qubit design, more stable control electronics and more "hygienic" EM environments, better cryogenic engineering to isolate qubits from the surrounding environment, etc. The goal of all of these methods is to reduce noise levels in order to improve qubit coherence times, which we discuss next.

Decoherence

There are two important characteristic timescales which define the "quantum coherence" of a qubit. The first timescale – called T_1 – describes how long a qubit will remain in an excited

[1] If the qubits are not far-detuned, then this will cause either direct Rabi driving on the spectator qubit via crosstalk, or indirect Rabi driving via the cross-resonance effect. Neighboring qubits are far-detuned by design in order to avoid unintended Rabi effects.



(a) Coherent errors (b) Longitudinal transitions (c) Dephasing noise

(d) Depolarizing noise (e) Stochastic Pauli noise

Figure 3.3: **Errors and noise.** (a) Coherent errors manifest as an unwanted unitary rotation by an angle ϵ (blue arrow) relative to the intended target state (black arrow). The axis of rotation and erroneous rotation angle are generally difficult to predict, denoted by the large blue region around the intended target state. (b) Longitudinal transitions result from spontaneous decay (cyan) or excitation (purple) of a qubit due to transverse noise. (c) Pure dephasing typically results from fluctuations in the qubit ω_{01} transition frequency due to longitudinal noise, such that the qubit Bloch vector along the equator precesses in and out of the rotating frame (light orange), shrinking the Bloch vector at a rate of Γ_ϕ (orange). (d) Depolarizing noise acts isotropically around the Bloch sphere, shrinking the length of the Bloch vector (red) relative to a pure quantum state on the surface of the Bloch sphere (black). (e) Stochastic Pauli noise does not act isotropically, impacting both the length of the Bloch vector and the direction that the vector points (orange) relative to the intended quantum state (black).

state before it decays to the ground state. From T_1, one can quantify a characteristic energy relaxation rate,

$$\boxed{\Gamma_1 \equiv \frac{1}{T_1} = \Gamma_{1\uparrow} + \Gamma_{1\downarrow}}, \tag{3.20}$$

which is composed of the rates of spontaneous excitation $\Gamma_{1\uparrow}$ and decay $\Gamma_{1\downarrow}$ (see Fig. 3.3b). Because superconducting qubits are cooled to temperatures approaching $\lesssim 10$ mK, $\Gamma_{1\uparrow}$ is exponentially suppressed compared to $\Gamma_{1\downarrow}$ by a Boltzmann factor,

$$\Gamma_{1\uparrow} = e^{-\hbar\omega_q/k_\mathrm{B}T}\Gamma_{1\downarrow}, \tag{3.21}$$

where T is the equilibrium temperature, k_B the Boltzmann constant, and ω_q the qubit frequency ($f_q = \omega_q/2\pi$ is typically around 5 GHz); this is known as the *detailed balance* equation. As such, the steady state of a qubit is typically the ground state $|0\rangle$, since $\Gamma_{1\downarrow}$ exponentially governs Γ_1 via Boltzmann statistics. Because spontaneous transitions are related to matrix elements connecting $|0\rangle$ to $|1\rangle$ (and vice-versa), one can connect the relaxation rate to the PSD at the qubit transition frequency via Fermi's Golden Rule:

$$\Gamma_1 = \frac{2\pi}{\hbar^2}|\langle 0|\hat{H}'|1\rangle|^2 S_{\hat{\lambda}\hat{\lambda}}(\omega_q), \tag{3.22}$$

where $\hat{H}'_q = \frac{\delta\hat{H}_q}{\delta\hat{\lambda}}$ is the perturbing Hamiltonian due to fluctuations from a noise source $\hat{\lambda}$. For example, if we take $\hat{H}_q$ to be the qubit Hamiltonian $\hat{H}_q = 4E_C\hat{n}^2 - E_J\cos\hat{\delta}$ and $\hat{\lambda} = \hat{n}$, then $\hat{H}'_q \sim \hat{n} \sim (a - a^\dagger)$, where a ($a^\dagger$) is the lowering (raising) operator[2]. Therefore, $\hat{H}'_q$ only connects off-diagonal matrix elements that differ by a single excitation (e.g. $|0\rangle$ and $|1\rangle$). Γ_1 is caused by noise that is transversal to the qubit's *longitudinal* quantization axis $\hat{z}$, and is often called the *longitudinal* relaxation rate because it causes depolarization along the qubit's quantization axis.

The second characteristic timescale of a qubit – called T_2 – describes how long a qubit can maintain phase coherence in a superposition state. From T_2, we can define a *transverse* relaxation rate

$$\boxed{\Gamma_2 \equiv \frac{1}{T_2} = \frac{\Gamma_1}{2} + \Gamma_\phi}, \tag{3.23}$$

which includes the pure dephasing rate Γ_ϕ (see Fig. 3.3c). As we can see, Γ_2 is due to both transverse noise causing T_1 decay, which completely erases all phase knowledge of the qubit state, and longitudinal noise causing pure dephasing, which is due to stochastic fluctuations in the qubit transition frequency ω_q such that the qubit precesses in and out of the rotating frame. Pure dephasing will eventually lead to complete loss of phase coherence of a superposition state, completely depolarizing of the Bloch vector along the equator. If $\Gamma_\phi = 0$, then T_1 intrinsically limits the maximum length of T_2 via $T_2 = 2T_1$.

[2]As a reminder, $\hat{\delta}$ and $\hat{n}$ are canonically conjugate variables which take the place of $\hat{x}$ and $\hat{p}$, respectively, in our qubit Hamiltonian. For the quantum harmonic oscillator, $\hat{x} = x_\mathrm{ZPF}(a + a^\dagger)$ and $\hat{p} = ip_\mathrm{ZPF}(a - a^\dagger)$, where ZPF denotes the zero point fluctuation uncertainty of these variables.

In a two-level system, T_1 and T_2 define the $1/e$ time for energy relaxation and phase decoherence, respectively. We can modify the ideal single-qubit density matrix (Eq. 3.8) to account for these decoherence factors:

$$\rho = \begin{pmatrix} 1 + (|\alpha|^2 - 1)e^{-t/T_1} & \alpha\beta^* e^{-t/T_2} \\ \alpha^*\beta e^{-t/T_2} & |\beta|^2 e^{-t/T_1} \end{pmatrix}. \tag{3.24}$$

This is known as the Bloch-Redfield model of two-level systems [77]. As we can see, in the limit that $t \longrightarrow \infty$, all matrix elements in ρ tend to zero except the upper left hand entry, which approaches 1. This reflects the fact that an undriven systems with initial $|\beta|^2 \neq 0$ will eventually undergo decoherence and relax to the ground state, assuming that the temperature is low enough that random thermal excitations are sufficiently suppressed.

Stochastic Noise

Stochastic noise describes a general class of processes resulting in random, uncontrolled dynamics in the system. The underlying mechanics of the process can be unitary, but noise manifests as the average of different random unitaries, resulting in purity decreasing processes that are characteristic of non-unitary operations. For example, pure dephasing is a process which replaces a quantum state $\rho \mapsto Z\rho Z$ with some probability p. Because Pauli operators are both Hermitian and unitary, dephasing is in theory reversible by application of a correction unitary operator. However, in practice, one cannot know how the qubit frequency will fluctuate, as these fluctuations are due to uncontrolled degrees of freedom either within the system or outside of it (i.e. unwanted couplings with the surrounding environment). In contrast, incoherent processes such as energy relaxation are *fundamentally* irreversible, and are therefore non-unitary. So, what differentiates unitary operations from stochastic noise which results from unitary operations is that the latter is due to the linear combination of many different unitary evolutions[3]. We can summarize this by stating that under a stochastic noise process, the quantum state ρ undergoes a transformation given by a convex mixture of unitaries [78]:

$$\rho \mapsto \sum_i p_i U_i \rho U_i, \tag{3.25}$$

where $\sum_i p_i = 1$. We note that the dynamics of this process is *not* generated by a linear combination of the Hamiltonians H_i which generate the unitary operators; rather, the dynamics of this process is a linear combination of unitary evolutions resulting from each Hamiltonian H_i separately. In the proceeding sections, we will describe various types of noise in quantum computations and different methods of representing these processes.

[3]By expanding the unitary operators up to second order, Ref. [78] shows how one can derive the non-unitary dynamics in the Linblad Master Equation.

3.4 Quantum Operations

Quantum operations are defined to be a linear transformation $\rho \mapsto \rho'$ that maps a density matrix $\rho \in \mathcal{H}$ in some Hilbert space $\mathcal{H}$ to a new density matrix $\rho' \in \mathcal{H}$ in the same Hilbert space. Quantum operations representing real, physical processes must satisfy two basic constraints:

1. Complete Positivity (CP): For any arbitrary initial state ρ, the final state ρ' must have non-negative probabilities for measuring ρ' in the eigenstate of any observable. This requirement is analogous to the positivity property of density matrices ($\rho \geq 0$).

2. Trace Preservation (TP): For any map $\rho \mapsto \rho'$, the total probability must be conserved. This requirement is analogous to the normalization property of density matrices [$\mathrm{Tr}(\rho) = 1$].

In order for a quantum operation to be physical and real, it must satisfy both the CP and TP constraints. Therefore, physical processes are called CPTP maps. Depending on how we define our system of interest, it may be necessary to relax the trace preservation constraint. For example, leakage out of the qubit subspace to higher energy levels is non-trace preserving. In this case, we instead require that total probability must not *increase*.

Quantum operations taking $\rho \mapsto \rho'$ are often called quantum *maps* or *channels*, which, as we will see in the proceeding sections, are often denoted with different symbols such as $\mathcal{E}$ or Λ. For example, unitary transformations are maps taking $\rho \mapsto U\rho U^\dagger = \mathcal{E}(\rho)$; here, $\mathcal{E}$ denotes the map and $\rho' = \mathcal{E}(\rho)$ the final state.

3.5 The Kraus Operator Formalism

Various representations of quantum operations exist [79], each of which has different advantages and disadvantages when modeling error processes. The most basic representation is the Kraus representation, also called the operator-sum representation [26]. We already encountered an example of this in the previous section in the form of a unitary transformation $\mathcal{E}(\rho) = U\rho U^\dagger$. However, in order to justify the use of this representation for capturing general quantum processes in open quantum systems, we must describe the dynamics of open quantum systems as arising from interactions between the quantum system of interest ρ and the surrounding environment ρ_{env}. If we take ρ_{env} to represent the state of the rest of the universe outside of the system of interest, then together $\rho_{\text{uni}} = \rho \otimes \rho_{\text{env}}$ represents the closed state of the entire universe. In a closed quantum system, all quantum dynamics can be described by unitary transformations. Therefore, interactions between ρ and ρ_{env} can be represented by a unitary operator U,

$$\mathcal{E}(\rho_{\text{uni}}) = U\left(\rho \otimes \rho_{\text{env}}\right) U^\dagger, \tag{3.26}$$

where U couples ρ to ρ_{env}, and vice-versa. However, we are often concerned with only ρ after the interaction and want to ignore the state of ρ_{env}. To do so, we introduce a set of

orthonormal basis states of the environment $\{|e_i\rangle\}$ and, without loss of generality, define the initial state of the environment as a pure state, $\rho_{\text{env}} = |e_0\rangle\langle e_0|$. By taking the partial trace of Eq. 3.26 over $\{|e_i\rangle\}$, we are left with a reduced density matrix of only the system of interest:

$$\mathcal{E}(\rho) = \text{Tr}_{\text{env}}\left[U\left(\rho \otimes \rho_{\text{env}}\right)U^\dagger\right] \tag{3.27}$$

$$= \sum_i \langle e_i| U\left(\rho \otimes |e_0\rangle\langle e_0|\right)U^\dagger |e_i\rangle\,. \tag{3.28}$$

If we define $K_i \equiv \langle e_i| U |e_0\rangle$ as a set of operational elements of $\mathcal{E}$ that act on the system of interest, then we may write the above equation as

$$\boxed{\mathcal{E}(\rho) = \sum_i K_i \rho K_i^\dagger}. \tag{3.29}$$

Eq. 3.29 is known as the Kraus (or operator-sum) representation, and the set of operators $\{K_i\}$ are known as Kraus operators.

The TP constraint for quantum operations is clearly defined in the Kraus representation. Consider a trace-preserving map $\rho \mapsto \mathcal{E}(\rho)$ for which $\text{Tr}(\rho) = \text{Tr}[\mathcal{E}(\rho)] = 1$. Using the above definition of the Kraus representation, we can write this as

$$\begin{aligned}\text{Tr}[\mathcal{E}(\rho)] &= \text{Tr}\left(\sum_i K_i \rho K_i^\dagger\right)\\ &= \text{Tr}\left(\sum_i K_i^\dagger K_i \rho\right)\\ &= 1.\end{aligned}$$

Since ρ is arbitrary, we see that for trace-preserving processes

$$\sum_i K_i^\dagger K_i = \mathbb{I}. \tag{3.30}$$

This is known as the *completeness relation*, and is analogous to the resolution of the identity for orthonormal bases. As outlined in the previous section, we generally relax the TP constraint for the less stringent requirement that the total probability of a quantum map simply cannot increase. This equates to the requirement that $\sum_i K_i^\dagger K_i \leq \mathbb{I}$ for non-TP processes such as leakage, in which information is lost outside of the qubit subspace. Another trivial example of a non-TP process is projective measurement. Suppose we want to measure our qubit ρ in the computational basis; in this case our Kraus operators are simply projectors $K_0 = |0\rangle\langle 0|$ and $K_1 = |1\rangle\langle 1|$, and the Kraus maps are $\mathcal{E}_0(\rho) = |0\rangle\langle 0|\,\rho\,|0\rangle\langle 0|$ and $\mathcal{E}_1(\rho) = |1\rangle\langle 1|\,\rho\,|1\rangle\langle 1|$ with respective probabilities $\text{Tr}[\mathcal{E}_0(\rho)]$ and $\text{Tr}[\mathcal{E}_1(\rho)]$ for measuring the qubit in $|0\rangle$ or $|1\rangle$. Because we can only measure our qubit in *either* $|0\rangle$ or $|1\rangle$, if there is a

finite probability of measuring either state then it must be true that $\text{Tr}[\mathcal{E}_0(\rho)] < 1$ and $\text{Tr}[\mathcal{E}_1(\rho)] < 1$. Therefore, our relaxed TP constraint states that for a real, physical process $\text{Tr}[\mathcal{E}(\rho)] \leq 1$, where $\text{Tr}[\mathcal{E}(\rho)] < 1$ denotes a non-TP process. Non-TP maps can intuitively be understood as those in which $\mathcal{E}$ does not provide a complete description of the full quantum process (i.e. non-deterministic processes).

While Eq. 3.29 is the general form of the Kraus representation, Kraus operators are not unique and it is often useful to express them in a specific basis. For example, in the Pauli basis we can write Eq. 3.29 as

$$\mathcal{E}(\rho) = \sum_{P \in \mathbb{P}^{\otimes n}} p_P P \rho P^\dagger \tag{3.31}$$

where the Kraus operators are $K = \sqrt{p_P} P$ with p_P the relative probability of an error due to P, and $\mathbb{P}^{\otimes n} = \{I, X, Y, Z\}^{\otimes n} = \mathbb{P}_n$ the set of 4^n generalized Pauli operators. Kraus maps in this form are often called *Pauli channels*.

Examples

In this section, we provide some examples of Kraus maps that capture common errors and noise in quantum computations. However, these examples are by no mean comprehensive, and generalized Kraus operators are not required to be unitary, Hermitian, or invertible.

- *Coherent error* by an angle θ:

$$\mathcal{E}(\rho) = U \rho U^\dagger = e^{-\frac{i}{2}\theta \hat{\mathbf{n}} \cdot \boldsymbol{\sigma}} \rho e^{\frac{i}{2}\theta \hat{\mathbf{n}} \cdot \boldsymbol{\sigma}}, \tag{3.32}$$

  with Kraus operator $K = U = e^{-\frac{i}{2}\theta \hat{\mathbf{n}} \cdot \boldsymbol{\sigma}}$; see Fig. 3.3a. In general, the Kraus representation for a unitary channel only contains one Kraus operator, and is thus not a sum over many Kraus operators. This distinction between one versus many Kraus operators is related to whether or not the Kraus map is purity-preserving.

- *Spontaneous emission* and *amplitude damping* with probability p:

$$\mathcal{E}(\rho) = K_0 \rho K_0^\dagger + K_1 \rho K_1^\dagger, \tag{3.33}$$

  with Kraus operators $K_0 = \sqrt{I - p\sigma^+\sigma^-} = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1-p} \end{pmatrix}$, $K_1 = \sqrt{p}\sigma^- = \begin{pmatrix} 0 & \sqrt{p} \\ 0 & 0 \end{pmatrix}$, and where $\sigma^+ = |1\rangle\langle 0|$ and $\sigma^- = |0\rangle\langle 1|$. Spontaneous emission describes the complete energy loss of a quantum system ρ in an excited state due to the emission of a photon (i.e. T_1 decay), with K_1 mapping $|1\rangle \mapsto |0\rangle$; see Fig. 3.3b. Amplitude damping is related to spontaneous emission and describes the relative change in the amplitude of $|1\rangle$ with respect to $|0\rangle$ when no spontaneous emission event occurs, with K_0 reducing the amplitude of $|1\rangle$ but leaving $|0\rangle$ unchanged. Spontaneous emission is an example of a non-unital channel, in which $\mathcal{E}(I) \neq I$.

- *Dephasing noise* with probability p:

$$\mathcal{E}(\rho) = \left(1 - \frac{p}{2}\right)\rho + \frac{1}{2}Z\rho Z, \tag{3.34}$$

with Kraus operators $K_0 = \sqrt{1-p/2}I$ and $K_1 = \sqrt{p/2}Z$; see Fig. 3.3c. Dephasing noise describes the process in which longitudinal noise acting along the qubit's quantization axis due to unwanted coupling to the environment results in fluctuations of the ω_{01} transition frequency, which causes the qubit's Bloch vector to precess in and out of the rotating frame. Dephasing noise is a phase-flip channel in which a quantum state ρ experiences complete loss of phase coherence with probability p, and is unchanged with probability $1-p$.

- *Depolarizing noise* with probability p:

$$\mathcal{E}(\rho) = \left(1 - \frac{3p}{4}\right)\rho + \frac{p}{4}(X\rho X + Y\rho Y + Z\rho Z), \tag{3.35}$$

with Kraus operators $K_0 = \sqrt{1-p/4}I$, $K_1 = \sqrt{p/4}X$, $K_2 = \sqrt{p/4}Y$, and $K_3 = \sqrt{p/4}Z$. Depolarizing noise describes the process in which a quantum state ρ is replaced by a completely mixed state with probability p, representing the complete loss of quantum information, and is unchanged with probability $1-p$; see Fig. 3.3d.

- *Stochastic Pauli noise* with probabilities p_X, p_Y, and p_Z:

$$\mathcal{E}(\rho) = (1 - p_X - p_Y - p_Z)I\rho I + p_X X\rho X + p_Y Y\rho Y + p_Z Z\rho Z, \tag{3.36}$$

with Kraus operators $K_0 = \sqrt{1-p_X-p_Y-p_Z}I$, $K_1 = \sqrt{p_X}X$, $K_2 = \sqrt{p_Y}Y$, and $K_3 = \sqrt{p_Z}Z$, and where $p_X, p_Y, p_Z \geq 0$ and $p_X + p_Y + p_Z \leq 1$. Stochastic Pauli noise is a generalized form of depolarizing noise which does not act isotropically around the Bloch sphere, thus X-, Y-, and Z-type errors have differing probabilities p_X, p_Y, and p_Z, respectively; see Fig. 3.3e.

3.6 The χ Matrix Superoperator Formalism

As previously stated, it is often convenient to represent Kraus operators in the Pauli basis. While Eq. 3.31 is the Kraus representation for Pauli operators, we can also expand any arbitrary Kraus operator K_i in the Pauli basis,

$$K_i = \sum_{i=1}^{d^2} c_{ij} P_j, \tag{3.37}$$

where $P_j \in \mathcal{P}^{\otimes n}$, $d^2 = 2^{2n}$ is the number of Pauli operators for n qubits, and c_{ij} the expansion coefficients. Inserting this into Eq. 3.29, we find

$$\mathcal{E}(\rho) = \sum_{i,j=1}^{d^2} \chi_{ij} P_i \rho P_j^\dagger, \tag{3.38}$$

where $\chi_{ij} = \sum_k c_{ki} c_{kj}^*$ is a $d^2 \times d^2$ complex-valued matrix (i.e. superoperator). χ completely captures the map $\mathcal{E}$ in a single matrix and is often called the *process matrix*[4].

The utility of the χ matrix is that it allows us to determine specific properties of $\mathcal{E}$ with relative ease. For example, if χ is Hermitian and positive semi-definite, then $\mathcal{E}$ is CP. The Hermiticity of χ is governed by the expansion coefficients, with $\chi_{ij} = \sum_k c_{ki} c_{kj}^* = \chi_{ji}^*$, and positive semi-definiteness can be determined for any d^2-dimensional complex vector $\mathbf{v}$: $\mathbf{v}^\dagger \chi \mathbf{v} = \sum_{ij} v_i^* \chi_{ij} v_j = \sum_{ijk} (c_{ki} v_i^*)(c_{kj}^* v_j) = \sum_j |\sum_i c_{ki} v_i^*|^2 \geq 0$. Additionally, using the completeness relation for Kraus maps (Eq. 3.30), we can easily determine the TP constraint for the χ matrix formalism:

$$\begin{aligned} \sum_i K_i^\dagger K_i &= \sum_{i,j,k} c_{ij}^* P_j^\dagger c_{ik} P_k^\dagger \\ &= \sum_{j,k} \left(\sum_i c_{ik} c_{ij}^* \right) P_j^\dagger P_k^\dagger \\ &= \sum_{j,k} \chi_{jk} P_j^\dagger P_k^\dagger = \mathbb{I}. \end{aligned} \tag{3.39}$$

Therefore, if $\mathcal{E}$ is CPTP, then the above completeness relation enforces d^2 constraints, therefore the $d^2 \times d^2$ Hermitian matrix χ has $d^2(d^2-1)$ free parameters.

Examples

Here, we provide the χ matrix of the example Kraus maps from Section 3.5.

- *Coherent error* about the z-axis by an angle θ:

$$\chi = \frac{1}{2} \begin{pmatrix} 1+\cos(\theta) & 0 & 0 & i\sin(\theta) \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -i\sin(\theta) & 0 & 0 & 1-\cos(\theta) \end{pmatrix} \tag{3.40}$$

- *Spontaneous emission* and *amplitude damping* with probability p:

$$\chi = \frac{1}{2} \begin{pmatrix} (1+\sqrt{1-p})^2 & 0 & 0 & p/2 \\ 0 & p/2 & -ip/2 & 0 \\ 0 & ip/2 & p/2 & 0 \\ p/2 & 0 & 0 & (1-\sqrt{1-p})^2 \end{pmatrix} \tag{3.41}$$

[4]We note that this term is also widely used for other superoperator maps, such as the Pauli-transfer matrix (see next section), so we avoid assigning this label specifically for the χ matrix.

- *Dephasing noise* with probability p:

$$\chi = \begin{pmatrix} 1-p/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p/2 \end{pmatrix} \tag{3.42}$$

- *Depolarizing noise* with probability p:

$$\chi = \begin{pmatrix} 1-3p/4 & 0 & 0 & 0 \\ 0 & p/4 & 0 & 0 \\ 0 & 0 & p/4 & 0 \\ 0 & 0 & 0 & p/4 \end{pmatrix} \tag{3.43}$$

- *Stochastic Pauli noise* with probabilities p_X, p_Y, and p_Z:

$$\chi = \begin{pmatrix} 1-p_X-p_Y-p_Z & 0 & 0 & 0 \\ 0 & p_X & 0 & 0 \\ 0 & 0 & p_Y & 0 \\ 0 & 0 & 0 & p_Z \end{pmatrix} \tag{3.44}$$

From the above examples, we see that if the Kraus map is diagonal in the Pauli basis, as it is for dephasing, depolarizing, and stochastic Pauli noise, then the diagonal components of the χ matrix can be determined directly from the probability coefficients in the Kraus representation.

3.7 The Pauli Transfer Matrix Superoperator Formalism

Of the various representations of quantum processes that exist, the Pauli transfer matrix (PTM) superoperator representation is perhaps one of the most useful for visualizing different error channels. The PTM representation is realized expanding any density matrix ρ in the n-qubit Pauli basis $\mathbb{P}^{\otimes n}$,

$$\rho = \sum_{P \in \mathbb{P}^{\otimes n}} \alpha_P P, \tag{3.45}$$

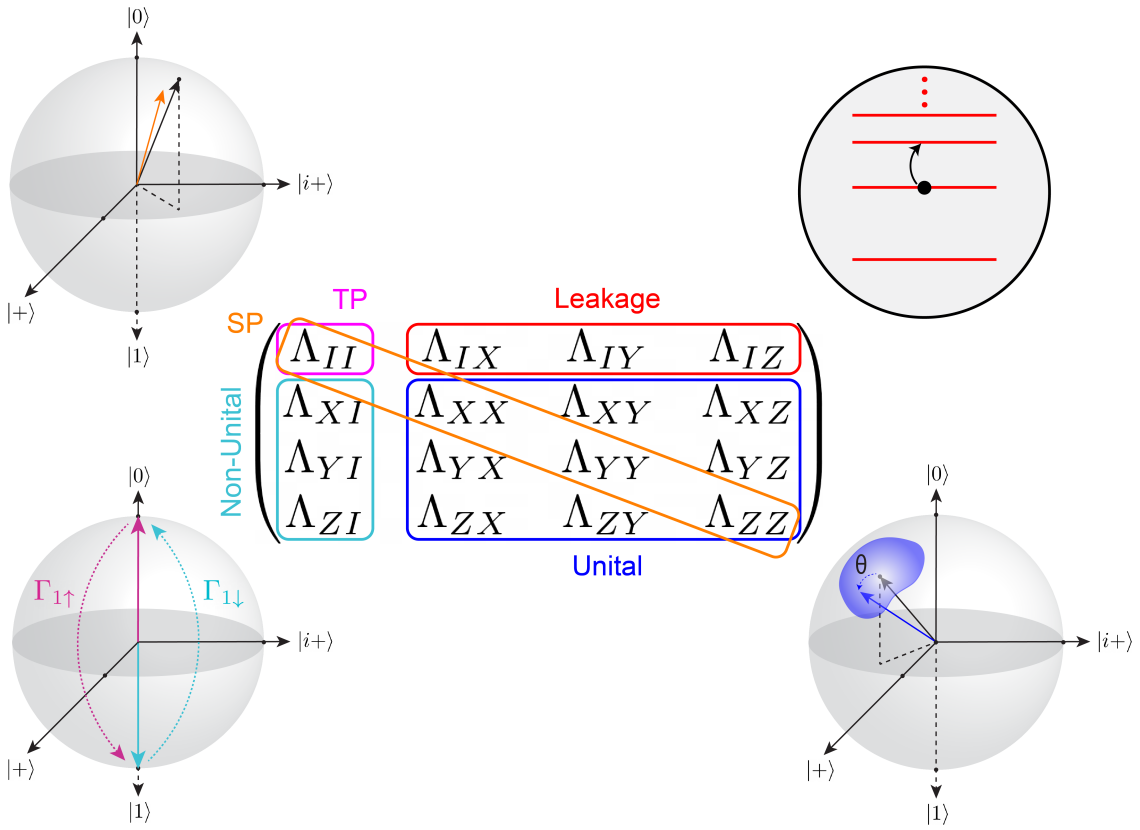


Figure 3.4: **Pauli Transfer Matrix.** The PTM is divided into four blocks: the upper left-hand entry represents trace-preservation (TP, purple). The lower right-hand block captures unital (blue) processes, such as Pauli noise or unitary errors. The column to the left of the unital block captures non-unital (cyan) processes, such as T_1 decay. And the the row above the unital block captures state-dependent leakage (red). The diagonal of the PTM represents state preservation (SP, orange), i.e. how well the state of each Pauli channel is preserved in a process.

where α_P are the expansion coefficients. This is sometimes referred to as a process matrix in the Pauli-Liouville representation. By vectorizing the expansion coefficients, we obtain a vectorized density matrix $|\rho\rangle\rangle = \begin{pmatrix} \alpha_{I^{\otimes n}} & \dots & \alpha_{Z^{\otimes n}} \end{pmatrix}^T$. A linear CPTP map $\rho \mapsto \rho' = \Lambda(\rho)$ in the PTM representation is completely defined by a $d^2 \times d^2 = 4^n \times 4^n$ matrix Λ, with values $\Lambda_{ij} = \mathrm{Tr}\{[P_i \mathcal{E}(P_j)]\}/d$ that can derived directly from the Kraus representation (Eq. 3.31). By definition, all entries in the PTM are real and are bounded by $\Lambda_{ij} \in [-1, 1]$. PTMs have the useful property that the composite map (i.e. the PTM of a quantum circuit) can be constructed by taking the matrix products of the individual maps (i.e. the PTMs of the individual gates).

For a single qubit, the vectorized density matrix is given as

$$|\rho\rangle\rangle = \begin{pmatrix} \alpha_I \\ \alpha_X \\ \alpha_Y \\ \alpha_Z \end{pmatrix}. \tag{3.46}$$

The quantum map $\rho' = \Lambda(\rho)$ can be expressed in vector form, $|\rho'\rangle\rangle = \Lambda\,|\rho\rangle\rangle$, or more explicitly,

$$\begin{pmatrix} \alpha'_I \\ \alpha'_X \\ \alpha'_Y \\ \alpha_Z \end{pmatrix} = \begin{pmatrix} \Lambda_{II} & \Lambda_{IX} & \Lambda_{IY} & \Lambda_{IZ} \\ \Lambda_{XI} & \Lambda_{XX} & \Lambda_{XY} & \Lambda_{XZ} \\ \Lambda_{YI} & \Lambda_{YX} & \Lambda_{YY} & \Lambda_{YZ} \\ \Lambda_{ZI} & \Lambda_{ZX} & \Lambda_{ZY} & \Lambda_{ZZ} \end{pmatrix} \begin{pmatrix} \alpha_I \\ \alpha_X \\ \alpha_Y \\ \alpha_Z \end{pmatrix}. \tag{3.47}$$

As seen in Fig. 3.4, the details of any given process represented by a PTM can be conveniently interpreted directly from the components of the matrix. The PTM can be divided into four blocks: the upper left-hand corner represents trace-preservation, with $\Lambda_{II} = 1$ if a process is trace-preserving (TP). This constraint can be succinctly summarized by stating that a process is TP if $\Lambda_{0j} = \delta_{0j}$ (i.e. the first row of the PTM is $[1, 0, ..., 0]$). The lower right-hand block is the unital block, which captures processes such as stochastic Pauli noise and unitary errors. A unital process is one that maps the identity operation to the identity operation $\Lambda(\mathbb{I}) = \mathbb{I}$ (it cannot purify a mixed state). The row above the unital block captures state-dependent leakage, represented by $\Lambda_{IP} \neq 0$ for $P \in \{X, Y, Z\}$; leakage is therefore not TP. The column to the left of the unital block is the non-unital block, which captures processes such as spontaneous emission (i.e. T_1 decay) or amplitude damping. This constraint can be summarized by stating that process is unital if $\Lambda_{i0} = \delta_{i0}$ (i.e. the first column of the PTM is $[1, 0, ..., 0]^T$). Finally, the diagonal of Λ represents state-preservation, with $\Lambda_{PP} = 1$ (< 1) for processes which (do not) preserve the Pauli channel $P \,\forall\, \{I, X, Y, Z\}$. One downside to the PTM representation is that it does not concisely capture whether or not a process is CP.

The process fidelity of a map in the PTM representation is the weighted average of the diagonal components,

$$F_\Lambda = \frac{1}{4^n} \sum_{P \in \mathbb{P}^{\otimes n}} \Lambda_{PP}. \tag{3.48}$$

Therefore, the process infidelity is given as $e_F = 1 - F_\Lambda$. For any CPTP map, e_F sets the upper bound on the infidelity (or *survival probability*) of any given Pauli channel,

$$0 \leq 1 - \Lambda_{PP} \leq 2e_F. \tag{3.49}$$

Examples

Below, we provide the PTM for the example Kraus maps from Section 3.5:

- *Coherent error* about the z-axis by an angle θ (Eq. 3.15):

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) & 0 \\ 0 & \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{3.50}$$

- *Spontaneous emission* and *amplitude damping* with probability p:

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \sqrt{1-p} & 0 & 0 \\ 0 & 0 & \sqrt{1-p} & 0 \\ p & 0 & 0 & 1-p \end{pmatrix} \tag{3.51}$$

- *Dephasing noise* with probability p:

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-p & 0 & 0 \\ 0 & 0 & 1-p & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{3.52}$$

- *Depolarizing noise* with probability p:

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-p & 0 & 0 \\ 0 & 0 & 1-p & 0 \\ 0 & 0 & 0 & 1-p \end{pmatrix} \tag{3.53}$$

- *Stochastic Pauli noise* with probabilities p_X, p_Y, and p_Z:

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-2(p_Y+p_Z) & 0 & 0 \\ 0 & 0 & 1-2(p_X+p_Z) & 0 \\ 0 & 0 & 0 & 1-2(p_X+p_Y) \end{pmatrix} \tag{3.54}$$

With these examples in hand, we can better understand the state-preservation constraint by considering the duality between quantum maps in the Kraus and PTM representations. Namely, a Pauli error Q in the Kraus representation that occurs with probability p_Q will reduce the value of Λ_{PP} for any anti-commuting Pauli channel (i.e. $\{P, Q\} = 0$). As a concrete example, consider a dephasing channel in the Kraus representation,

$$\mathcal{E}(\rho) = \left(1 - \frac{p}{2}\right)\rho + \frac{p}{2} Z\rho Z,$$

in which a Z error (no error) occurs with probability p $(1-p)$. The corresponding PTM is given in Eq. 3.52, where we see that a Z error in the Kraus representation reduces the survival probability of the X and Y Pauli channels, but leaves the Z Pauli channel unchanged. We can generalize the connection between the Kraus and PTM representations for the diagonal terms in Eq. 3.54:

$$\Lambda_{PP} = \sum_{Q \forall [P,Q]=0} p_Q - \sum_{Q \forall \{P,Q\}=0} p_Q, \tag{3.55}$$

where p_Q is the probability of an error due to the Pauli Q in the Kraus representation.

3.8 Non-Markovian Errors

In our previous discussion of errors and noise processes, we have made the assumption that errors impacting a qubit are local to the affected qubit and have no impact on other qubits. Furthermore, while we have not explicitly discussed the notion of "time" in the Kraus, χ, or PTM representations, the time dependence of the error or noise process is built into the probability parameter p appearing in the above examples, which is taken to be proportional to the time t of a finite quantum operation, $p \sim t$. In all of the above examples, ρ' is completely determined by ρ according to the map $\mathcal{E} : \rho \mapsto \rho'$ or $\Lambda : \rho \mapsto \rho'$, and has no dependence on the previous gates or gates applied to other qubits. Therefore, by definition, all of the error processes we have discussed up until now can be considered "Markovian," as they obey assumptions of locality in space and are temporally restricted to a single quantum operation (i.e. gate). While non-Markovianity is well-defined in the classical regime in terms of temporal correlations, a rigorous definition of what constitutes a "non-Markovian" process in the quantum domain does not exist for all systems, as this definition is dependent upon the system size and timescales under consideration. This is certainly true for NISQ processors, in which the term "non-Markovian error" is often used yet rarely well defined [80, 81, 82, 83]. Furthermore, for infinitely large systems and infinitely small timescales, all processes are Markovian.

For the purposes of this work, we define a non-Markovian error as one that cannot be modeled as occurring within a given cycle of gates, for a system defined to contain a given number of qubits (see Fig. 3.5). We use a "cycle" as the characteristic timescale for Markovian processes, as this term is derived from the notion of "clock cycles" on a quantum computer, which defines the timescale within which discrete operations act on qubits (i.e. quantum

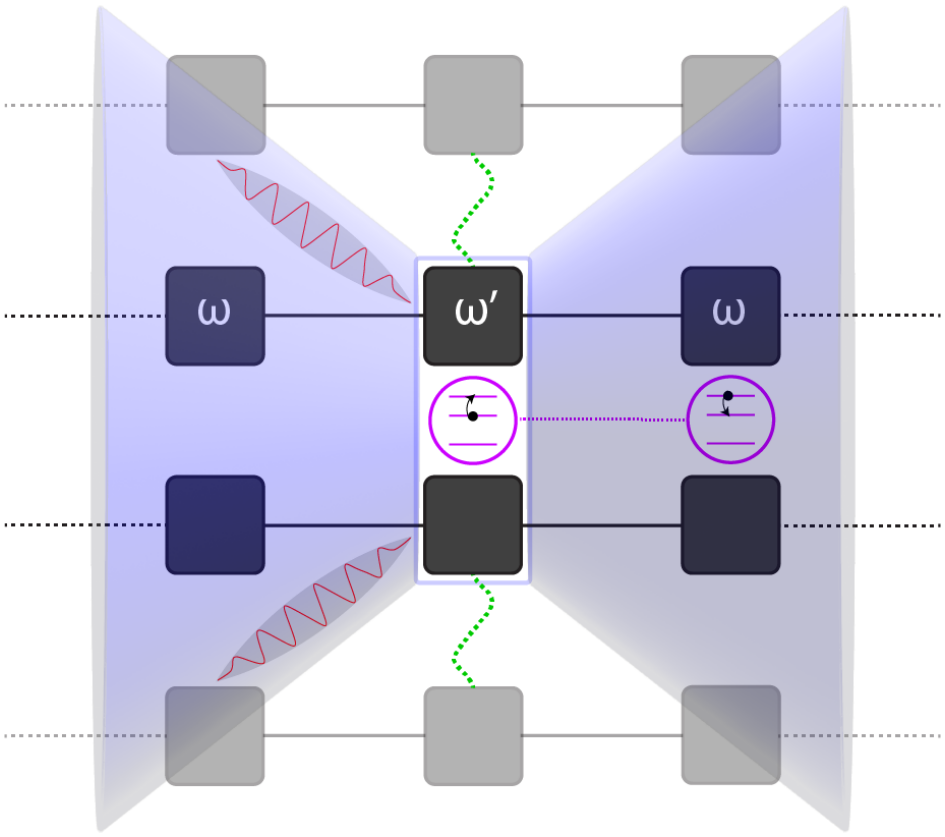


Figure 3.5: **Non-Markovian errors in gate-based quantum computing.** Markovian errors for a given system of qubits (black) are defined to occur within the timescale of a given cycle of gates (blue rectangle). Sources of non-Markovianity include leakage out of (and later seepage back into) the computational basis states (purple), unintended entanglement (green) with qubits outside of the defined system (grey) or two-level fluctuators in the environment, drift in the system properties (e.g. fluctuations in the qubit transition frequency ω, with $\omega' = \omega + \delta\omega$), and classical EM signals from outside of the defined system (red) that reach the system qubits within their pseudo light cone (blue).

gates). Furthermore, the Kraus, χ, and PTM representations of quantum errors and noise are constructed on the assumption that every Markovian error can be captured by a process matrix for any cycle of quantum gates. Therefore, violations of this assumption (i.e. errors that cannot be captured by a process matrix) must – by definition – be non-Markovian. Various common sources of non-Markovianity in NISQ systems include leakage out of the computational basis states [84, 85, 86, 87, 88, 89], unwanted entanglement between qubits (e.g. static ZZ coupling), drift in qubit parameters [68], unwanted coupling to sources of noise in the environment (e.g. two-level fluctuators and nonequilibrium quasiparticles [90, 91, 92, 93]), and qubit heating [94].

Chapter 4

Noise Tailoring via Twirling

The notion of "twirling" a quantum channel is by now ubiquitous in the field of quantum characterization, verification, and validation (QCVV). However, it is by no means well-known in QIS, even though many contemporary methods of benchmarking error rates in quantum gates utilize it. The basic concept is to average a quantum channel Λ over some unitary group, which has the benefit of simplifying the structure of Λ, often taking it from dense to sparse. As we will see in the next chapter, a sparse, diagonal process matrix condenses information about the physical process into a single number – the process fidelity. However, twirling can suffer from the unforeseen consequence of being an inaccurate representation of Λ in any actual quantum application. Below, we give a precise definition of twirling, and discuss the difference between twirling over the Pauli and Clifford gate sets.

4.1 Twirling Quantum Channels

Consider a quantum channel Λ representing some quantum gate or process (e.g. in the PTM representation). Next, consider a unitary operator $\hat{U}$ which belongs to the set $\mathcal{U}(d)$ of all $d \times d$ unitary operators in d-dimensional Hilbert space, where $d = 2^n$ for n qubits. Suppose that Λ is conjugated by $\hat{U}$, which has been randomly sampled from $\mathcal{U}(d)$, mapping $\Lambda \mapsto \hat{U} \circ \Lambda \circ \hat{U}^\dagger$. Using this notation, a twirled channel $\bar{\Lambda}$ is given by

$$\bar{\Lambda} = \int_{\mathcal{U}(d)} d\mu(\hat{U}) \hat{U} \circ \Lambda \circ \hat{U}^\dagger, \tag{4.1}$$

where $\mu(\hat{U})$ is the uniform Haar measure [95] over the space of $d \times d$ unitary matrices, ensuring that $\hat{U}$ is uniformly sampled from $\mathcal{U}(d)$. Thus, $\bar{\Lambda}$ can be thought of as the expected value of Λ conjugated with all possible unitaries $\hat{U} \in \mathcal{U}(d)$. Because $\Lambda = \Lambda(\rho)$ is a linear map on a quantum state ρ, $\hat{U}$ also acts on ρ by conjugation:

$$\hat{U}(\rho) = U \rho U^\dagger, \ \hat{U}^\dagger(\rho) = U^\dagger \rho U. \tag{4.2}$$

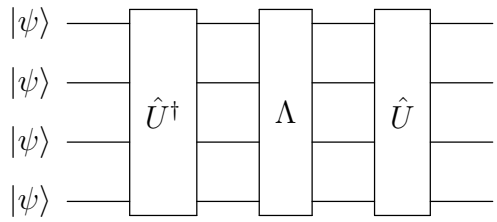


Figure 4.1: **Twirling.** Twirling a quantum channel results in the map $\Lambda \mapsto \hat{U} \circ \Lambda \circ \hat{U}^\dagger$.

Therefore, the twirled channel on a density operator $\bar{\Lambda}(\rho)$ can be written as

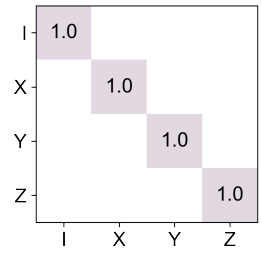
$$\bar{\Lambda}(\rho) = \int_{\mathcal{U}(d)} d\mu(U) U^\dagger \Lambda(U\rho U^\dagger) U. \tag{4.3}$$

It is often the case that one does not actually twirl over all of $\mathcal{U}(d)$ because of the impracticality of uniformly sampling from $\mathcal{U}(d)$. Rather, it is much more common to twirl a channel over a discrete distribution of unitaries, such as the Pauli or Clifford group. For example, consider the channel $\Lambda(\rho) = P\rho Q$, where $\{P, Q\}$ are arbitrary linear operators. Next, consider some group $\mathcal{G} = \{U_1, ..., U_K\}$ consisting of K different unitary operators. In the discrete case, the twirled channel $\bar{\Lambda}(\rho)$ can be written as the weighted average over all K operators [96]:

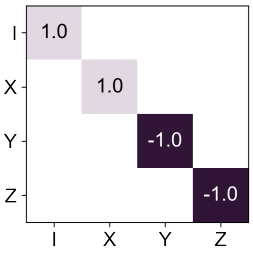
$$\bar{\Lambda}(\rho) = \frac{1}{K} \sum_{k=1}^{K} U_k^\dagger P U_k \rho U_k^\dagger Q U_k. \tag{4.4}$$

4.2 Pauli Twirling

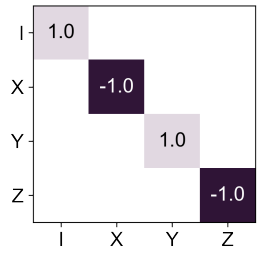
One can twirl a channel Λ over any group. However, it is often convenient to choose a particular group – such as the Pauli or Clifford group. Because we often represent our channels in the Pauli basis, it is educational to first understand the basics of Pauli twirling, before considering twirling over any other unitary group. As we show in Fig. 4.2, Pauli twirling a channel Λ in the PTM representation can be easily understood by considering the PTM for a simple $R_x(10°)$ rotation, which contains off-diagonal terms only in the lower right-hand block of the PTM. Under conjugation with any Pauli $P \in \{I, X, Y, Z\}$ (also represented in PTM form), the signs of the off-diagonal terms remain the same (are reversed) if P commutes (anti-commutes) with Λ. Therefore, for $\Lambda[R_x(10°)]$, the off-diagonal terms remain the same for $I \circ \Lambda[R_x(10°)] \circ I$ or $X \circ \Lambda[R_x(10°)] \circ X$, but are reversed for $Y \circ \Lambda[R_x(10°)] \circ Y$ and $Z \circ \Lambda[R_x(10°)] \circ Z$. Thus, the off-diagonal terms change sign with a 50% probability upon conjugation with a randomly-selected Pauli.



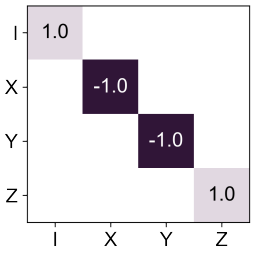
(a) PTM for Pauli I



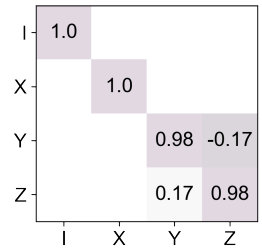
(b) PTM for Pauli X



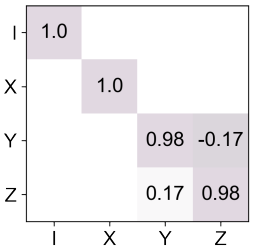
(c) PTM for Pauli Y



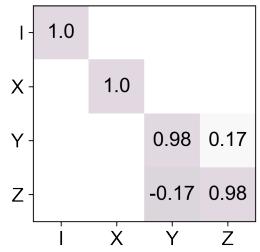
(d) PTM for Pauli Z



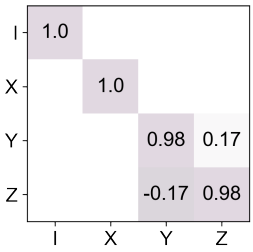
(e) PTM for $R_x(10°)$



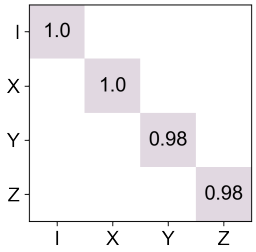
(f) $X \circ \Lambda[R_x(10°)] \circ X$



(g) $Y \circ \Lambda[R_x(10°)] \circ Y$



(h) $Z \circ \Lambda[R_x(10°)] \circ Z$



(i) $\frac{1}{2}\Big(X \circ \Lambda[R_x(10°)] \circ X + Z \circ \Lambda[R_x(10°)] \circ Z\Big)$

Figure 4.2: **Basics of Pauli twirling.** (a) PTM for the Pauli I gate. (b) PTM for the Pauli X gate. (c) PTM for the Pauli Y gate. (d) PTM for the Pauli Z gate. (e) PTM for an $R_x(10°)$ rotation, $\Lambda[R_x(10°)]$. (f) $\Lambda[R_x(10°)]$ conjugated with X gates; because $R_x(10°)$ commutes with X, this is identical to $\Lambda[R_x(10°)]$. (g) $\Lambda[R_x(10°)]$ conjugated with Y gates; because $R_x(10°)$ does not commutes with Y, the signs of the off-diagonal terms have been flipped relative to $\Lambda[R_x(10°)]$. (h) $\Lambda[R_x(10°)]$ conjugated with Z gates; because $R_x(10°)$ does not commutes with Z, the signs of the off-diagonal terms have been flipped relative to $\Lambda[R_x(10°)]$. (i) Sum of the Pauli-X $\Lambda[R_x(10°)]$ and Pauli-Z twirled $\Lambda[R_x(10°)]$, resulting in a PTM in which the off-diagonal terms have been exactly averaged to zero.

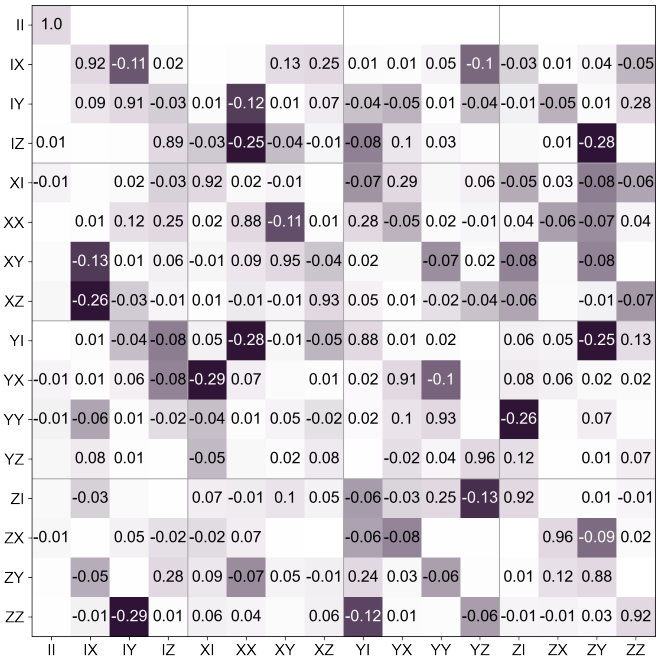
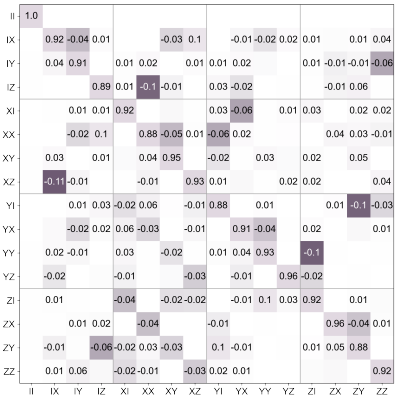


Figure 4.3: **Random CPTP channel Λ in the PTM representation.** The color (transparency) of each cell is determined by the complex angle (magnitude) of each entry.

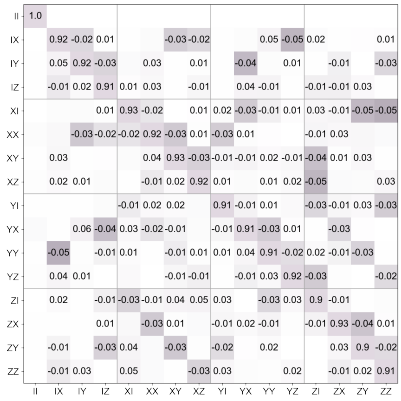
In general, Pauli twirling is implemented by averaging over N randomly sampled Paulis,

$$\bar{\Lambda}(\rho) = \frac{1}{N} \sum_{P \in_R \mathbb{P}_n}^{N} P\Lambda(\rho)P^\dagger, \tag{4.5}$$

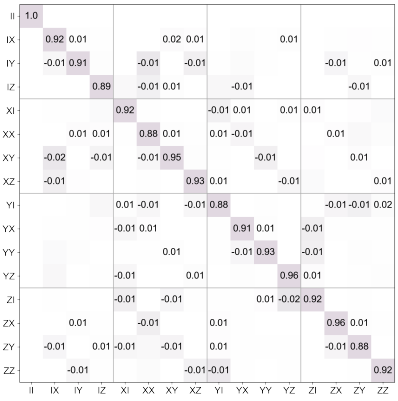
where R denotes that P is chosen at random from the n-qubit Pauli group $\mathbb{P}_n$ each time. In this case, the magnitude of the off-diagonal terms scale as $\sim (\theta/\sqrt{N})$, vanishing as $N \longrightarrow \infty$ or if by luck the correct Paulis are sampled which average to zero (see Fig. 4.2i). This is what is meant by "noise tailoring via twirling." As a concrete example, in Fig. 4.3 we plot a random two-qubit PTM and show how the off-diagonal terms are averaged to zero under Pauli twirling as N is increased from 10, to 100, to 10,000 (see Fig. 4.4). However, note that the diagonal terms remain unchanged for all N. More specifically, Pauli twirling tailors a physical map into Pauli channels, in which the diagonal entries of the PTM remain distinct (Eq. 3.54).



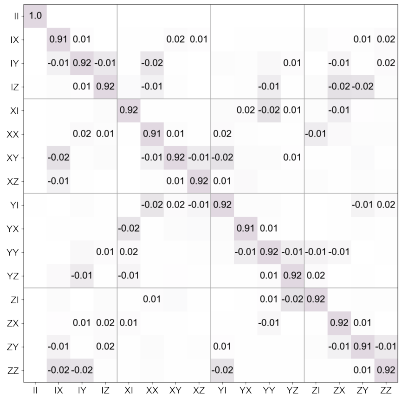
(a) Pauli twirling, $N = 10$



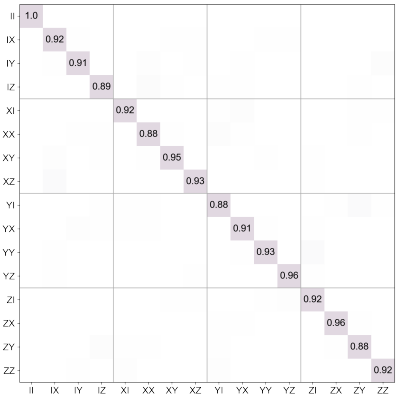
(b) Clifford twirling, $N = 10$



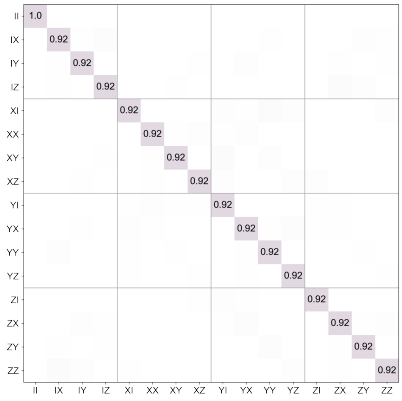
(c) Pauli twirling, $N = 100$



(d) Clifford twirling, $N = 100$



(e) Pauli twirling, $N = 10$k



(f) Clifford twirling, $N = 10$k

Figure 4.4: **Pauli twirling vs. Clifford twirling.** Pauli twirling vs. Clifford twirling as a function of the number N of randomly-selected gates acting on the random unitary in Fig. 4.3. Pauli twirling preserves the eigenvalues along the diagonal of the PTM, whereas Clifford twirling averages them together into a global depolarizing channel.

P	$HPH^\dagger$
I	I
X	Z
Y	$-Y$
Z	X

Table 4.1: $\mathbb{P}_1$ **under** H **conjugation.**

P	$SPS^\dagger$
I	I
X	Y
Y	$-X$
Z	Z

Table 4.2: $\mathbb{P}_1$ **under** S **conjugation.**

4.3 Clifford Twirling

The n-qubit Clifford group [97], denoted $\mathbb{C}_n$, is a group of unitary operators that map Pauli operators to Pauli operators under conjugation:

$$\forall C \in \mathbb{C}_n : CPC^\dagger \mapsto P' \in \mathbb{P}_n,\ \forall P \in \mathbb{P}_n. \tag{4.6}$$

Formally, it is said that the Clifford group is the normalizer of the Pauli group for n qubits. The single-qubit Clifford group $\mathbb{C}_1$ contains 24 single-qubit gates; these consist of any integer number of $\pi/2$ rotations about any of the six cardinal axes of the Bloch sphere ($\pm\hat{x}$, $\pm\hat{y}$, and $\pm\hat{z}$), which includes all single-qubit Pauli gates ($\mathbb{P}_n \subset \mathbb{C}_n$). The two-qubit Clifford group contains 11,520 two-qubit gates, which can be constructed from C_1 and a two-qubit entangling gate, such as a CX, CZ, or (i)SWAP gate. The three-qubit Clifford group contains 92,897,280 elements.

Typical examples of Clifford gates which are not in the Pauli group are the Hadamard H, $S = \sqrt{Z} = Z_{\pi/2}$, CNOT-like, and SWAP-like gates. In fact, the subgroup of Clifford gates $\mathcal{G} = \{H, S, \text{CNOT}\}$ is sufficient to generate the full Clifford group between any two pairs of qubits. In Tables 4.1 and 4.2, we show the action of all single-qubit Pauli operators under conjugation by the Hadamard H and S gates, respectively. In Tables 4.3 and 4.4, we show the action of all two-qubit Pauli operators under conjugation by the CNOT and iSWAP gates, respectively. In all cases, we find that the resulting gate is a Pauli which belongs to $\mathbb{P}_n$ (up to a global phase).

As we can see, the action of conjugating a Pauli operator by a Clifford gate is to map the Pauli operator to the same (a different) Pauli operator if the Pauli commutes (does not commute) with the Clifford gate. In general, under Clifford twirling, a quantum channel $\Lambda(\rho)$ is averaged over N randomly sampled Cliffords:

$$\bar{\Lambda}(\rho) = \frac{1}{N} \sum_{C \in_R \mathbb{C}_n}^{N} C\Lambda(\rho)C^\dagger. \tag{4.7}$$

P	CNOT(P)CNOT†
$I \otimes I$	$I \otimes I$
$I \otimes X$	$I \otimes X$
$I \otimes Y$	$Z \otimes Y$
$I \otimes Z$	$Z \otimes Z$
$X \otimes I$	$X \otimes X$
$X \otimes X$	$X \otimes I$
$X \otimes Y$	$Y \otimes Z$
$X \otimes Z$	$-Y \otimes Y$
$Y \otimes I$	$Y \otimes X$
$Y \otimes X$	$Y \otimes I$
$Y \otimes Y$	$-X \otimes Z$
$Y \otimes Z$	$X \otimes Y$
$Z \otimes I$	$Z \otimes I$
$Z \otimes X$	$Z \otimes X$
$Z \otimes Y$	$I \otimes Y$
$Z \otimes Z$	$I \otimes Z$

Table 4.3: $\mathbb{P}_2$ **under CNOT conjugation.**

P	iSWAP(P)iSWAP†
$I \otimes I$	$I \otimes I$
$I \otimes X$	$Y \otimes Z$
$I \otimes Y$	$-X \otimes Z$
$I \otimes Z$	$Z \otimes I$
$X \otimes I$	$Z \otimes Y$
$X \otimes X$	$X \otimes X$
$X \otimes Y$	$Y \otimes X$
$X \otimes Z$	$I \otimes Y$
$Y \otimes I$	$-Z \otimes X$
$Y \otimes X$	$X \otimes Y$
$Y \otimes Y$	$Y \otimes Y$
$Y \otimes Z$	$-I \otimes X$
$Z \otimes I$	$I \otimes Z$
$Z \otimes X$	$Y \otimes I$
$Z \otimes Y$	$-X \otimes I$
$Z \otimes Z$	$Z \otimes Z$

Table 4.4: $\mathbb{P}_2$ **under iSWAP conjugation.**

If we express $\Lambda(\rho)$ in the PTM representation, for which the diagonal entries represent different Pauli channels, then the act of Clifford twirling $\Lambda(\rho)$ results in a PTM in which the survival probabilities of all of the Pauli channels are averaged together into the same value; this is called a *global depolarizing channel*,

$$\bar{\mathcal{E}}(\rho) = (1-p)\rho + p\frac{\mathbb{I}}{d}, \tag{4.8}$$

where $p \in [0,1]$ is the probability that ρ is mapped to a maximally-mixed state (see Eq. 3.35). From p, we can define an effective depolarizing parameter $f = 1-p$, which is equal to the average value along the diagonal of the PTM Λ:

$$f(\Lambda) = \frac{\mathrm{Tr}\,\Lambda_u}{d^2-1}, \tag{4.9}$$

where Λ_u denotes the unital bloch of Λ (see Fig. 3.4). In Fig. 4.4, we demonstrate the impact of Clifford twirling on the random PTM shown in Fig. 4.3. We show that as N increases from 10, to 100, to 10,000, the off-diagonal entries are averaged to zero (like Pauli twirling), and the diagonal entries slowly converge to the same value.

4.4 Unitary t-designs

The Clifford group holds a special place in quantum computing. According to the Gottesman-Knill theorem [98], quantum circuits containing only Clifford gates and measurements of Pauli operators can be efficiently simulated in polynomial time on a classical computer. Therefore, Clifford circuits are insufficient to realize the full potential of quantum computers over classical computers. In fact, in order to perform universal quantum computation, one requires a gate set which also contains a non-Clifford gate such the $T = \sqrt{S} = Z_{\pi/4}$ gate (sometimes called the "$\pi/8$" for historical reasons). Nonetheless, Clifford gates are ubiquitous in quantum computations and are essential to a number of important applications. For example, stabilizer codes in quantum error correction use Clifford gates for encoding and decoding. Additionally, basic benchmarking procedures for measuring average error rates of quantum gates, such as randomized benchmarking, are constructed entirely of Clifford gates.

One of the primary reasons that the Clifford group remains important is because it forms what is known as a *unitary 2-design*. Classically, the notion of spherical t-designs defines a finite collection of points on the surface of a unit sphere which provide a "good" approximation to the integral over the entire unit sphere [99]. Unitary t-designs are the extension of spherical t-designs to the quantum domain, for which we desire to reproduce the basic properties of the entire unitary group $\mathcal{U}(d)$. More formally, a unitary t-design in d-dimensions is a finite set of unitary operators $\{U_1, ..., U_K\}$ such that the sum over every polynomial $P_{t,t}(U_k) = U_k^{\otimes t} \otimes (U_k^*)^{\otimes t}$ of degree no larger than t in the matrix elements of U and their complex conjugates is equal to the integral of $P_{(t,t)}(U)$ over $\mathcal{U}(d)$,

$$\frac{1}{K}\sum_{k=1}^{K} P_{(t,t)}(U_k) = \int_{\mathcal{U}} d\mu(U) P_{(t,t)}(U), \tag{4.10}$$

where $d\mu(U)$ is the unitarily invariant Haar measure. Statistically, a unitary t-design simulates the properties of uniformly distributed (Haar) random matrices up to the t'th moment. In lay terms, this means that unitaries in a t-design are evenly spaced around the unit sphere defining $\mathcal{U}(d)$, with larger values of t defining more densely spaced points. For example, the set of Pauli operators in d-dimensions forms a unitary 1-design.

Generating a random unitary from $\mathcal{U}(d)$ using only single- and two-qubits gates is an exponentially difficult task, taking of order $\mathcal{O}(n^2 2^{2n})$ quantum gates for n-qubits [100], making an integration over the uniform Haar distribution practically impossible. However, it is important to be able to approximate the properties of $\mathcal{U}(d)$; for many applications, it is sufficient to approximate $\mathcal{U}(d)$ up to second order. Returning to the topic of twirling,

for a unitary group $\mathcal{G} = \{U_1, ..., U_K\}$ to be a unitary t-design, it must satisfy the following equality:

$$\frac{1}{K}\sum_{k=1}^{K} U_k^\dagger \Lambda(U_k \rho U_k^\dagger) U_k = \int_{\mathcal{U}(d)} d\mu(U) U^\dagger \Lambda(U\rho U^\dagger) U. \tag{4.11}$$

More specifically, for $\mathcal{G}$ to be a 2-design, it must satisfy the following [101]:

$$\frac{1}{K^2}\sum_{k,k'=1}^{K} \left|\mathrm{Tr}\left(U_{k'}^\dagger U_k\right)\right|^4 = 2. \tag{4.12}$$

The Clifford group satisfies both requirements for $t = 2$ [96], with

$$\frac{1}{|\mathbb{C}_n|}\sum_{k=1}^{|\mathbb{C}_n|} C_k^\dagger P C_k \rho C_k^\dagger Q C_k = \int_{\mathcal{U}(d)} d\mu(U) U^\dagger P U \rho U^\dagger Q U, \ \forall \rho, \tag{4.13}$$

where we use the notation of Eq. 4.4 containing arbitrary linear operators $\{P, Q\}$.

This is a particularly useful result, because arbitrary Clifford gates can be decomposed into single- and two-qubit gates using only $\mathcal{O}(n^2)$ quantum gates for n-qubits [96] (see also the discussion in the supplementary information of Ref. [102]). We note in passing that while it is considered "common knowledge" that the Clifford group is a unitary 2-design, this in fact not universally true, as it depends on whether the dimension d is prime or not [103]. In fact, for any even prime-power dimension d, the Clifford group also forms a unitary 3-design [104, 105].

Chapter 5

Quantum Characterization, Verification, and Validation

Utilizing quantum computers to solve classically-intractable problems like Shor's algorithm [106] will require quantum error correction (QEC) [107, 108, 109, 110, 111], which can only protect against arbitrary errors if the error rate of each physical qubit is below a given fault tolerance (FT) threshold [112, 113, 114, 115, 116, 117]. With the rapid advancements being made in the development of better qubits, better quantum gates, and larger quantum processors — all working toward the ultimate goal of fault-tolerant QEC — the field of quantum characterization, verification, and validation (QCVV) emerged out of the necessity for benchmarking this progress, as well as understanding what types of noise and errors limit our computations. In the past several decades, a vast number of different QCVV protocols have been developed, ranging widely from methods which are scalable but provide less information to methods which provide more information but are less scalable. These methods can be broadly placed into four different categories:

1. Characterization of quantum states
2. Characterization of quantum processes or gates
3. Holistic characterization of the capabilities of quantum processors
4. Application benchmarks of quantum algorithms

In this chapter, we aim to give a first-principles understanding of contemporary methods of benchmarking quantum gates, what these benchmarks can and cannot tell us about algorithm performance, and what we can learn about how close we are to FT thresholds. We will start by giving a broad overview of different ways to define the accuracy (or, conversely, error) of a quantum state or process, and discuss to what extent benchmarked error rates can inform us about FT thresholds. Then we will cover common benchmarking protocols such as randomized benchmarking and gate set tomography, as well as introducing more advanced methods, such as cycle benchmarking and cycle error reconstruction. Finally, we

will comment on other contemporary benchmarks in the field, and where these fit within the vast array of current methods available.

5.1 An Overview of Fidelity Measures and Error Metrics

Various definitions exist for quantifying the "error" of a quantum state or operation. Broadly speaking, these can be divided into fidelity-based error *measures* and norm-based error *metrics* (e.g. L_1-norm, L_2-norm, etc.)[1]. Which measure or metric to use will depend on many factors, including what information is desired, which benchmarking method is used, what application does one want to implement, etc. In this section, we aim to give a broad overview of the various error measures and metrics that are commonly used in QIS.

Fidelity

State Fidelity

Consider two quantum states ρ and σ. The *state fidelity* between these two density matrices is defined[2] to be

$$\boxed{F(\rho,\sigma) \equiv \left(\mathrm{Tr}\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right)^2}. \tag{5.1}$$

The fidelity has several important properties, namely, it is bounded by $0 \leq F(\rho,\sigma) \leq 1$ for any pair $\{\rho,\sigma\}$, and it is symmetric in ρ and σ, with $F(\rho,\sigma) = F(\sigma,\rho)$. If $F(\rho,\sigma) = 1$, then ρ and σ are identical. If $F(\rho,\sigma) = 0$, then ρ and σ are orthogonal. For the case in which $\rho = |\psi\rangle\langle\psi|$ is a pure state, Eq. 5.1 simplifies to

$$F(\rho,\sigma) = \mathrm{Tr}[\sigma\,|\psi\rangle\langle\psi|] = \langle\psi|\,\sigma\,|\psi\rangle\,. \tag{5.2}$$

If both ρ and σ are pure states, with $\sigma = |\phi\rangle\langle\phi|$, then F can be written as

$$F(\rho,\sigma) = |\,\langle\psi|\phi\rangle\,|^2. \tag{5.3}$$

From Eq. 5.3, we can see that the fidelity is the inner product of ρ and σ. In fact, the physical motivation behind the fidelity in quantum mechanics is that it represents a measure of the *overlap* between two quantum states. While it is convenient to think about the fidelity between pure states, if ρ or σ are potentially mixed states after passing through a noise channel, then $F(\rho,\sigma)$ can be interpreted as being the lower bound on the overlap between the two, assuming they were initially pure states [119]. The fidelity also has the convenient

[1] See Chapter 9 in Ref. [26] for a good overview of this topic.

[2] Note that our definition of fidelity differs from that of Ref. [26], which defines the fidelity as $\sqrt{F(\rho,\sigma)}$, the square root of our definition. We, however, adhere to the definition of fidelity presented in Ref. [118].

operational interpretation as being the probability that ρ and σ are indistinguishable from one another. For example, suppose ρ represents an ideal (pure) quantum state, and σ is our attempt to experimentally prepare ρ in the lab; in this case, $F(\rho, \sigma)$ represents the probability that we were able to successfully prepare ρ, assuming perfect measurement of σ. The connection between fidelity and probability will become useful later on when we discuss the how it relates to benchmarks of quantum operations.

A special case of $F(\rho, \sigma)$ exists when ρ and σ commute, in which case they can be diagonalized in the same basis. Writing $\rho = \sum_i p_i \left|i\right\rangle\!\left\langle i\right|$ and $\sigma = \sum_i q_i \left|i\right\rangle\!\left\langle i\right|$ for some orthonormal basis $\{\left|i\right\rangle\}$, $F(\rho, \sigma)$ can be written as

$$\boxed{F(p_i, q_i) = \left(\sum_i \sqrt{p_i q_i}\right)^2}, \tag{5.4}$$

where p_i (q_i) is the probability of measuring ρ (σ) in the state $\left|i\right\rangle$ (see Eq. 3.2). Therefore, when $[\rho, \sigma] = 0$, $F(\rho, \sigma)$ reduces to the classical definition of fidelity $F(p_i, q_i)$ between two probability distributions. This is sometimes referred to as the *Hellinger fidelity*, and can be used to compute the fidelity between two bit string distributions.

Average Gate Fidelity

When we speak of quantum gates, we must discuss the fidelity of quantum operations, instead of the fidelity of quantum states. Various metrics and measures exist for quantifying the fidelity of a quantum gate. The most common method for characterizing gate fidelities is via randomized benchmarks (discussed in the following section), which defines the *average gate fidelity* as

$$\mathcal{F}(G, \tilde{G}) \equiv \int d\psi \, \mathrm{Tr}\Big[G^\dagger(\psi)\tilde{G}(\psi)\Big], \tag{5.5}$$

where G ($\tilde{G}$) is the ideal (noisy) gate acting on the state $\left|\psi\right\rangle$, and the integral is over the uniform Haar measure $d\psi$. Because $\tilde{G} = G$ only in the noiseless limit, $\mathcal{F} = 1$ indicates that the quantum gate is operationally perfect over all possible input states. By defining G as an ideal unitary channel $\mathcal{U}(\rho) = U \left|\psi\right\rangle\!\left\langle\psi\right| U^\dagger$ acting on the quantum state $\rho = \left|\psi\right\rangle\!\left\langle\psi\right|$, and by defining $\tilde{G}$ as the noisy channel $\mathcal{E}(\rho)$, we can re-write Eq. 5.5 as

$$\mathcal{F}(\mathcal{U}, \mathcal{E}) = \int d\psi \, \mathrm{Tr}\big[\mathcal{U}^\dagger(\rho)\mathcal{E}(\rho)\big] \tag{5.6}$$

$$= \int d\psi \, \left\langle\psi\right| U^\dagger \mathcal{E}(\rho) U \left|\psi\right\rangle . \tag{5.7}$$

Eq. 5.6 is useful for understanding the meaning of $\mathcal{F}$ in terms of noisy quantum operations, since $\mathcal{F} \longrightarrow 1$ only if $\mathcal{U}(\rho)$ is approximated well by $\mathcal{E}(\rho)$, which is the case only when the magnitude of noise and errors impacting U is small.

Many randomized benchmarking procedures are constructed such that $U = \mathbb{I}$, and therefore the noise channel $\mathcal{E}$ is (by construction) the identity channel in the noiseless limit. In this case, the average gate fidelity is typically written as

$$\boxed{\mathcal{F}(\mathcal{E}) = \int d\psi \, \langle\psi| \, \mathcal{E}(\rho) \, |\psi\rangle}, \tag{5.8}$$

where $\mathcal{F}$ is an integral over all pure states, and is thus a measure of the overlap of $\mathcal{E}$ with $\mathbb{I}$ that is independent of the input state. Therefore, $\mathcal{F}$ defines the success probability that evolving an arbitrary pure state $\rho = |\psi\rangle\langle\psi|$ through a noisy channel $\mathcal{E}(\rho)$ returns to the original state upon measurement, where $\mathcal{F} = 1$ indicates that quantum information is perfectly preserved by the channel $\mathcal{E}$. From Eq. 5.8, we can define the *average error rate* $r(\mathcal{E})$ in terms of $\mathcal{F}(\mathcal{E})$,

$$r(\mathcal{E}) = 1 - \mathcal{F}(\mathcal{E}), \tag{5.9}$$

where $r(\mathcal{E})$ defines the probability of an error occurring per gate, or average gate infidelity[3].

It is interesting to note that the average gate fidelity is unchanged by twirling. To see this, we utilize the definition of a twirled channel (Eq. 4.3) in place of $\mathcal{E}(\rho)$ in Eq. 5.8:

$$\begin{aligned}
\mathcal{F}(\bar{\mathcal{E}}) &= \int d\psi \int d\mu(U) \, \langle\psi| \, U^\dagger \mathcal{E}(U\rho U^\dagger) U \, |\psi\rangle && (5.10)\\
&= \int d\mu(U) \int d\psi \, \langle\psi| \, U^\dagger \mathcal{E}(U\rho U^\dagger) U \, |\psi\rangle && (5.11)\\
&= \int d\mu(U) \mathcal{F}(\mathcal{E}) && (5.12)\\
&= \mathcal{F}(\mathcal{E}), && (5.13)
\end{aligned}$$

where in lines 2-3 we make a change of variables $|\psi'\rangle \equiv U \, |\psi\rangle$ and subsequently drop the primes. One way to understand this result is to recognize that the average gate fidelity is only sensitive to the diagonal terms in $\mathcal{E}(\rho)$ due to the matrix element $\langle\psi| \, \mathcal{E}(\rho) \, |\psi\rangle$, which is unaffected by the twirling of off-diagonal terms in the process.

Process Fidelity

A related measure, known as the *process* (or *entanglement*) *fidelity*, defines how well an entangled state is preserved if the quantum channel $\mathcal{E}$ acts only on a subset of a maximally-entangled state $\rho = |\psi\rangle\langle\psi|$:

$$\boxed{F(\mathcal{E}) = \langle\psi| \, (\mathbb{I} \otimes \mathcal{E})(\rho) \, |\psi\rangle}, \tag{5.14}$$

where $\mathbb{I}$ is the identity operation acting on the subsystem not passed through $\mathcal{E}$. Operationally, F can be estimated experimentally by measuring and computing the weighted

[3]In this work, we do not distinguish between the average error rate and average gate infidelity measured by randomized benchmarks. However, it should be noted that more precise definitions of these terms are given in the literature [120, 121]

	$\mathcal{F}$	r	F	e_F	f
$\mathcal{F}$	$\mathcal{F}$	$1-r$	$\frac{dF+1}{d+1}$	$1-e_F\frac{d}{d+1}$	$\frac{(d-1)f+1}{d}$
r	$1-F$	r	$(1-F)\frac{d}{d+1}$	$e_F\frac{d}{d+1}$	$(1-f)\frac{d-1}{d}$
F	$\frac{(d+1)\mathcal{F}-1}{d}$	$1-r\frac{d+1}{d}$	F	$1-e_F$	$\frac{(d^2-1)f+1}{d^2}$
e_F	$(1-F)\frac{d+1}{d}$	$r\frac{d+1}{d}$	$1-F$	e_F	$(1-f)\frac{d^2-1}{d^2}$
f	$\frac{d\mathcal{F}-1}{d-1}$	$1-r\frac{d}{d-1}$	$\frac{d^2F-1}{d^2-1}$	$1-e_F\frac{d^2}{d^2-1}$	f

Table 5.1: **Summary of the linear relationship between the average gate fidelity $\mathcal{F}$, the average gate infidelity r, the process fidelity F, the process infidelity e_F, and the effective depolarizing parameter f.** $d = 2^n$ for n qubits. (Table adapted from Refs. [124, 125].)

average of the diagonal components of the Pauli Transfer Matrix (PTM) of an error process (see Eq. 3.48). The average gate fidelity and process fidelity are related by a simple dimensionality factor [122, 123],

$$\mathcal{F}(\mathcal{E}) = \frac{dF(\mathcal{E}) + 1}{d+1}, \tag{5.15}$$

where $d = 2^n$ (n qubits). Since the process *infidelity* $e_F(\mathcal{E})$ is given as

$$e_F(\mathcal{E}) = 1 - F(\mathcal{E}), \tag{5.16}$$

the average gate infidelity and process infidelity are also related by a simple dimensionality factor:

$$\boxed{e_F(\mathcal{E}) = r(\mathcal{E})\frac{d+1}{d}}. \tag{5.17}$$

The process infidelity is also related to the effective depolarizing parameter (Eq. 4.9) via

$$f = 1 - e_F(\mathcal{E})\frac{d^2}{d^2-1}. \tag{5.18}$$

In Table 5.1, we include a summary of the linear relationships between the average gate (in)fidelity, the process (in)fidelity, and the effective depolarizing parameter.

The process infidelity is a useful measure for quantifying the error rate of composite processes in cycles (or layers) of gates, as it is stable under tensor products of quantum gates of differing dimensions (e.g. single- and two-qubit gates) — whereas the average gate infidelity is not. To see this, we note that the fidelity of a composite process acting on two qubits (say, qubit 1 and qubit 2) can be written as

$$F(\mathcal{E}_1 \otimes \mathcal{E}_2) = 1 - e_F(\mathcal{E}_1 \otimes \mathcal{E}_2) = [1 - e_F(\mathcal{E}_1)][1 - e_F(\mathcal{E}_2)],$$

where

$$e_F(\mathcal{E}_1 \otimes \mathcal{E}_2) = e_F(\mathcal{E}_1) + e_F(\mathcal{E}_2) - e_F(\mathcal{E}_1)e_F(\mathcal{E}_2) \approx e_F(\mathcal{E}_1) + e_F(\mathcal{E}_2),$$

assuming that both $e_F(\mathcal{E}_1)$ and $e_F(\mathcal{E}_2)$ are small. Note that these relationships do not hold for the average gate infidelity, which must be converted to the process fidelity via Eq. 5.17 to compute the average error rate of composite processes.

Trace Distance

Consider two quantum states ρ and σ. The *trace distance* between these two density matrices is defined to be

$$\boxed{D(\rho, \sigma) \equiv \frac{1}{2} \operatorname{Tr} |\rho - \sigma|}, \tag{5.19}$$

where $|A| = \sqrt{A^\dagger A}$ and $0 \leq D(\rho, \sigma) \leq 1$. The trace distance can be understood as being related to the Euclidean distance between ρ and σ for the special case of single qubits states on the Bloch sphere. To see this, we can write ρ and σ in terms of their respective Bloch vectors $\mathbf{r}$ and $\mathbf{q}$,

$$\rho = \frac{1}{2}(I + \mathbf{r} \cdot \boldsymbol{\sigma})$$

and

$$\sigma = \frac{1}{2}(I + \mathbf{q} \cdot \boldsymbol{\sigma})$$

where $\boldsymbol{\sigma}$ is the single-qubit Pauli vector. The trace distance can be written as

$$D(\rho, \sigma) = \frac{1}{2} \operatorname{Tr} |\rho - \sigma| = \frac{1}{4} \operatorname{Tr} |(\mathbf{r} - \mathbf{q}) \cdot \boldsymbol{\sigma}|. \tag{5.20}$$

Because the eigenvalues of $\boldsymbol{\sigma}$ are ± 1, the trace of $|(\mathbf{r} - \mathbf{q}) \cdot \boldsymbol{\sigma}| = 2|(\mathbf{r} - \mathbf{q})|$, and thus

$$D(\rho, \sigma) = \frac{1}{2}|\mathbf{r} - \mathbf{q}|. \tag{5.21}$$

In other words, the trace distance between two single-qubit states on the Bloch sphere is exactly equal to one-half the Euclidean distance between their vectors. This definition of the trace distance provides some physical intuition behind the difference between the state fidelity $F(\rho, \sigma)$ (Eq. 5.1) and $D(\rho, \sigma)$: while the state fidelity is a measure of the *overlap* between two quantum states, the trace distance is a measure of the *distance* between the two states. Therefore, in contrast to the state fidelity being a measure of state *indistinguishability*, the trace distance is a measure of the *distinguishability* between two quantum states. In many ways the two are dual to each other, and they are in fact closely related via the following inequalities [126]:

$$1 - \sqrt{F(\rho, \sigma)} \leq D(\rho, \sigma) \leq \sqrt{1 - F(\rho, \sigma)}, \tag{5.22}$$

where $D(\rho, \sigma)$ saturates the upper bound only if ρ and σ are both pure states. If either ρ or σ is a pure state, then

$$1 - F(\rho, \sigma) \leq D(\rho, \sigma). \tag{5.23}$$

The trace distance has several important properties. Firstly, it is a proper distance metric on the space of density matrices because it satisfies the following three properties:

1. $D(\rho, \sigma) = 0$ i.f.f. $\rho = \sigma$
2. Symmetry: $D(\rho, \sigma) = D(\sigma, \rho)$
3. Triangle inequality: $D(\rho, \sigma) \leq D(\rho, \tau) + D(\tau, \sigma)$

While the state fidelity is a symmetric function of its inputs, $F(\rho, \sigma) = F(\sigma, \rho)$, it does not satisfy properties 1 and 3 and therefore is not a proper distance metric. For this reason, we do not refer to the fidelity as a *distance metric*, but rather a *measure* of the overlap between two quantum states. It is worth noting that the state fidelity does give rise to a proper metric known as the *Bures metric*, which is related to the Euclidean distance and angle between two quantum states.

Secondly, the trace distance is contractive for any trace-preserving quantum process $\mathcal{E}$:

$$D[\mathcal{E}(\rho), \mathcal{E}(\sigma)] \leq D(\rho, \sigma). \tag{5.24}$$

This contractivity property expresses the fact that any TP quantum operation on both qubits cannot *increase* the distinguishability of two quantum states. A limiting example of Eq. 5.24 is that of a unitary transformation U, in which case $D(U\rho U^\dagger, U\sigma U^\dagger) = D(\rho, \sigma)$. Therefore, the trace distance is invariant under unitary transformations.

Finally, the quantum trace distance is equivalent to the classical trace distance for the case in which ρ and σ commute and are thus diagonal in the same basis. Writing $\rho = \sum_i p_i |i\rangle\langle i|$ and $\sigma = \sum_i q_i |i\rangle\langle i|$ for some orthonormal basis $\{|i\rangle\}$, the trace distance can be written as

$$D(\rho, \sigma) = \frac{1}{2} \operatorname{Tr} \left| \sum_i (p_i - q_i) |i\rangle\langle i| \right| \tag{5.25}$$

$$= \frac{1}{2} \sum_i |p_i - q_i| = D(p_i, q_i), \tag{5.26}$$

where p_i (q_i) is the probability of measuring ρ (σ) in the state $|i\rangle$. This quantity is often referred to as the *total variation distance* (TVD) or *Kolmogorov distance* between classical probability distributions. We will use this metric extensively in the remainder of this work and choose the former name, defining it to be

$$\boxed{d_{\mathrm{TV}}(p_i, q_i) = \frac{1}{2} \sum_i |p_i - q_i|} \tag{5.27}$$

in order to distinguish it from the quantum trace distance. For a countable set, the TVD is related to the L_1-norm:

$$d_{\mathrm{TV}}(p,q) = \frac{1}{2}||p-q||_1, \tag{5.28}$$

where $p = \{p_i\}$ and $q = \{q_i\}$ are classical probability distributions, showing that the TVD is a proper distance metric on the space of classical distributions.

While the trace distance is a fundamental measure of distance between two quantum states, the TVD is a classical distance metric which depends on the measurement basis. In fact, for any quantum states ρ and σ, it can be shown that the TVD is upper-bounded by the trace distance between for any positive operator-valued measure (POVM) $\{M_i\}$:

$$d_{\mathrm{TV}}(p,q) \leq D(\rho,\sigma), \tag{5.29}$$

where

$$D(\rho,\sigma) = \max_{\{M_i\}} d_{\mathrm{TV}}(p_i,q_i). \tag{5.30}$$

Therefore, the trace distance can interpreted as the maximum possible probability of distinguishing ρ from σ upon measurement. The TVD of probabilities corresponding to a measurement $\{M_i\}$ will saturate the trace distance if and only if $\{M_i\}$ are projectors on the eigenbasis of $(\rho-\sigma)$. In general, if ρ and σ are close in trace distance, then any measurement of the two density matrices in the same basis will result in probability distributions that are close in total variation distance.

The equivalence between the quantum trace distance and the classical TVD allows us to define the *convexity* of the trace distance. In fact, the trace distance is convex in its first input,

$$D\left(\sum_i p_i\rho_i, \sigma\right) \leq \sum_i p_i D(\rho_i,\sigma), \tag{5.31}$$

and *doubly convex* in both inputs,

$$D\left(\sum_i p_i\rho_i, \sum_i p_i\sigma_i\right) \leq \sum_i p_i D(\rho_i,\sigma_i) \tag{5.32}$$

for probabilities p_i. The convexity of the trace distance can be interpreted in the following way: distinguishing ρ from σ when the measurement basis $|i\rangle$ is not known cannot be greater than distinguishing ρ from σ when the measurement basis is known but randomly chosen from p_i.

Diamond Norm

While the average gate infidelity $r(\mathcal{E})$ and process infidelity $e_F(\mathcal{E})$) quantify *average* error rates, they do not guarantee the threshold below which *all* errors occur, which is required for

fault-tolerant QEC. Therefore, FT thresholds are typically defined via the worst-case error rate (i.e. *diamond norm* [127])[4],

$$\boxed{\epsilon_\diamond(\mathcal{E}) = \frac{1}{2}\left|\left|\mathcal{E} - \mathbb{I}\right|\right|_\diamond = \frac{1}{2}\sup_{\rho\in\mathcal{H}_{d^2}} \left|\left|\left[\mathbb{I}_d \otimes (\mathcal{E} - \mathbb{I}_d)\right](\rho)\right|\right|_1}, \tag{5.33}$$

where the supremum is taken over all pure states in Hilbert space $\mathcal{H}$ of dimension d^2 and $\left|\left|X\right|\right|_1 = \mathrm{Tr}\sqrt{X^\dagger X}$ is the trace norm. The diamond norm defines the maximum distance between a noise channel $\mathcal{E}$ and the ideal channel, in this case the identity operation $\mathbb{I}$; therefore, it is sometimes referred to as the *diamond distance from the identity*. The diamond norm can also be written in terms of a quantum map $\mathcal{G}$ induced by an ideal gate and its noisy implementation $\tilde{\mathcal{G}}$,

$$\epsilon_\diamond(\mathcal{G}, \tilde{\mathcal{G}}) = \frac{1}{2}\sup_\rho \left|\left|\left[\mathbb{I}_d \otimes (\mathcal{G} - \tilde{\mathcal{G}})\right](\rho)\right|\right|_1, \tag{5.34}$$

conveniently capturing the maximum probability that the output of a noisy gate can be distinguished from the ideal output. In practice, $\epsilon_\diamond(\mathcal{E})$ represents the worst-case performance of a quantum gate, whereas $r(\mathcal{E})$ represents the average-case performance for a single instance of the gate.

Average error rates defined via $r(\mathcal{E})$ or $e_F(\mathcal{E})$ can be measured directly via randomized benchmarks, but there exists no known scalable method for measuring $\epsilon_\diamond(\mathcal{E})$. However, the worst-case error rate is bounded by the average error rate [120, 128, 129, 130] via

$$r(\mathcal{E})\frac{d+1}{d} \leq \epsilon_\diamond(\mathcal{E}) \leq \sqrt{r(\mathcal{E})}\sqrt{d(d+1)}, \tag{5.35}$$

or in terms of the process infidelity,

$$\boxed{e_F(\mathcal{E}) \leq \epsilon_\diamond(\mathcal{E}) \leq \sqrt{e_F(\mathcal{E})}d}, \tag{5.36}$$

where the lower bound is set by a purely stochastic noise channel, and the upper bound is set by a purely unitary error channel. Therefore, average error rates and worst-case error rates can differ by orders of magnitude in the presence of coherent errors.

To provide some intuition behind what sets the upper and lower bounds on the diamond norm, consider a unitary error by an angle θ about the x-axis for a single qubit,

$$U(\hat{x}, \theta) = \exp\left(-i\frac{\theta}{2}\sigma_x\right) = \begin{pmatrix} \cos(\theta/2) & -i\sin(\theta/2) \\ i\sin(\theta/2) & \cos(\theta/2) \end{pmatrix}, \tag{5.37}$$

[4]Note that we do not distinguish between the worst-case error rate and the diamond norm. However, they are occasionally differentiated in the literature [128].

or in the PTM representation,

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos(\theta) & -\sin(\theta) \\ 0 & 0 & \sin(\theta) & \cos(\theta) \end{pmatrix}. \tag{5.38}$$

For small θ, the diagonal components of Λ scale as $\cos(\theta) \approx 1 - \frac{1}{2}\theta^2$, and the off-diagonal terms scale as $\sin(\theta) \approx \theta$. Since state-preservation in PTMs equates to $\Lambda_{PP} = 1$, and since average gate fidelities are only sensitive to the diagonal terms in Λ, then one can equate $\cos(\theta)$ with the process fidelity F of this error, and therefore $e_F \approx \theta^2$. However, being a distance metric that is sensitive to the maximum deviation of an error from the identity operation, the diamond norm *is* sensitive to the off-diagonal terms in Λ, which are quadratically larger, with $\sqrt{e_F} \approx \theta$. Here, we see that the off-diagonal terms due to coherent errors set the $\sim \sqrt{e_F}$ upper bound of Eq. 5.36.

We see that error assessments based on the average gate infidelity or process infidelity are only sensitive to the diagonal terms of the error process, whereas norm-based error-metrics, such as the TVD and the diamond norm, will generally be sensitive to the off-diagonal terms in the error process. However, while the diamond norm is *always* sensitive to the off-diagonal terms, the TVD is a basis-dependent metric and is only sensitive to the off-diagonal terms if the (ideal) target state is orthogonal to the measurement basis, where the orthogonality implies that a greater number of off-diagonal terms can contribute to the error metric. The TVD and the diamond norm are related via by the following inequality,

$$d_{\mathrm{TV}}(p, q) \leq \epsilon_{\diamond}(\mathcal{E}). \tag{5.39}$$

The upper bound in Eq. 5.39 follows from considering the definition of $\epsilon_{\diamond}(\mathcal{E})$, when defining $\rho_{\mathcal{E}} = \mathbb{I} \otimes \mathcal{E}(\rho)$ and $\rho_{\mathbb{I}} = \mathbb{I} \otimes \mathbb{I}(\rho)$, so that

$$\frac{1}{2}||\rho_{\mathcal{E}} - \rho_{\mathbb{I}}||_1 \leq \epsilon_{\diamond}(\mathcal{E}). \tag{5.40}$$

Then, defining the partial trace taking any ρ (on the dilated Hilbert space) to σ, a density operator on the (reduced) base Hilbert space, and noting that the partial trace is a CPTP map and that the norm $||\cdot||_1$ is contractive under CPTP maps, it follows that

$$||\sigma_{\mathcal{E}} - \sigma_{\mathbb{I}}||_1 \leq ||\rho_{\mathcal{E}} - \rho_{\mathbb{I}}||_1. \tag{5.41}$$

Finally, if we define p (q) as the probability distribution induced by the measurement of $\sigma_{\mathcal{E}}$ ($\sigma_{\mathbb{I}}$), then the definition of the TVD implies

$$d_{\mathrm{TV}}(p, q) \leq \frac{1}{2}||\sigma_{\mathcal{E}} - \sigma_{\mathbb{I}}||_1 \leq \epsilon_{\diamond}(\mathcal{E}). \tag{5.42}$$

By extension, by considering the definition of the diamond norm $\epsilon_\diamond(\mathcal{G}, \tilde{\mathcal{G}})$ in terms of the noisy and ideal quantum gates $\mathcal{G}$ and $\tilde{\mathcal{G}}$, respectively, it can be shown that the TVD between the real and ideal probability distributions of some circuit is upper-bounded by the sum of the diamond errors $\epsilon_\diamond(\mathcal{G}, \tilde{\mathcal{G}})$ over all gates in the circuit [127]. Because of this, FT thresholds are defined in terms of the diamond norm, which captures the worst-case error rate of a gate under repeated operation in a circuit.

5.2 Randomized Benchmarking

Randomized benchmarking (RB) was developed in the mid- to late-2000s [96, 131, 132, 133, 134] in order to provide a method of measuring average gate fidelities of quantum operations. The idea is based on the notion of the reversibility of unitary operators, and was born out of the necessity for finding more scalable benchmarks than quantum process tomography (QPT) [135] that decouple gate errors from state-preparation and measurement (SPAM) errors. While early papers proposed producing an average fidelity by twirling over Haar-random unitaries [i.e. randomly sampling unitaries $U \in \mathcal{U}(d)$], modern variants are implemented by averaging over the n-qubit Clifford group $\mathbb{C}_n$. Numerous different flavors of RB and related methods have been developed since the early proposals; however, RB remains to this day the *de facto* standard for benchmarking error rates in quantum gates.

In Protocol 1, we outline the basic procedure for performing RB, and highlight the procedure in Fig. 5.1. While the constants A and B in the exponential decay function $\bar{p}(m) = Af^m + B$ (Eq. 5.43) are both SPAM-dependent constants, Ref. [136] proposed a simple modification to the standard RB procedure that allows one to fit an exponential decay function without B, which they call a "nuisance parameter." Therefore, in the simplified decay function $\bar{p}(m) = Af^m$, A can be regarded as the SPAM constant which quantifies the offset from 1 for $\bar{p}(0)$. A additionally contains the error of the RB inversion gate (see Fig. 5.1); therefore, if the gate errors are significant, this will distort the SPAM parameter. We note that the decay parameter f that appears in Eq. 5.43 is the same as the effective depolarizing parameter (Eq. 4.9), and that Eq. 5.44 relates f to the average gate fidelity $\mathcal{F}(\mathcal{E})$, first shown in Table 5.1. We are justified in equating the average gate fidelity to the effective depolarizing parameter because, as outlined in the previous chapter, Clifford twirling tailors all gate errors into a global depolarizing channel. Furthermore, as outlined in Section 5.1, the average gate fidelity is unchanged by twirling, even though the individual Pauli channel eigenvalues may change.

Because the RB circuit is constructed from Clifford gates, the average gate infidelity $r = 1 - \mathcal{F}$ represents the *error per Clifford* (EPC). The average number of single-qubit (SQ) gates per single-qubit Clifford is 1.875 [102]. Therefore, the EPC for $\mathbb{C}_1$ is given as

$$r_{\mathbb{C}_1} = 1.875 r_{\mathrm{SQ}}, \tag{5.45}$$

where we assume that all single-qubit gates have the same error rate r_{SQ}. The average number of gates per two-qubit Clifford is 1.5 two-qubit gates and 8.25 single-qubit gates

Protocol 1 Randomized Benchmarking

1. Select m Clifford gates $C \in \mathbb{C}_n$ sampled uniformly at random: $C_m \circ C_{m-1} \circ ... \circ C_2 \circ C_1$
2. Compute the inversion gate $C_{m+1} = (C_m \circ C_{m-1} \circ ... \circ C_2 \circ C_1)^{-1} = C_{1:m}^{-1}$.
3. Construct a circuit $\mathcal{C}$ composed of the randomly sampled Clifford gates and the inversion gate: $\mathcal{C} = C_{1:m}^{-1} \circ C_m \circ C_{m-1} \circ ... \circ C_2 \circ C_1$
4. Prepare an n-qubit quantum state ρ, typically chosen to be the ground state: $\rho = {}^{n\otimes}|0\rangle\langle 0|^{\otimes n}$
5. Implement the circuit $\mathcal{C}$ on a quantum computer and perform a POVM $\{M, \mathbb{I} - M\}$, where typically $\{M = |0\rangle\langle 0|, \mathbb{I} - M = |1\rangle\langle 1|\}$ corresponds to a measurement in the computational basis. The result can be labeled as having returned to the original state or not.
6. Repeat steps 4 and 5 N times (e.g. $N = 1000$ "shots") to construct a probability distribution $p(m, \mathcal{C}) = \Pr(M|\mathcal{C}, \rho)$ to some appropriate precision, where p denotes the success probability of returning to the original state for the circuit $\mathcal{C}$.
7. Repeat steps 1 - 6 for s different circuits $\mathcal{C}$ composed of m different randomly sampled gates $C \in \mathbb{C}_n$ (e.g. $s = 30$)
8. Repeat steps 1 - 7 for l different values of depth-m random circuits (e.g. $l = 3$).
9. Compute the average success probability $\bar{p}(m)$ for each depth m and fit the data to an exponential decay function:
$$\bar{p}(m) = Af^m + B, \tag{5.43}$$
where A and B are related the state-preparation and measurement (SPAM) error.
10. Compute the average gate fidelity $\mathcal{F}(\mathcal{E})$ from f:
$$\boxed{\mathcal{F} = \frac{(d-1)f + 1}{d}}, \tag{5.44}$$
where $\mathcal{F}(\mathcal{E})$ is defined in Eq. 5.8.

[102]. Therefore, the EPC for $\mathbb{C}_2$ is given as

$$r_{\mathbb{C}_2} = \frac{3}{2} r_{CZ} + \frac{33}{4} r_{\mathrm{SQ}}, \tag{5.46}$$

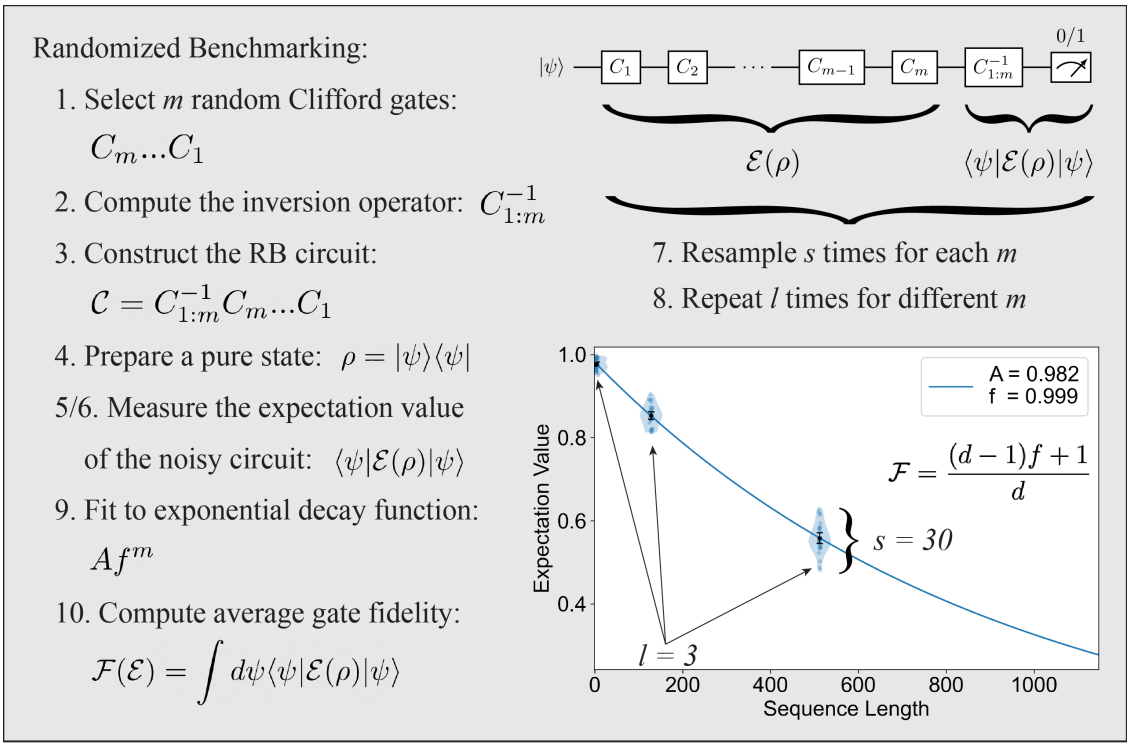
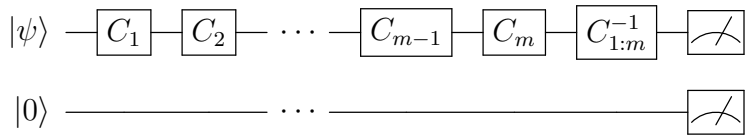


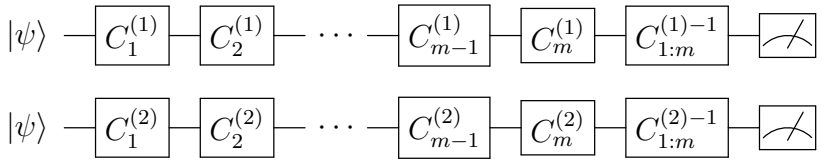
Figure 5.1: **Randomized Benchmarking.** Summary of the randomized benchmarking procedure outlined in Protocol 1.

where we take the CZ gate as the native two-qubit operation on the device.

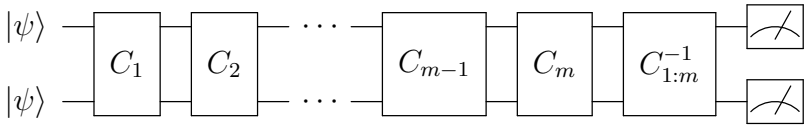
The utility of RB is that it provides a single number which quantifies the average error rate of a quantum gate. One can perform RB for isolated single-qubit gates, simultaneous single-qubit gates (sRB), or two-qubit gates (see Fig. 5.2). Due to the large number of Clifford gates in the three-qubit Clifford group (92,897,280), it is infeasible to perform three-qubit RB (or beyond), as it is not possible to uniformly sample from the entire group in any reasonable amount of time. In Fig. 5.3, we plot some example RB results to show how the exponential fit changes when benchmarking different subsets of qubits. For example, in comparing isolated RB performed on two qubits separately to sRB performed on both qubits at the same time, we see stark differences in exponential decay rates (note the difference in the x-axis between the two subsets of plots). This indicates that the infidelity of single-qubit gates is worse when performed in parallel than in isolation; this discrepancy is likely due to crosstalk-induced coherent errors acting on both qubits under simultaneous operation. Furthermore, we note that the exponential decay for two-qubit RB is much faster than single-qubit RB, a reflection of the fact that two-qubit gates generally have a larger gate infidelity than single-qubit gates.



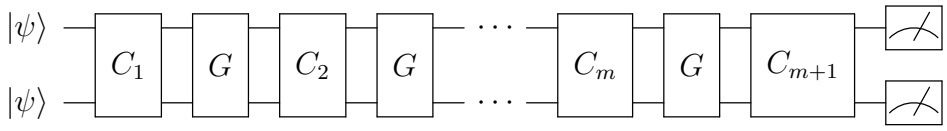
(a) Isolated single-qubit RB



(b) Simultaneous single-qubit RB



(c) Two-qubit RB

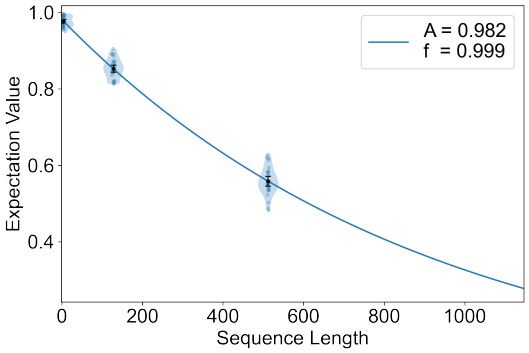


(d) Interleaved RB

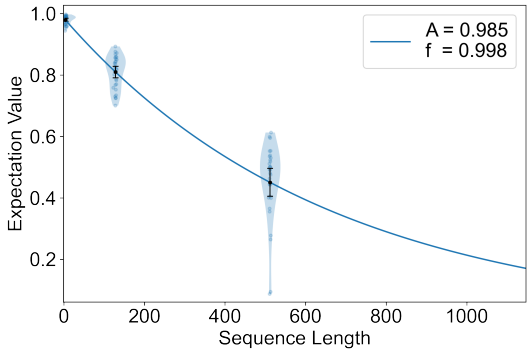
Figure 5.2: **Randomized benchmarking sequences:** (a) Isolated single-qubit gates, (b) simultaneous single-qubit gates (sRB) on qubits (1) and (2), (c) two-qubit gates, and (d) interleaved RB for a two-qubit gate G. For isolated single-qubit RB, all other qubits are assumed to be idling in the ground state. For two-qubit RB, all C_m are sampled from $\mathbb{C}_2$, which must be decomposed into available single- and two-qubit gates on the local device. For interleaved RB, G is the gate of interest whose infidelity can be estimated from (c) and (d), and C_{m+1} represents the inversion gate for the entire sequence.

Interleaved RB

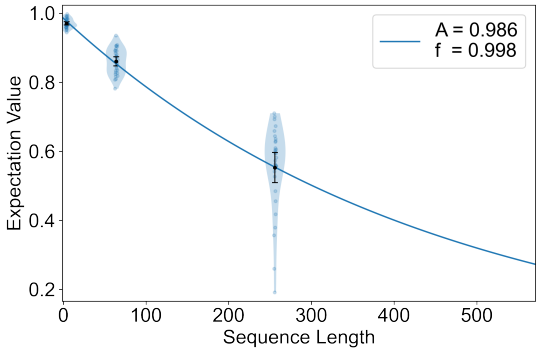
While standard RB captures the error per Clifford gate, it is possible to estimate the average error rate of an individual quantum gate G using interleaved RB (IRB) [137], where G is typically some native gate on a quantum device. By interleaving G between every layer of random Cliffords in RB (see Fig. 5.2d), the extracted average gate fidelity $\mathcal{F}_{\tilde{G}}$ now represents the fidelity of a dressed gate $\tilde{G} = G\bar{C}_n$ composed of the bare gate G and an average random n-qubit Clifford $\bar{C}_n$. However, because RB requires an inversion operator, the interleaved

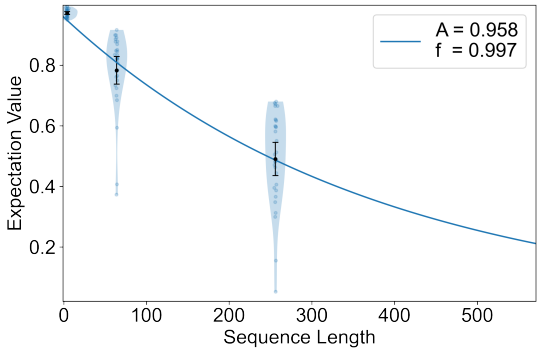


(a) Isolated RB on qubit 5

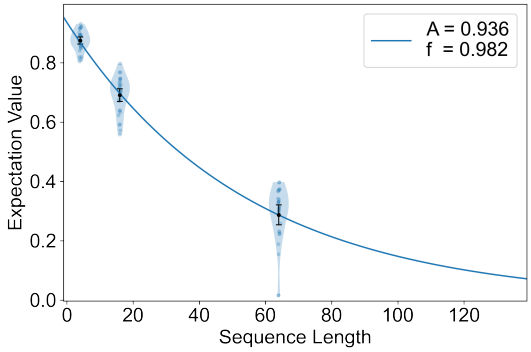


(b) Isolated RB on qubit 6





(c) Simultaneous RB on qubits 5 & 6



(d) Two-qubit RB on qubits 5 & 6

Figure 5.3: **Randomized benchmarking results.** Randomized benchmarking results on two qubits (labeled 5 & 6) for the RB circuits in Fig. 5.2.

gate G must itself be a Clifford gate (or locally-equivalent to a Clifford) in order to efficiently compute the inversion gate $C_{m+1} = (G \circ C_m \circ G \circ C_{m-1} \circ ... \circ G \circ C_2 \circ G \circ C_1)^{-1}$ for the

entire sequence. Because we add an interleaved gate for every cycle in the circuit, the EPC interleaved with a native two-qubit gate (e.g. $G = CZ$) is given as

$$r_{\mathbb{C}_2+CZ} = \frac{5}{2} r_{CZ} + \frac{33}{4} r_{\mathrm{SQ}}. \tag{5.47}$$

By measuring both RB and IRB, one may obtain an estimate of the error rate of G alone via

$$\boxed{r_G = \frac{d-1}{d}\left(1 - \frac{f_{\tilde{G}}}{f}\right)}, \tag{5.48}$$

where f ($f_{\tilde{G}}$) is the depolarizing constant measured from RB (IRB), and $d = 2^n$ (n qubits). In Fig. 5.4, we plot the RB and IRB decay curves for a controlled-Z (CZ) gate implemented on the quantum processor used in this work [60], showing how the IRB sequence decays faster than the RB sequence due to the inclusion of the interleaved gate; we estimate the CZ gate fidelity to be $\mathcal{F}_{CZ} = 1 - r_G = 99.44(9)\%$ from measured values of $f = 0.9744(9)$ and $f_{\tilde{CZ}} = 0.9672(7)$.

One issue with IRB is that the estimate of r_G is subject to a large statistical uncertainty when f and $f_{\tilde{G}}$ are comparable. According to Ref. [137], the estimate of r_G is only guaranteed in the range $[r_G - E,\ r_G + E]$, where

$$E = \min \begin{cases} \dfrac{(d-1)}{d}\left[\left|f - \dfrac{f_{\tilde{G}}}{f}\right| + (1-f)\right] \\ \dfrac{2(d^2-1)(1-f)}{fd^2} + \dfrac{4\sqrt{1-f}\sqrt{d^2-1}}{f}. \end{cases} \tag{5.49}$$

In fact, it has been shown that the upper- and lower-bounds on the fidelity estimate of IRB can span orders of magnitude [138] depending on the type of errors impacting the gate. For the CZ gate benchmarking results in Fig. 5.4, the upper- and lower-bounds on r_{CZ} were determined to be 8.1(2)% and 0.04(1)%, spanning nearly 2 orders of magnitude.

Purity Benchmarking

One disadvantage of RB is that it does not by default provide any information regarding the contribution to the total error rate due to coherent errors versus incoherent noise, since all gate errors are tailored into a global depolarizing channel. However, purity benchmarking (PB) [139] is able to separate the two by measuring the decay in the purity as a function of sequence depth. To do so, Ref. [139] introduces the *unitarity* $u(\mathcal{E})$ of a noise channel $\mathcal{E}$,

$$u(\mathcal{E}) = \frac{d}{d-1} \int d\psi \, \mathrm{Tr}\left[\mathcal{E}\left(|\psi\rangle\langle\psi| - \frac{\mathbb{I}}{d}\right)\right]^2. \tag{5.50}$$

The unitarity can also be written in terms of the unital block of the corresponding PTM Λ_u,

$$u(\Lambda) = \frac{1}{d^2-1} \mathrm{Tr}\left[\Lambda_u^\dagger \Lambda_u\right]. \tag{5.51}$$

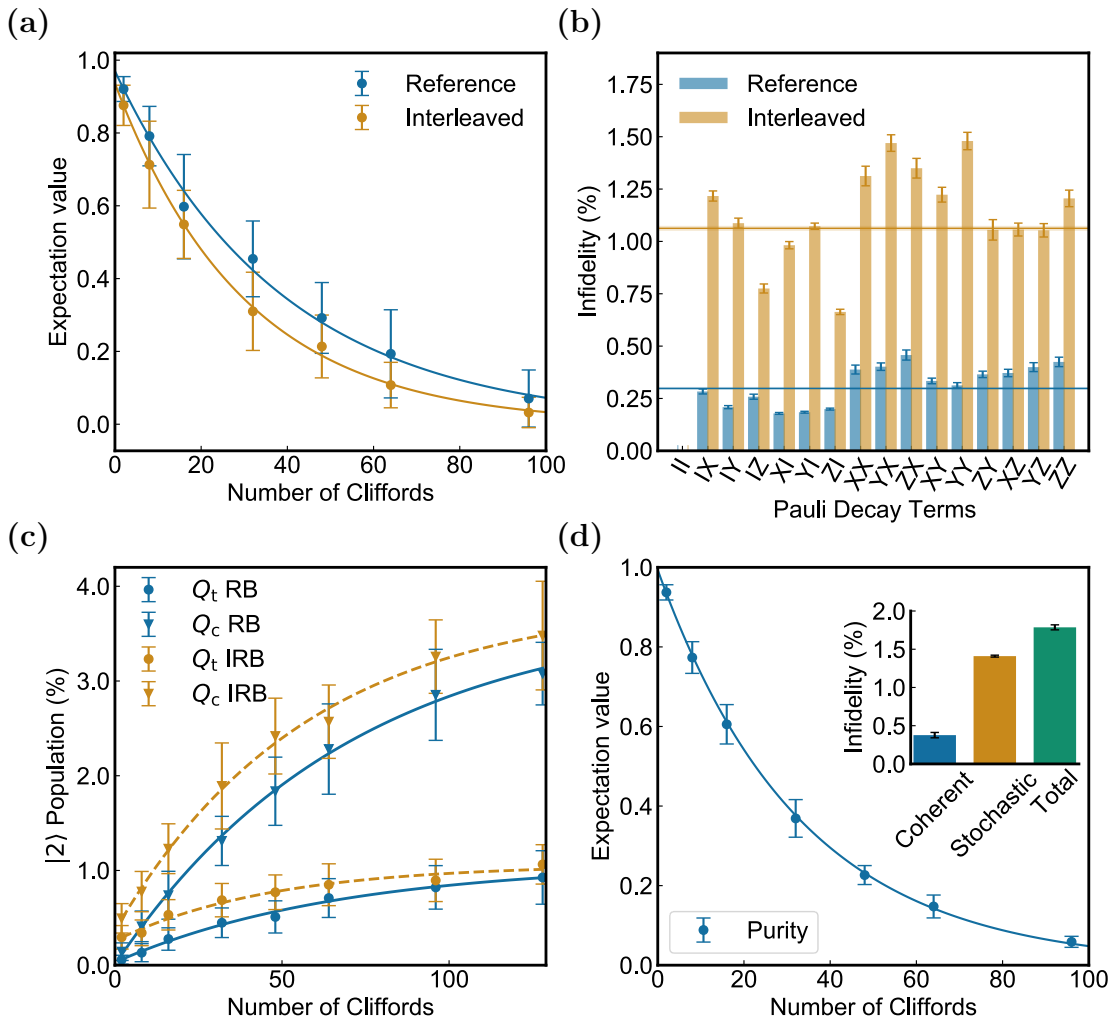


Figure 5.4: CZ **gate benchmarking results.** **(a)** Interleaved RB. Exponential decay curves for RB (Reference, blue) and IRB (Interleaved, gold), with $f_{RB} = 0.9744(9)$ and $f_{IRB} = 0.9672(7)$, respectively. From these values, we estimate $\mathcal{F}_{CZ} = 99.44(9)\%$. **(b)** Cycle Benchmarking. The y-axis lists error rate e_P for each Pauli eigenstate P (x-axis) for CB performed on the identity cycle (Reference, blue) and the CZ gate (Interleaved, gold). The horizontal lines denote the average process infidelities across all Pauli channels; from these values, we extract a gate fidelity of 99.43(1)%. **(c)** Leakage RB. By monitoring the $|2\rangle$ state population for both the target (t) and control (c) qubit (Q) during RB and IRB, we are able to fit the data to an exponential model to extract the leakage-per-gate for each qubit. **(d)** Purity Benchmarking. By measuring the average purity at the end each RB sequence, one can determine the breakdown between coherent and stochastic contributions to the total RB error rate (inset). We find stochastic noise to be the dominant source of error, showing that the gate is nearly coherence limited. (This figure has been reprinted with permission from Ref. [60].)

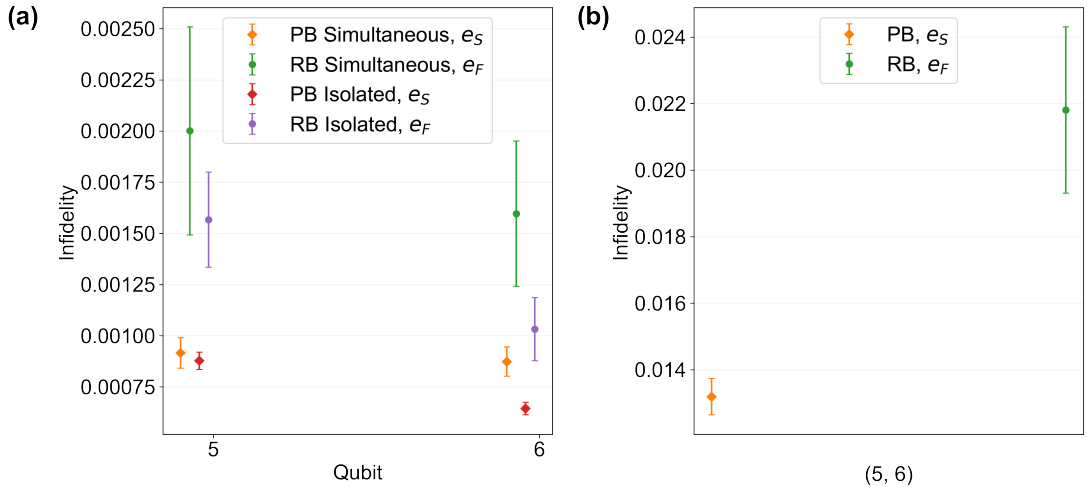


Figure 5.5: **Randomized benchmarking and purity benchmarking.** Combining RB and PB allows one to determine which fraction of the total error rate is due to coherent errors versus incoherent noise: the RB infidelity e_F is the total error rate, the PB infidelity e_S is the error rate due to incoherent (or stochastic) noise, and the difference between the two $e_U = e_F - e_S$ is the error rate due to coherent errors. (a) Isolated vs. Simultaneous RB and PB. The RB error rate is higher under simultaneous operation, whereas the infidelity from PR is approximately the same, indicating that crosstalk-induced coherent errors constitute a larger fraction of the total error rate during simultaneous gates. (b) Two-qubit RB and PB. The discrepancy between the RB and PB infidelity indicates the presence of calibration errors in the two-qubit gate. The PB infidelity represents a lower bound on the RB error rate for coherence-limited operations.

The unitarity is equivalent to the Euclidean norm (i.e. average squared length) of a d-dimensional Bloch vector after undergoing the error process $\mathcal{E}(|\psi\rangle\langle\psi|)$, excluding the identity component of the error map, which has been subtracted off. In other words, the unitarity is a generalized measure of the purity of a quantum state after applying $\mathcal{E}$. If $u(\mathcal{E}) = 1$, then $\mathcal{E}$ is a unitary channel.

To perform PB, only a small modification to the standard RB protocol is needed: after each RB sequence, we perform state tomography to estimate the purity γ of the final state ρ', where

$$\gamma = \langle\sigma_x\rangle^2 + \langle\sigma_y\rangle^2 + \langle\sigma_z\rangle^2 \tag{5.52}$$

for a single qubit, or more generally

$$\gamma = \frac{d}{d-1}||\mathbf{r}(\rho')||^2 \tag{5.53}$$

for a d-dimensional Bloch vector $\mathbf{r}$ in which the identity component has been subtracted off.

Averaging over many random sequences of length m allows one to fit the average purity to an exponential decay function,

$$\langle \gamma(m) \rangle = Au^m, \tag{5.54}$$

where u^m is the unitarity after m gates and A the SPAM parameter. Therefore, $\langle \gamma(m) \rangle$ represents that average purity after m gates, including a reduction in purity due to SPAM errors.

In Fig. 5.5, we plot RB and PB results for isolated single-qubit gates, simultaneous single-qubit gates, and two-qubit gates. The RB infidelity $e_F(\mathcal{E})$ represents the total error rate due to $\mathcal{E}$, the PB infidelity $e_S(\mathcal{E})$ represents the error rate due to incoherent (or stochastic) noise[5], and the difference between the two $e_U(\mathcal{E}) = e_F(\mathcal{E}) - e_S(\mathcal{E})$ is the error rate due to coherent errors. We see that the total RB error rate is worse for simultaneous RB than isolated RB, but the PB infidelity is approximately the same in both cases. This discrepancy is likely due to crosstalk-induced coherent errors, which make up a larger fraction of the total error budget under simultaneous operation. Furthermore, a measure of the unitarity of a noise channel can be useful for determining whether or not a quantum gate is *coherence limited* (i.e. all gate errors are due to incoherent noise, such as T_1 and T_2). Figure 5.5 shows that $e_S \neq e_F$ for both isolated single-qubit RB and two-qubit RB, suggesting that the gates are not coherence limited and that the residual coherent error contribution e_U is likely due to calibration errors, rather than crosstalk[6]. We additionally include PB results for the CZ gate in Fig. 5.4d, with the exponential decay showing a decrease in purity as a function of sequence depth; we find that the total process infidelity $e_F = 1.78(3) \times 10^{-2}$ is split between stochastic $e_S = 1.41(1) \times 10^{-2}$ and coherent $e_U = 0.37(3) \times 10^{-2}$ contributions, showing that the gate is nearly coherence limited.

Leakage Randomized Benchmarking

While PB is useful for determining the ratio of stochastic and coherent errors in an error process, it does not quantify the rate at which excitations "leak" out of the computational subspace $\{|0\rangle, |1\rangle\}$ and occupy higher energy levels (see Fig. 3.4). Leakage is a common source of error in systems whose energy spacings are not sufficiently well-separated in energy space to isolate the $|0\rangle \longrightarrow |1\rangle$ transition from transitions to higher energy levels. For example, for typical transmon qubits the anharmonicity $\alpha \in [100, 300]$ MHz, which can lead to non-negligible leakage into the $|2\rangle$ state.

By monitoring leakage into the $|2\rangle$ at the end of a benchmarking sequence via qutrit classification (Fig. 2.15), we are able to perform leakage RB (LRB) [86, 87] to quantify the

[5]The stochastic process infidelity $e_S(\mathcal{E})$ is related to the unitarity $u(\mathcal{E})$ via

$$e_S(\mathcal{E}) = 1 - \sqrt{\frac{(d^2-1)u(\mathcal{E}) + 1}{d^2}}.$$

[6]This is not entirely true for two-qubit RB, which includes simultaneous single-qubit gates in the decomposition of random Cliffords.

leakage rate (or leakage-per-gate) of our quantum gates by fitting the data to an exponential model. As an example, in Fig. 5.4c we plot the leakage rate of the control and target qubits during the CZ gate for both RB and IRB as a function of the $|2\rangle$ state population as a percentage of the total number of experimental measurements. We see that the control qubit has a much higher leakage rate than the target qubit.

Leakage can depend on many factors. For example, the strength of a driving signal can increase leakage rates (larger amplitudes equate to larger non-zero matrix elements coupling $|1\rangle$ to $|2\rangle$). Additionally, if the ω_{01} transition frequency of one qubit is near the ω_{12} transition frequency of another, then crosstalk between the two qubits under simultaneous operation can result in larger leakage rates for the second qubit. Pulse shaping is also important, since sharper pulse features can lead to the generation of higher frequency harmonics, potentially leading to more signals and thus more leakage on the quantum processor. One important pulse shaping method for single-qubit gates is the derivative reduction by adiabatic gate (DRAG) [86, 140] protocol, which adds a derivative to the quadrature component of the pulse envelop to combat either leakage or phase errors.

5.3 Cycle Benchmarking

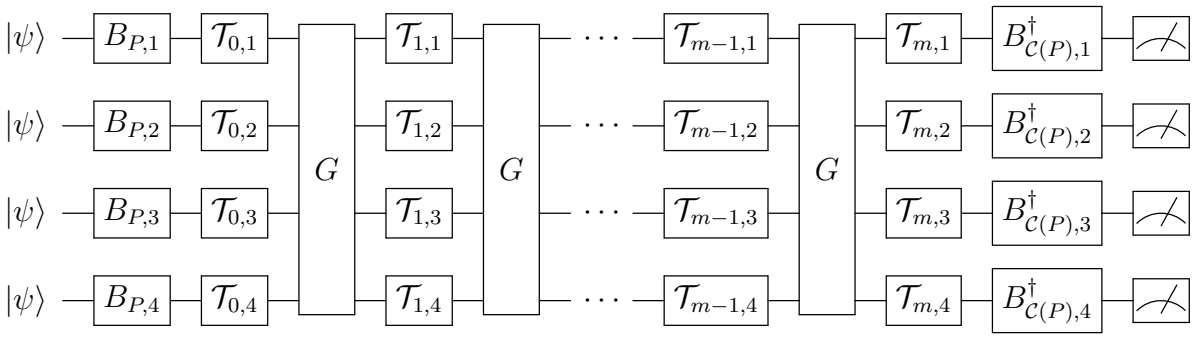
Many different variants of RB have been developed over the years proceeding the development of the original method in Ref. [131], and we could exhaust ourselves discussing the different methods and their various nuances. However, despite RB's tremendous success and continued popularity within the quantum computing community, it has serious limitations. First and foremost, RB is not scalable beyond two qubits due to the vast size of the three-qubit Clifford group. Furthermore, RB is limited to isolated single- or two-qubit gates, or capturing the marginalized (not global) performance of such gates under simultaneous operation. For these reasons, we instead turn to cycle benchmarking (CB) [141], a scalable protocol that measures errors affecting all qubits during any parallel gate cycle G. In Protocol 2, we outline the basic procedure for performing CB, and depict the basic sequence structure in Fig. 5.6a. CB differs from RB in two key ways: (1) it utilizes Pauli twirling instead of Clifford twirling, which maps gate errors into stochastic Pauli channels (Eq. 3.36) instead of a global depolarizing channel (Eq. 3.35); (2) CB benchmarks the global performance of quantum gates performed in parallel, providing a measure of their infidelity in the context of multi-qubit quantum algorithms. In contrast, benchmarking the individual constituent gates of multi-qubit cycles has been shown to be a poor predictor of the global performance of quantum circuits [142] due to the presence of coherent errors and crosstalk between qubits, and because such benchmarks fail to capture errors on (or incurred by) idling spectator qubits [75].

As outlined in Protocol 2, under CB the process fidelity of any parallel gate cycle G can be measured by preparing the system in a Pauli basis state P, interleaving the cycle of interest between cycles of n-qubit Pauli operators $\mathcal{T}$ randomly sampled from the full n-qubit Pauli group $\mathbb{P}_n$, and measuring the error rate as a function of sequence depth. For each

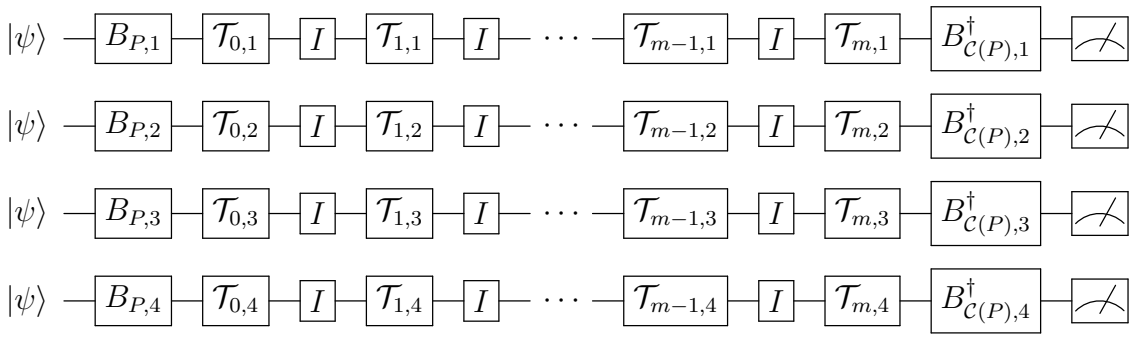
Protocol 2 Cycle Benchmarking

1. Select K random Paulis P from the n-qubit Pauli group $\mathbb{P}_n$. The set $P \in \mathcal{P}$ constitutes the preparation and measurement basis states, with $K = |\mathcal{P}|$.
2. Select a sequence length m at which the application of the interleaved gate or cycle G returns to the identity, $G^m = \mathbb{I}$.
3. For each random Pauli P, do:
   a) Select $m+1$ random Pauli cycles $\mathcal{T}_0, ..., \mathcal{T}_{m-1}, \mathcal{T}_m$, and construct the following circuit (shown in Fig. 5.6a),
   $$\mathcal{C}(P) = B^{\dagger}_{\mathcal{C}(P)} \circ \mathcal{T}_m \circ G \circ \mathcal{T}_{m-1} \circ G \circ ... \circ G \circ \mathcal{T}_0 \circ B_P,$$
   where $\mathcal{T}$ denotes the cycles of twirling operators, B_P prepares the Pauli state P, and $B^{\dagger}_{\mathcal{C}(P)}$ rotates the system back to P at the end of the sequence.
   b) Measure $\mathcal{C}(P)$ N times (e.g. $N = 1000$ "shots") and compute the overlap
   $$f_{P,m} = \mathrm{Tr}\Big[\mathcal{C}(P)\tilde{\mathcal{C}}(\rho)\Big] \tag{5.55}$$
   between the ideal result of $\mathcal{C}(P)$ and the noisy implementation $\tilde{\mathcal{C}}(\rho)$, where ρ is a $+1$-eigenstate of P.
   c) Repeat for L different random sequences (e.g. $L = 20 - 30$) and compute the average overlap $\sum_{l=1}^{L} f_{P,m,l}$.
4. Do steps 2 - 3 for at least two different sequence lengths m which satisfies $G^m = \mathbb{I}$.
5. Compute the average process fidelity from any two sequence lengths $m \in [m_1, m_2]$:
   $$F = \frac{1}{K} \sum_{P \in \mathcal{P}} \left(\frac{\sum_{l=1}^{L} f_{P,m_2,l}}{\sum_{l=1}^{L} f_{P,m_1,l}} \right)^{\frac{1}{m_2 - m_1}}. \tag{5.56}$$

preparation/measurement basis P (i.e. Pauli channel), an exponential decay of the form Af^m can be fit to the dressed cycle (composition of the cycle of interest with the random Pauli gates), where A is the SPAM parameter, f is the decay parameter, and m is the sequence length. Therefore, the exponential decays in CB are often referred to as "Pauli decays" and are labeled by the basis preparation and measurement state P.



(a) CB on the gate cycle G



(b) CB on the all-identity reference cycle

Figure 5.6: **Cycle benchmarking sequences:** (a) Four-qubit gate cycle G, and (b) the all-identity "reference" cycle. For (a), the gate cycle G can be composed of any combination of single- and multi-qubit gates, as long as $G^m = \mathbb{I}$ for a sequence depth of m. $\mathcal{T}_{n,a}$ denotes the nth twirling operator acting on qubit q. $B_{P,q}$ denotes the basis preparation gate on qubit n for the Pauli P, and $B^\dagger_{\mathcal{C}(P),q}$ rotates qubit q back to the eigenstate of P at the end of the sequence. For (b), the all-identity cycle has been inserted for visual clarity, but this cycle is either skipped in compilation, or the identity gates can be implemented as true idles for the normal duration of a single-qubit gate; the choice is up to the experimenter.

In effect, CB measures the eigenvalues Λ_{PP} of the Pauli transfer matrix (PTM) of G. For example, CB performed on a two-qubit cycle would produce a PTM with the following diagonal components,

$$\Lambda = \begin{pmatrix} 1 & & & & & & & & & & & & & & & \\ & f_{IX}^m & & & & & & & & & & & & & & \\ & & f_{IY}^m & & & & & & & & & & & & & \\ & & & f_{IZ}^m & & & & & & & & & & & & \\ & & & & f_{XI}^m & & & & & & & & & & & \\ & & & & & f_{XX}^m & & & & & & & & & & \\ & & & & & & f_{XY}^m & & & & & & & & & \\ & & & & & & & f_{XZ}^m & & & & & & & & \\ & & & & & & & & f_{YI}^m & & & & & & & \\ & & & & & & & & & f_{YX}^m & & & & & & \\ & & & & & & & & & & f_{YY}^m & & & & & \\ & & & & & & & & & & & f_{YZ}^m & & & & \\ & & & & & & & & & & & & f_{ZI}^m & & & \\ & & & & & & & & & & & & & f_{ZX}^m & & \\ & & & & & & & & & & & & & & f_{ZY}^m & \\ & & & & & & & & & & & & & & & f_{ZZ}^m \end{pmatrix}, \tag{5.57}$$

where f_P^m is the eigenvalue of the Pauli basis P at a sequence depth of m (for simplicity we label the eigenvalues by Λ_P instead of Λ_{PP}, since they are diagonal in the Pauli basis). The process fidelity of the above PTM can be computed as

$$F = \frac{1}{4^n} \sum_{P \in \mathbb{P}_n} \Lambda_{PP} = \frac{1}{16} \sum_{P \in \{I,X,Y,Z\}^{\otimes 2}} f_P, \tag{5.58}$$

where we have set $m = 1$. In practice it is not necessary to measure all 4^n Pauli operators, as the precision of the fidelity estimate is set by the number of Pauli channels $K \leq 4^n$ (n qubits) that are sampled out of the full 4^n possible states [141]. This is justified by the fact that for any CPTP map, the process infidelity $e_F = 1 - F$ sets the upper bound on the infidelity of any given Pauli channel,

$$0 \leq 1 - \Lambda_{PP} \leq 2e_F, \tag{5.59}$$

so statistically each Pauli infidelity $e_P = 1 - \Lambda_{PP}$ cannot deviate significantly from the average. Therefore, while Eq. 5.56 can be used to compute the process fidelity from two different sequence depths $[m_1, m_2]$, by fitting each Pauli channel to an exponential decay, the total process fidelity under CB can also be computed as the weighted average over all K Pauli channels for $m = 1$,

$$F = \frac{1}{K} \sum_{P \in \mathcal{P}} f_P. \tag{5.60}$$

Much like IRB, in which the target gate is interleaved between random Clifford gates, CB interleaves the target cycle between cycles of random single-qubit Pauli gates. Therefore, CB measures the process fidelity of a dressed cycle $G\bar{\mathcal{T}}$ composed of the cycle (or gate) of interest G and an average Pauli twirling cycle $\bar{\mathcal{T}}$,

$$F_D\left(G\bar{\mathcal{T}}, \tilde{G}\bar{\tilde{\mathcal{T}}}\right) = \frac{1}{K} \sum_{P \in \mathcal{P}} f_P\left(G\bar{\mathcal{T}}, \tilde{G}\bar{\tilde{\mathcal{T}}}\right) \tag{5.61}$$

$$= \frac{1}{K}\frac{1}{L} \sum_{P \in \mathcal{P}} \sum_{l=1}^{L} \mathcal{T}_l^\dagger G(P)^\dagger \tilde{G}(P) \tilde{\mathcal{T}}_l \tag{5.62}$$

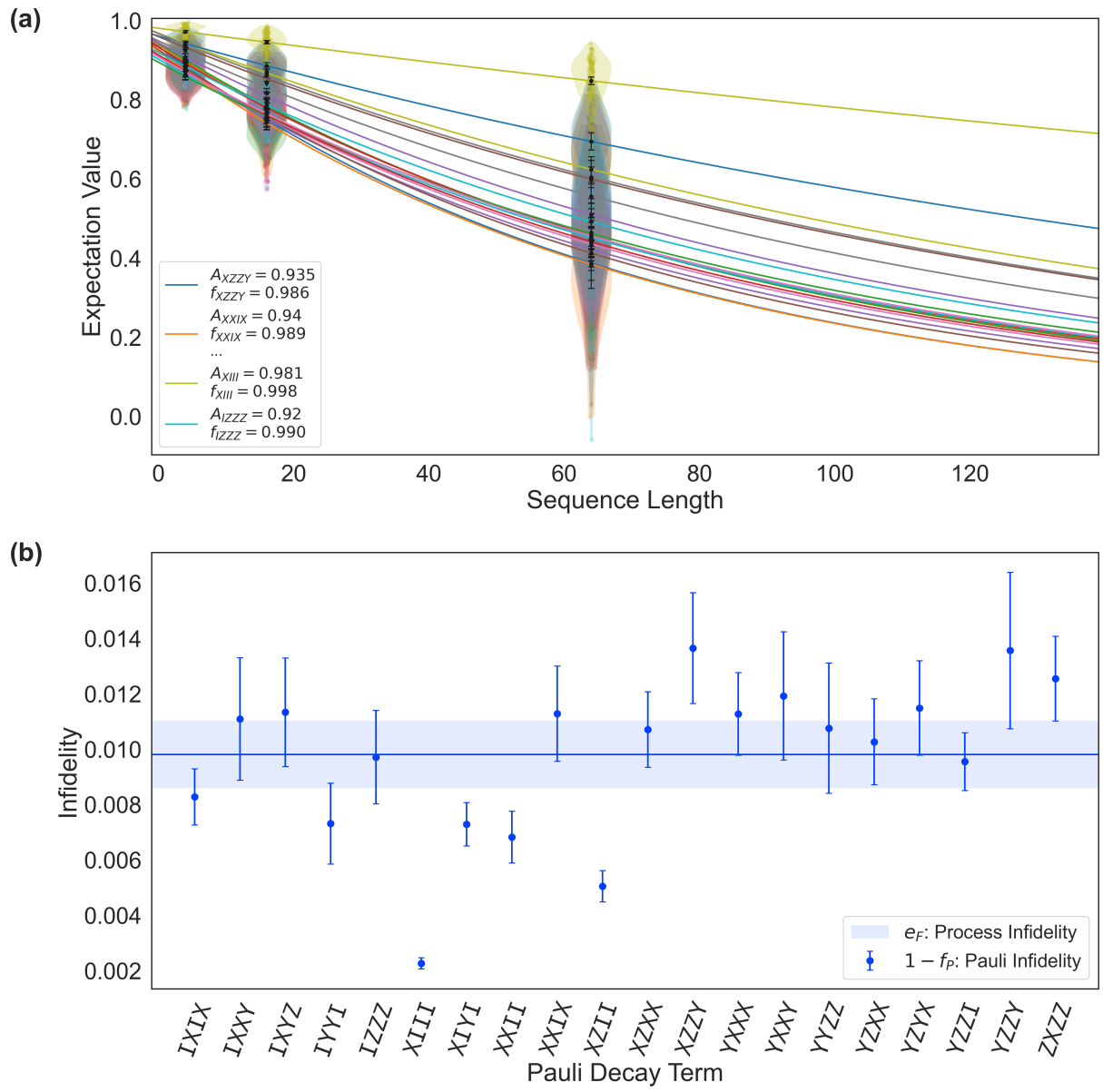


Figure 5.7: **Cycle Benchmarking of the all-identity "reference" cycle for four qubits.** (a) Pauli decays for each Pauli channel P, with the SPAM parameter A_P and the exponential fit f_P listed in the legend for a subset of P. (b) Pauli infidelities $e_P = 1 - f_P$ for each Pauli channel, and the average process infidelity e_F. The highlighted region denotes the 95% confidence interval of e_F.

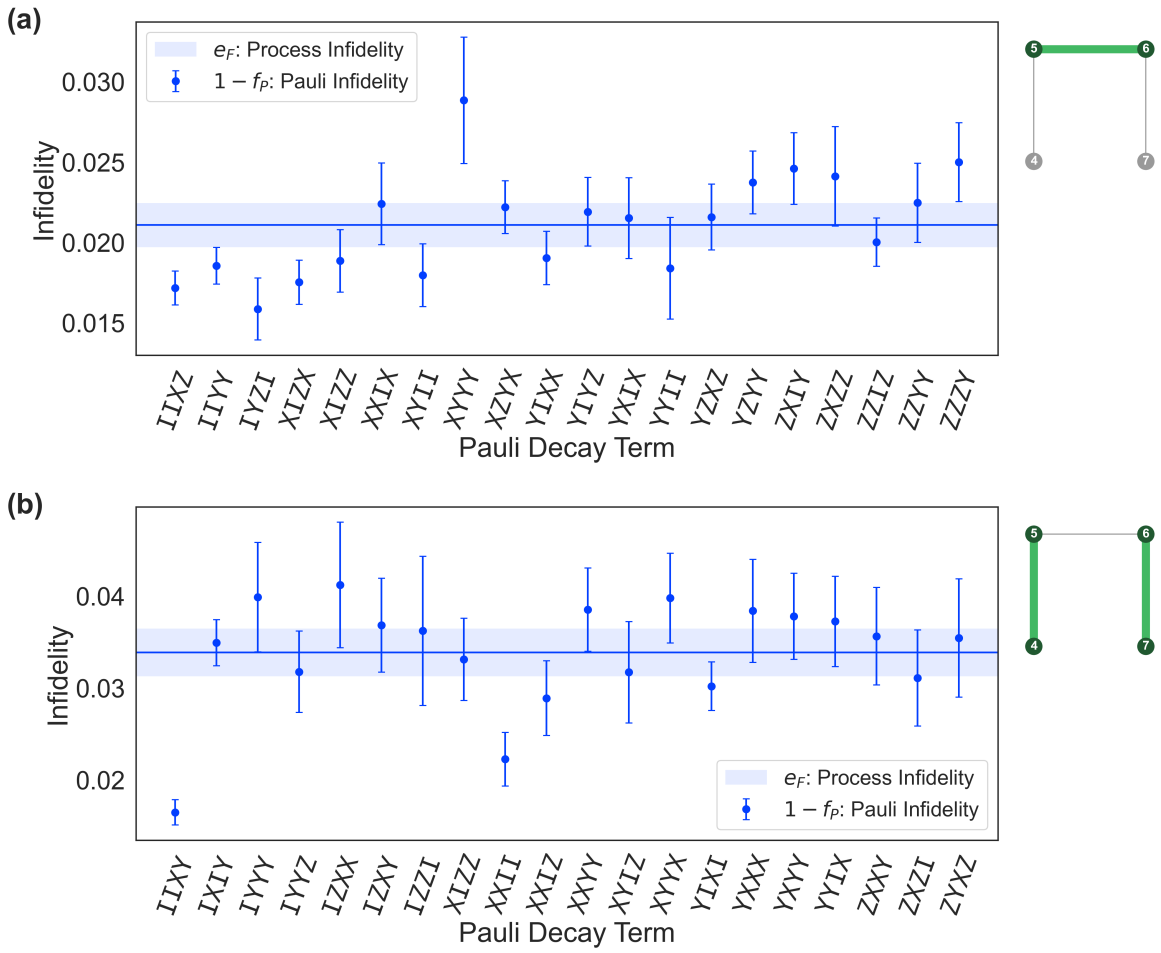


Figure 5.8: **Cycle Benchmarking of multi-qubit cycles containing CZ gates.** (a) CZ gate between qubits (5, 6), with qubits 4 and 7 idling. (b) Simultaneous CZ gates between qubits (4, 5) and (6, 7).

where $G(P)$ denotes the gate prepared in the Pauli basis P, and the tilde denotes the noisy implementation of the gates. Therefore, the process infidelity of the dressed cycle $e_D = 1 - F_D(G\bar{\mathcal{T}}, \tilde{G}\bar{\tilde{\mathcal{T}}})$ contains the errors due to the interleaved target cycle G as well as the Pauli twirling gates $\mathcal{T}$. In the limit that the twirling operators can be perfectly implemented, then the process infidelity $e_F = 1 - F(G, \tilde{G})$ is given in terms of the fidelity of the cycle of interest alone. However, in any realistic scenario the infidelity of the twirling operators will contribute to the total process infidelity. Therefore, to separate the infidelity of the interleaved cycle from the twirling gates, we measure the CB fidelity of the "all-identity" reference cycle (see Figs. 5.6b and 5.7), which equates to benchmarking the average performance of only the Pauli twirling gates. Similar to IRB, we can use this to estimate

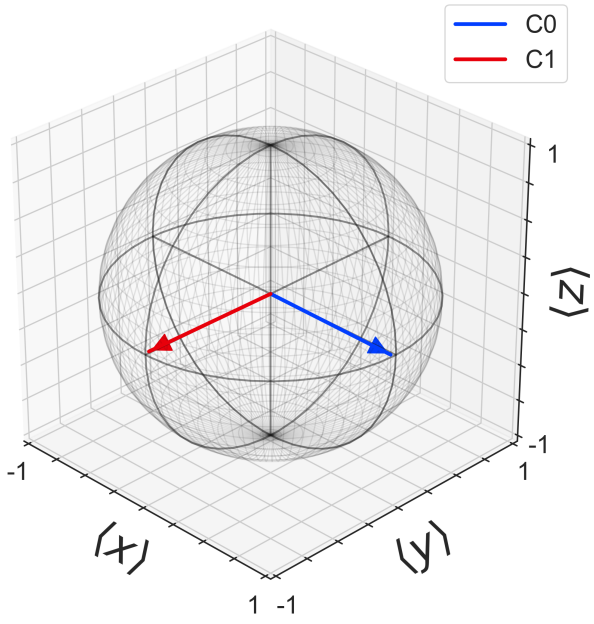


Figure 5.9: $CS^\dagger$ **gate.** State tomography performed on the target qubit at the end of the $CS^\dagger$ = CPHASE(-90) gate for the target qubit initially prepared in the $|+\rangle$ state. When the control qubit is prepared in $|0\rangle$ (C0), the target qubit remains in $|+\rangle$; when the control qubit is prepared in $|1\rangle$ (C1), the target qubit is rotated -90 degrees to $|i-\rangle$. The CS gate performs the opposite rotation, with the target qubit ending in $|i+\rangle$ when the control is prepared in $|1\rangle$.

the process infidelity of the interleaved cycle e_G by taking the ratio of the process fidelities of the dressed (D) and reference (I) cycles,

$$e_G = \frac{d-1}{d}\left(1 - \frac{F_D}{F_I}\right), \tag{5.63}$$

where $d = 2^n$. Using CB has been shown to tighten the upper- and lower-bounds on the fidelity estimate of the interleaved cycle relative to IRB [60], which can span orders of magnitude [138]. For example, we measure a lower [upper] bound on the fidelity estimate of $F_{CZ} = 97.52(2)\%$ [$F_{CZ} = 99.764(5)\%$] for the CZ gate shown in Fig. 5.4b, compared to the lower [upper] bound of $F_{CZ} = 91.9(2)\%$ [$F_{CZ} = 99.96(1)\%$] measured via IRB.

While RB requires that the native two-qubit gate be locally-equivalent to a Clifford, CB has no such requirement. Rather, CB has two basic requirements: (1) that the gate cycle G performed m times composes the identity, $G^m = \mathbb{I}$, and (2) the ability to prepare and measure in the same Pauli basis state P. These basic requirements are easiest to satisfy if the interleaved gate cycle contains only Clifford gates, making it trivial to determine which basis operators $B^\dagger_{\mathcal{C}(P)}$ are required to rotate the system back to P at the end of the

Gate / Qubits:		(Q4, Q5)	(Q5, Q6)	(Q6, Q7)
CZ	RB e_F (1×10^{-2})	1.9(1)	2.04(8)	1.95(6)
	CB e_D (1×10^{-2})	1.09(1)	1.05(1)	1.26(1)
	CB e_{CZ} (1×10^{-3})	5.8(1)	4.8(1)	5.9(2)
CS	CB e_D (1×10^{-2})		0.98(1)	
	CB e_{CS} (1×10^{-3})		4.3(1)	
$CS^\dagger$	CB e_D (1×10^{-2})	0.98(1)		0.91(1)
	CB $e_{CS^\dagger}$ (1×10^{-3})	5.0(1)		3.3(1)
Ref.	CB e_I (1×10^{-3})	3.24(5)	4.12(8)	4.8(1)

Table 5.2: **Benchmarking non-Clifford gates with CB.** CB performed on Clifford (CZ) and non-Clifford (CS and $CS^\dagger$) two-qubit gates on the same pairs of qubits. We include the RB error rate for the CZ for completeness, as well as the reference infidelity used to compute the interleaved gate infidelities. (Table reprinted with permission from Ref. [143].)

sequence. However, these requirements do not exclude benchmarking non-Clifford gates with CB. In Table 5.2, we include CB results for Clifford (CZ) and non-Clifford [$CS = \sqrt{CZ}$ = CPHASE(90) and $CS^\dagger$ = CPHASE(-90), see Fig. 5.9] two-qubit gates. We note that with the CS and $CS^\dagger$ gates, requirement (1) is trivial to satisfy for m being an integer multiple of four rotations. While satisfying requirement (2) is less trivial and sometimes requires compiling multi-qubit gates into the final cycle $B^\dagger_{\mathcal{C}(P)}$ in order to rotate the full system back to P, the cost of doing so is beneficial as long as the fidelity of the multi-qubit gates are reasonably high so that the final cycle of gates does not add significantly to the estimated process fidelity of the interleaved cycle. As a concrete example, Ref. [144] reports benchmarking a fidelity of 98.26(2)% for a three-qubit iToffoli gate.

The utility of CB is that it allows one to benchmark the errors across an entire register of qubits, including spectator qubits that should be idling. For example, Table 5.2 shows that the error rate of a CZ gate performed on qubits (5, 6) is $e_{CZ} = 0.48(1)\%$. However, in Table 5.3 we instead show CB results of a CZ gate performed on qubits (5, 6) with qubits 4 and 7 idling, for which the error rate was estimated to be $e_G = 1.09(9)\%$. We see that by including the qubits that are not supposed to be operated on during the two-qubit gate, we see a $2\times$ increase in the error rate. This type of information is not captured by RB and related methods, which marginalize errors to be local on the benchmarked qubits. In Table 5.3 we also show the benchmarked error rates of various other parallel gate cycles, such as the all-Hadamard cycle for preparing a register of qubits in a superposition state, simultaneous $CS^\dagger$ gates, and cycles involving SWAP gates, which decompose into three native CZs on the device. These results demonstrate the power of CB for characterizing the errors across a full quantum quantum processor. In fact, the complexity of fully characterizing an n-qubit

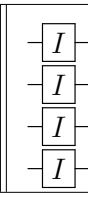
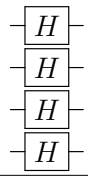
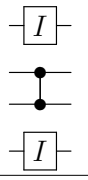
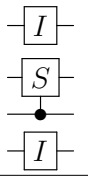
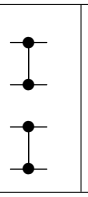
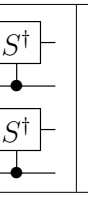
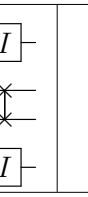
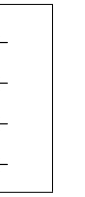
Interleaved Cycle G:								
e_I (1×10^{-3})	9.6(6)							
e_D (1×10^{-2})		1.16(6)	2.11(7)	1.67(8)	3.4(1)	2.09(8)	6.3(2)	10.4(4)
e_G (1×10^{-2})		0.19(8)	1.09(9)	0.68(9)	2.3(1)	1.07(9)	5.1(2)	9.0(4)

Table 5.3: **CB performed on various different parallel gate cycles G.** Examples include (from left to right) the all-identity reference cycle, the all-Hadamard cycle, a CZ gate between idling qubits, a CS gate between idling qubits, parallel CZ gates, parallel $CS^\dagger$ gates, a SWAP gate between idling qubits, and parallel SWAP gates. The (measured) dressed e_D and (estimated) bare e_G infidelities are listed for each cycle. (Table reprinted with permission from Ref. [143].)

system via CB is only polynomially expensive, because it depends only on the number of distinct cycles one wishes to characterize for any given application.

5.4 Cycle Error Reconstruction

Cycle benchmarking measures the eigenvalues of the PTM of some physical process (Eq. 5.57). This provides information about what Pauli channels are impacted by the error process (see the state preservation block of the PTM in Fig. 3.4), but this does not immediately provide information about which types of errors are causing the degradation of the survival probability of each Pauli channel. However, as discussed at the end of Section 3.5, there exists a relationship between Pauli channel eigenvalues and the Kraus operators that cause errors – namely, the survival probability Λ_{PP} will be impacted by any Pauli error Q which does not commute with P (see Eq. 3.55). In this section, we introduce a cycle error reconstruction (CER) protocol [145, 146, 147] (also called k-body noise reconstruction [148]) based on CB measurements to produce an error map of the Pauli error rates in our system.

Fig. 5.10a outlines the process by which CB can be used to reconstruct k-body gate errors that occur during any gate cycle. CER results are based on targeted CB measurements in which specific Pauli decays are chosen to estimate the error rates afflicting subsets of the gates or idle qubits in the specific cycle of interest. The Pauli error rates estimated by CER are the coefficients p_P in the Kraus representation of a Pauli channel in Eq. 3.31. One can convert between the Pauli eigenvalues f_P from CB and the Kraus coefficients p_P using a linear transformation,

$$\mathbf{f}_P = \mathbb{W}\mathbf{p}_P, \tag{5.64}$$

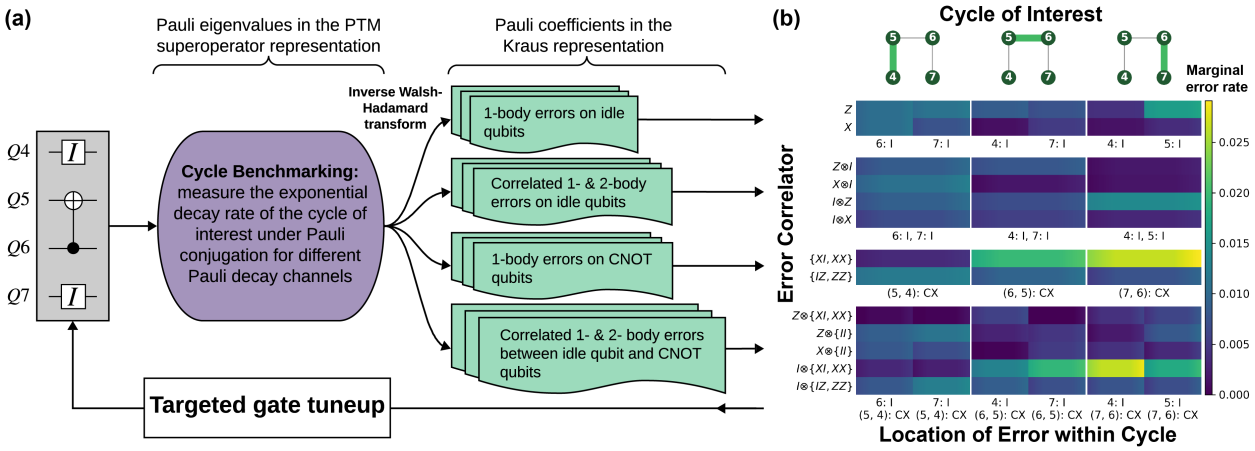


Figure 5.10: **Cycle error reconstruction of the Pauli noise measured via cycle benchmarking.** (a) Schematic of the process by which 1- and 2-body gate errors can be reconstructed using targeted CB measurements of a parallel gate cycle (e.g. CNOT between Q5 and Q6, with Q4 and Q7 idling), providing detailed information about the Pauli (Kraus) errors occurring during the cycle, as shown in (b). (b) CER results of four-qubit cycles containing a single CNOT gate and identity gates on the spectator qubits. The y-axis (x-axis) labels the type of error (the intended gate acting on the qubit), and the color (gradient) indicates the marginal error rate from all Pauli contributions (95% confidence interval). The first (third) row of subplots shows 1-body errors acting in idling (CNOT) qubits, the second row of subplots shows correlated 1- and 2-body errors between idling qubits, and the last row of subplots shows correlated 1- and 2-body errors between an idling spectator qubit and qubits participating in the CNOT. Here, a k-body error is labeled by k non-identity Paulis acting on n qubits (e.g. $k = 1$ for CNOT qubits, because these errors occur within a single gate body acting on 2 qubits). The tensor notation $\otimes$ between k-body errors indicates correlators between product states, whereas the lack of tensors indicates errors on entangled qubits. Curly brackets indicate error types that cannot be distinguished due to degeneracies, since any local error acting on either qubit in the entangling operation will be transformed by the CNOT. All rows in which all errors are below 30% of the maximum value have been omitted for clarity. This detailed information can be used to perform targeted gate tuneup to address the most harmful errors. The residual errors in our system are broadly distributed among many pathways, so any further targeted tuneup will come with diminishing returns. (Figure reprinted with permission from Ref. [145].)

where $\mathbf{f}_P$ is the vector of Pauli eigenvalues f_P, $\mathbf{p}_P$ is the vector of Pauli error rates, and $\mathbb{W}$ is an n-qubit Walsh-Hadamard transform (see Ref. [146], Section B for further details); $\mathbb{W}$ is a generalization of the transformation given in Eq. 3.55. The Pauli decays in CB are dual to the Pauli operators which cause errors, so to measure the error rate p_P of some fixed Pauli P, we measure a set of Pauli decays that commute and anti-commute with P, and then use this

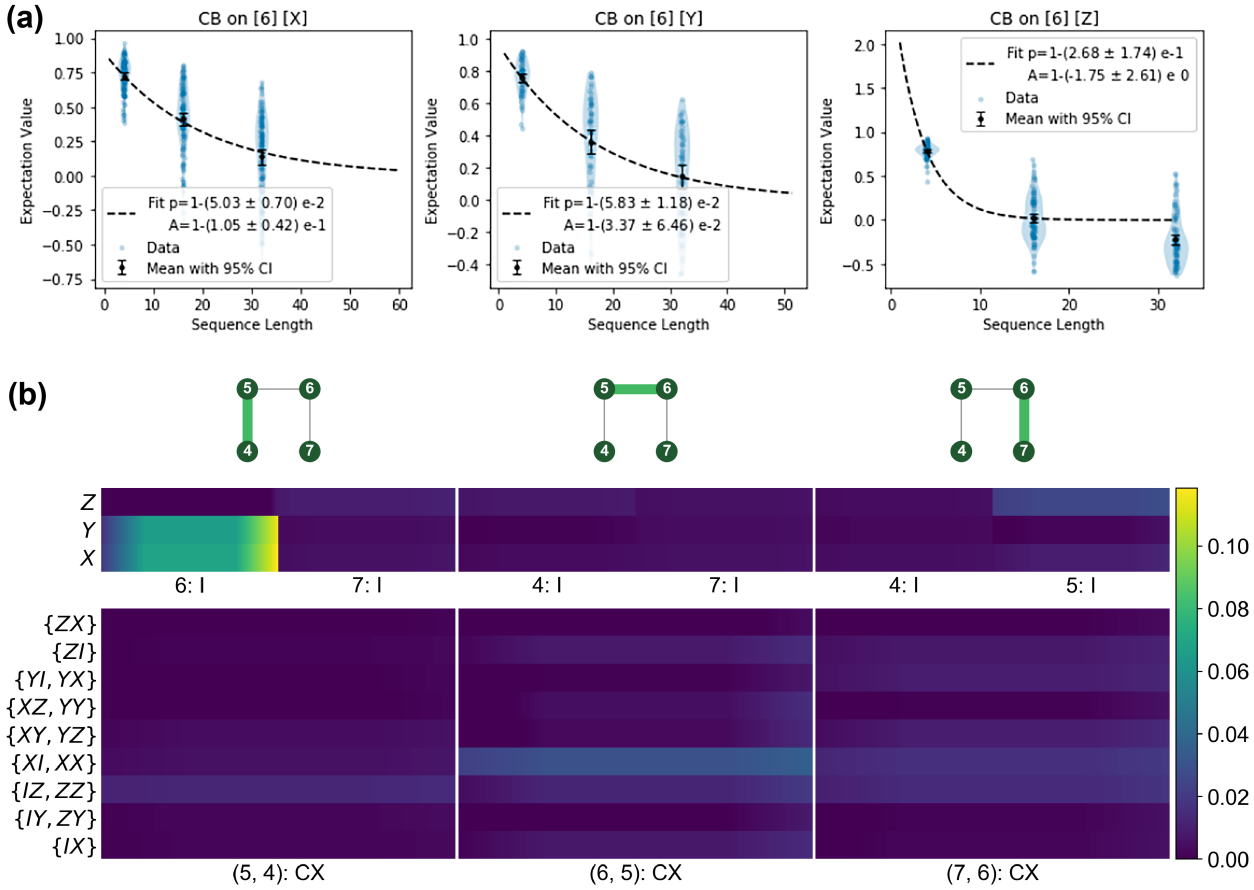


Figure 5.11: **Duality of Pauli decays to Pauli error rates.** (a) CB results marginalized over qubit 6 showing the different rates of decay for the X, Y, and Z channels during the CNOT between qubits 4 and 5. The sharp decay in Z suggests that the dominant errors are due to non-commuting Pauli operators (e.g. X or Y). (b) CER results show that the most dominant Pauli errors for all multi-qubit cycles in our system were due to both X and Y errors on qubit 6 during the CNOT between 4 and 5.

info to reconstruct the error of P via linear inversion: $\mathbf{p}_P = \mathbb{W}^{-1}\mathbf{f}_P$. In Fig. 5.10b, we plot 1- and 2-body Pauli (Kraus) errors reconstructed via CER for the gate cycles containing two-qubit CNOT $= CX$ gates in our system. Using CER, we identify the major sources of errors in our system and compensate the most harmful effects with targeted decoupling pulses or virtual phase gates (see Supplemental Material of Ref. [145] for more details.) An example of the duality of Pauli decays to Pauli errors is given in Fig. 5.11.

As previously discussed, the cost of characterizing a full quantum processor via CB depends on the number of distinct cycles one wishes to characterize. For this reason, we restrict the number of possible two-qubit gate cycles in our system by serializing all single-

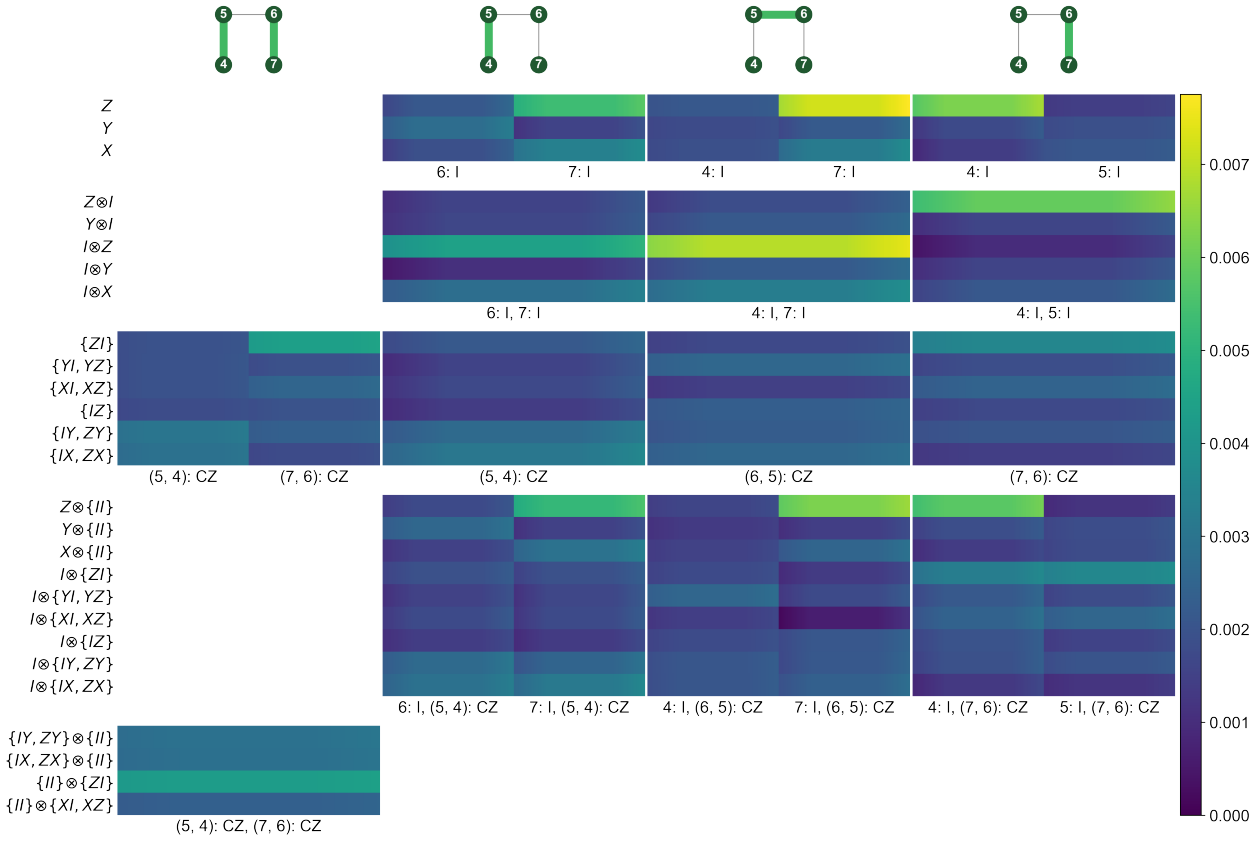


Figure 5.12: **Full processor characterization via cycle error reconstruction.** We characterize the four distinct two-qubit gate cycles available on our processor to produce an error map of the dominant errors in our system. We see that the most dominant errors in our system are 1-body errors (i.e. weight-1 errors on single-qubit gates, weight-1 or -2 errors on two-qubit gates, etc.), and that 2-body errors are largely suppressed in comparison. The largest observable error in our entire system is a Z error on qubit y during the CZ gate between qubits (5, 6). (Figure reproduced with permission from Ref. [149].)

qubit gates with respect to two-qubit gates; in other words, single- and two-qubit gates are never performed in the same cycle. For the four-qubit processor used in this work, the possible distinct two-qubit gate cycles are: (1) CZ gate between (4, 5), with qubits 6 and 7 idling; (2) CZ gate between (5, 6), with qubits 4 and 7 idling; (3) CZ gate between (6, 7), with qubits 4 and 5 idling; and (4) simultaneous CZ gates between qubits (4, 5) and (5, 6). In Fig. 5.12, we plot a complete error map of our quantum processor using CER, and show that the most dominant error in our system is not an error acting on a qubit involved in a two-qubit gate; rather, it is a Z error on qubit 7 during the CZ gate between (5, 6). This type of information cannot be gathered from conventional randomized benchmarking methods.

By limiting measurements to only 1- and 2-body errors, and making use of marginalized probability distributions, it is not necessary to measure all $K = 4^n$ Pauli decay channels under CER. More generally, it is possible to reconstruct all 1-body marginals with $\mathcal{O}(1)$ Pauli decays and all 2-body marginals with $\mathcal{O}[\log(n)]$ Pauli decays for n qubits. However, while some k-body errors can efficiently be estimated for $k > 2$, it becomes exponentially expensive to estimate all n-body errors. The results in Fig. 5.12 show that the dominant residual errors in our system are 1-body errors. Therefore, even though only 1- and 2-body errors were measured, we are justified in only measuring $k \leq 2$-body errors, since we observe that 2-body terms are largely suppressed compared to 1-body terms. Moreover, the probabilities of 2-body errors are the sum of the probabilities of all errors that act non-trivially on the corresponding two bodies, irrespective of their action on other bodies. Therefore, the fact that 2-body errors are negligible shows that 3- or more body errors are also negligible.

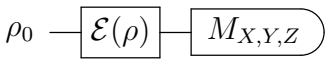
5.5 Gate Set Tomography

Gate set tomography (GST) [79, 150, 151, 152, 153] is a robust method for tomographically reconstructing errors and noise impacting all gate operations within a defined gate set. GST was developed to improve upon quantum process tomography (QPT) [135], which itself grew out of quantum state tomography (see Fig. 5.13, which outlines the differences in these methods). Like traditional QPT, GST fully characterizes the process matrix of a quantum gate; however, it does so in a self-consistent manner which decouples state-preparation and measurement (SPAM) errors from gate errors. By decomposing SPAM gates in terms of native operations, and by amplifying a noise process $\mathcal{E}(\rho)$ many times for better accuracy, GST fully characterizes the set $\{|\rho\rangle\rangle, \langle\langle M|\} \cup \{G_i\}$, where $|\rho\rangle\rangle$ is the initial state, $\langle\langle M|$ is the POVM, and $\mathbb{G} = \{G_i\}$ the set of native gate operations.

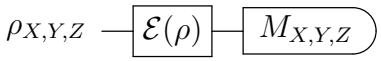
In Fig. 5.13d, all gates are decomposed in terms of the native operations in $\mathbb{G}$, including the gates $\mathbb{F} = \{F_i\}$ needed to prepare and measure in the Pauli basis states (termed the preparation and measurement "fiducials"). The set $\mathbb{F}$ constitutes the SPAM gates, such that $F_i |\rho\rangle\rangle = |\rho_i\rangle\rangle$ and $\langle\langle M_i| = \langle\langle M| F_i$. Therefore, a process matrix Λ_i for each gate $G_j \in \mathbb{G}$ can be reconstructed by measuring $\langle\langle M_i| G_j |\rho_k\rangle\rangle$ for all $F_i, F_k \in \mathbb{F}$. This means that all gates are self-consistently characterized, including SPAM, which allows GST to characterize a noise process $\mathcal{E}(\rho)$ in a manner that is independent of SPAM. An example of a typical native two-qubit gate set is $\mathbb{G} = \{G_1 \otimes G_2 : G_1, G_2 \in \{I, X_{\pi/2}, Y_{\pi/2}\}\} \cup \{CX\}$, where the CX gate can be trivially replaced with any locally-equivalent two-qubit gate. In Protocol 3, we describe the GST protocol for a single qubit[7].

Using the open-source Python library `pyGSTi` [154, 155], one can obtain a best-fit model of the process matrix for each gate in $\mathbb{G}$ using maximum-likelihood estimation [152, 153]. The quality of the model is quantified by computing the log-likelihood ratio λ of the likelihood

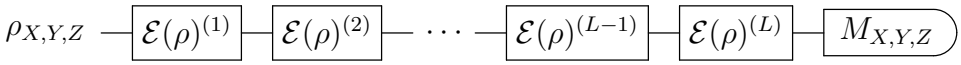
[7] This follows Ref. [79].



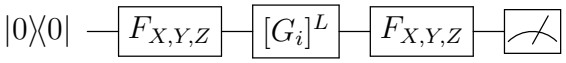
(a) Quantum State Tomography



(b) Quantum Process Tomography



(c) Gate Set Tomography, pulse amplification



(d) Gate Set Tomography, SPAM independence

Figure 5.13: **Tomographic reconstruction methods.** (a) Quantum state tomography characterizes the final state of a noisy process $\mathcal{E}(\rho)$ by measuring in all Pauli basis states $P \in \mathbb{P}_n$ for an initial n-qubit state ρ_0, typically taken to be the ground state. For a single qubit, this equates to preparing in $|0\rangle\langle 0|$ and measuring in X, Y, and Z (it is not necessary to explicitly measure in I, as this term can be inferred from any of the other measurements). (b) Quantum process tomography full characterizes the noisy process $\mathcal{E}(\rho)$ by preparing ρ and measuring $\mathcal{E}(\rho)$ in all basis states $P \in \mathbb{P}_n$. (c) Gate set tomography characterizes $\mathcal{E}(\rho)$ by preparing and measuring in all Pauli basis states $P \in \mathbb{P}_n$, but repeats $\mathcal{E}(\rho)$ L times for better accuracy. (d) Gate set tomography completely characterizes all gates in a gate set $\mathbb{G}$, as well as SPAM. $F_{X,Y,Z}$ denotes the SPAM layers (called "fiducials"), and $G_i \in \mathbb{G}$ denotes the gate that is being amplified.

$\mathcal{L}$ of the GST model with the likelihood $\mathcal{L}_{\max}$ of the "maximal model",

$$\lambda = -2 \ln \left(\frac{\mathcal{L}}{\mathcal{L}_{\max}} \right). \tag{5.65}$$

The maximal model is the one in which each independent measurement outcome in the data set is assigned a distinct probability equal to the observed frequencies. Wilks' theorem [156] states that if $\mathcal{L}$ and $\mathcal{L}_{\max}$ are both valid models, then the log-likelihood ratio is a χ^2_k random variable, where $k = N_{\max} - N_p$ is the difference in the number of free parameters between the maximal and GST model. If λ is not consistent with χ^2_k distribution (i.e. it does not lie within the interval $[k - \sqrt{2k}, k + \sqrt{2k}]$, with mean k and standard deviation $\sqrt{k}$), then this indicates that the GST data are inconsistent with the GST model. One can further quantify the model violation, or "goodness of fit," by the number of standard deviations that λ is

Protocol 3 Gate Set Tomography

For each $F_i, F_k \in \mathbb{F}$ and $G_j \in \mathbb{G}$, do:

1. Initialize the qubit in the state $|\rho\rangle\rangle$, typically taken to be the ground state of the qubit, $\rho = |0\rangle\langle 0|$.
2. Apply the circuit $\mathcal{C} = F_i \circ G_j \circ F_k$. Since all $F_i \in \mathbb{F}$ are composed of native gates in $\mathbb{G}$, the entire circuit $\mathcal{C}$ only contains native operations defined in the gate set $\mathbb{G}$.
3. Perform the POVM M. Both M (e.g. $|0\rangle\langle 0|$) and $\mathbb{I} - M$ (e.g. $|1\rangle\langle 1|$) must be positive semidefinite Hermitian operators.
4. Repeat steps (1) - (3) a large number of times (e.g. $n = 1,000$ – $10,000$ "shots"). For each repetition r, record the success probability p_r, where $p_r = 1$ or 0 if the measurement was a success (e.g. resulted in $|0\rangle\langle 0|$) or failure (e.g. resulted in $|1\rangle\langle 1|$), respectively.
5. Compute the average result of step (3): $\langle p_{ijk} \rangle = \frac{1}{n} \sum_{r=1}^{n} n_r$, which represents an estimation of the expectation value $p_{ijk} = \langle\langle M | F_i G_j F_k | \rho \rangle\rangle$, which is a random variable with mean p_{ijk} and variance $p_{ijk}(1 - p_{ijk})/n$.
6. Optional: repeat all steps for G_j amplified L times for better accuracy (e.g. $L = 128$).

from the expected mean k,

$$N_\sigma = \frac{\lambda - k}{\sqrt{2k}}. \tag{5.66}$$

If $N_\sigma \leq 1$, then the GST model faithfully captures all of the behavior of the device. However, on actual NISQ hardware, $N_\sigma > 1$ (or even $N_\sigma >> 1$) has been observed [152, 157, 158], indicating the presence of significant model violation. GST makes an assumption of Markovianity[8] (i.e. any Markovian process can – by definition – be captured in a generalized model based on process matrices), therefore large N_σ indicates the presence of non-Markovian errors, but it does not quantify the amount of non-Markovianity in the model or the magnitude of such errors. Because N_σ will grow linearly with the number of shots and depth of the GST circuits (which set the accuracy of the model), a large N_σ simply indicates that the assumption of Markovianity was violated with high confidence, even if the underlying non-Markovian errors are small in magnitude. Therefore, large N_σ does not mean that the GST estimate cannot be trusted; rather, it simply indicates that some non-Markovianity is present, which cannot be fit by a process matrix model.

[8] One can additionally add trace-preserving (TP) and completely-positive (CP) constraints as well.

Model Fitting in GST

GST enables the comparison of nested error models in a self-consistent manner [158, 159]. To compare two nested models, we utilize the *evidence ratio* γ [159],

$$\gamma = \frac{\lambda_S - \lambda_L}{N_{p,L} - N_{p,S}}, \tag{5.67}$$

where λ is the model's log-likelihood ratio, and L (S) denotes the larger (smaller) model defined in terms of the number of free parameters N_p describing the model. If $\gamma \leq 1$, then we automatically choose the smaller model, as it describes the data at least as well as the larger model without extra (potentially unused) parameters. If $1 < \gamma \leq 2$, there is weaker evidence for rejecting the smaller model, but Akaike's information criterion [160] suggests that the smaller model is still preferable. Even if $\gamma > 2$, it may still be preferable to choose the smaller model due to its simplicity, as it may still accurately capture the vast majority of the data.

In Fig. 5.14, we include a schematic of a typical GST sequence for two qubits under simultaneous operation, and show how nested models can be used to fit different kinds of crosstalk errors to the data. The different models of crosstalk are based upon the conditions that must be satisfied for quantum operations to be crosstalk-free (see Ref. [69] and the discussion in Sec. 3.2). However, one can go further and assume that a model is free of coherent errors entirely, in which case data can be fit to a purely stochastic Pauli model. Therefore, one can fit GST data to the following parameterized error models:

1. General completely-positive and trace-preserving (CPTP) model: each gate's errors are modeled by a general CPTP two-qubit PTM $\Lambda_{A,B}^{(1,2)}$, where A (B) denotes the gate acting on qubit 1 (2).

2. General Pauli stochastic (S) model: the errors in the general model are restricted to be diagonal in the Pauli basis $\left(\Lambda_{S:A,B}^{(1,2)}\right)$.

3. Context-dependent (CD) model: local gate errors are described by single-qubit PTMs which can depend on the gate acting on a different qubit $\left(\Lambda_{A;B}^{(1)} \otimes \Lambda_{B;A}^{(2)}\right)$.

4. Stochastic context-dependent (SCD) model: the error generators in the CD model are restricted to be weight-1 stochastic Pauli errors $\left(\Lambda_{S:A;B}^{(1)} \otimes \Lambda_{S:B;A}^{(2)}\right)$.

5. Context-free (CF) model: local gate errors are described by single-qubit PTMs, unconditional on the gate acting on the other qubit $\left(\Lambda_{A}^{(1)} \otimes \Lambda_{B}^{(2)}\right)$.

6. Stochastic context-free (SCF) model: the error generators in the CF model are restricted to be weight-1 stochastic Pauli errors $\left(\Lambda_{S:A}^{(1)} \otimes \Lambda_{S:B}^{(2)}\right)$.

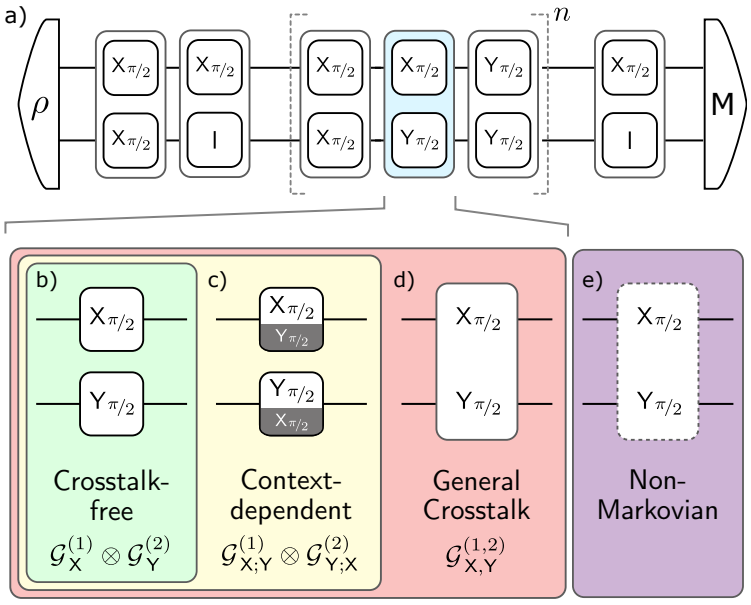


Figure 5.14: **Fitting crosstalk errors with simultaneous GST.** (a) sGST circuits can be broken down into an initialization layer ρ (taken to be the ground state), a set of gates to perform state-preparation, a block of operations repeated n times, a set of measurement basis gates, and finally measurement M in the computational basis. The process matrix for any combination of parallel operations (e.g. $X_{\pi/2} \otimes Y_{\pi/2}$ acting on qubits 1 and 2) can be fit to three different models of crosstalk: (b) the crosstalk free model assumes that each gate can be described by a single-qubit process matrix. (c) The context-dependent model also assumes that each gate can be described by a single-qubit process matrix, but that the errors on each qubit can be conditioned on the gate that is being applied to the neighboring qubit. (d) The general crosstalk model makes no assumption of locality and fits the data to a full two-qubit process matrix. (c) Non-Markovian errors cannot be fit by a two-qubit process matrix. (Figure reproduced with permission from Ref. [158].)

The hierarchy of all of the nested GST models can be seen in Fig. 5.15.

GST can be used to compare nested families of models in a self-consistent manner. For example, the set of circuits needed to evaluate the CPTP model contains all of the information needed to evaluated the S (CD) model; similarly, the set of circuits required to evaluate the S (CD) model contains all of the information needed to evaluate the SCD (CF) model; etc. In general, models with increasing complexity are able to capture more complex dynamics; however, they also require fitting a larger number of free parameters. The general CPTP model makes no assumption of locality and can completely capture all two-qubit interactions, including quantum crosstalk errors. The CD model assumes that any errors acting on qubits must be local (weight-1), since the PTMs decompose as a tensor product of operations, but allows the errors to be classically correlated (i.e. the error impacting qubit 1

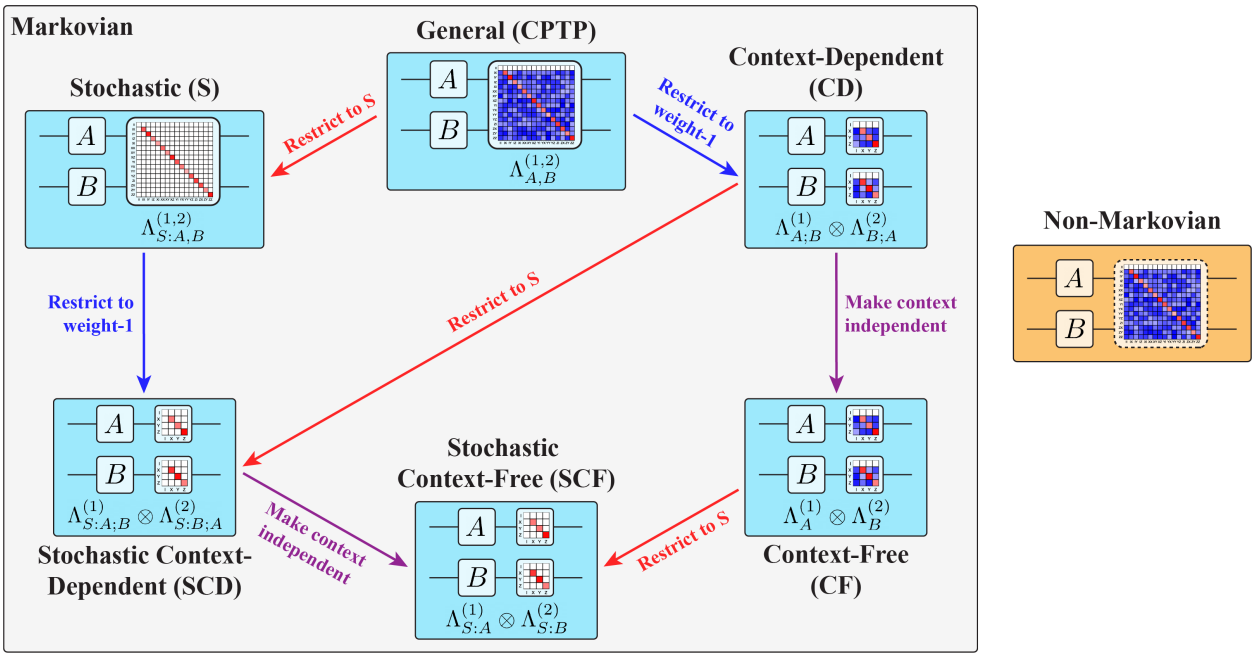


Figure 5.15: **Hierarchy of nested GST models.** We model all physical gates as an ideal gate followed by an error generator in the PTM representation [colored (white) PTM cells denote parameters that are (not) included in the model]. For simultaneous single-qubit gates, the general CPTP model contains all weight-1 and weight-2 error generators. The S model restricts the errors to be stochastic (i.e. the error generator is diagonal in the Pauli basis) [red arrows]. The CD (SCD) model restricts the error generators to be weight-1 (stochastic) errors [blue arrows], but allows contextual dependence. The CF (SCF) model restricts the error generators to be context-independent weight-1 (stochastic) errors [purple arrows]. If one of these models fits the data, the errors are dominantly Markovian [large grey region]. If none of these models fit the data (i.e. if the errors cannot be fit by a full two-qubit process matrix), the errors are taken to be non-Markovian [orange region with dashed border around the PTM]. (These models can also be generalized to two-qubit gates with the inclusion of both weight-1 and weight-2 error generators in the CD and SCD models.)

can be correlated with the gate being applied to qubit 2, but it cannot depend on the state of qubit 2); this can model errors due to classical crosstalk, but not entangling quantum crosstalk. In the CF model, the single-qubit GST data is marginalized to weight-1 errors regardless of the gate acting on the other qubit. The corresponding S-type models make the same assumptions, but restrict the errors to be diagonal in the Pauli basis. All models make an assumption of Markovianity.

Quantifying Non-Markovian Errors in GST

While N_σ is useful for providing evidence of the existence of non-Markovian errors, to quantify the magnitude of such errors, GST utilizes a wildcard error model [161]. The wildcard error rate $w_G \in [0, 1]$ quantifies the unmodeled error per logic gate operation; in other words, it quantifies how much the observed data deviates from the predictions made by the model. A wildcard error can also be assigned to a circuit $\mathcal{C}$ containing gates by summing over the wildcard error rates for all gates $G \in \mathcal{C}$: $w_\mathcal{C} = \sum_{G\in\mathcal{C}} w_G$. The wildcard model is chosen to be minimal, such that assigning w_G to a gate G is just sufficient to make the model's predictions consistent with the observed data. This is enforced by requiring that the total variation distance (TVD)

$$d_{\mathrm{TV}}(p, q) = \frac{1}{2}||p - q||_1 \tag{5.68}$$

between the observed probability distribution p_C and the wildcard-augmented probability distribution q_C be bounded by the total wildcard error for circuit $\mathcal{C}$,

$$d_{\mathrm{TV}}(p_\mathcal{C}, q_\mathcal{C}) \leq w_\mathcal{C}. \tag{5.69}$$

The wildcard-augmented model is therefore not unique, as $q_\mathcal{C}$ can be chosen from any distribution that satisfies Eq. 5.69. Because the wildcard error quantifies the magnitude of unmodeled error per gate, and because unmodeled errors are often attributed to non-Markovianity in the system, it stands to reason that the per-gate wildcard error budget is a good estimate of the magnitude of non-Markovian errors impacting the gate.

The TVD is a useful metric for quantifying the magnitude of unmodeled error because it captures the rate at which measurement outcomes are incorrectly predicted by a model. Because the TVD is upper-bounded by the diamond norm [145], one can compare the wildcard error w_G to the diamond error $\epsilon_\diamond$ for any gate G to inform whether unmodeled errors in the GST estimate are dominant or negligible, and thus whether the GST model can be trusted. By extension, comparing w_G to $\epsilon_\diamond$ quantifies whether Markovian or non-Markovian errors dominate the total error. If $w_G \ll \epsilon_\diamond$, then Markovian errors dominate and non-Markovian errors are negligible; in this case, the model captures the majority of the errors in the gate, despite the fact that it is rigorously incomplete. On the other hand, if $w_G > \epsilon_\diamond$, the non-Markovian errors dominate and Markovian errors are negligible; in this case, the GST estimate is unreliable and should be discarded.

5.6 Overview of other contemporary benchmarking methods

In this chapter, we have outlined the basics of randomized benchmarks and tomographic methods of reconstructing quantum processes. These can broadly be distinguished by differentiating classes of QCVV methods which are scalable but provide limited information about a quantum processes (e.g. CB), versus methods which are not scalable but provide complete

information about a quantum process (e.g. GST[9]). Another class of QCVV methods which was not discussed here is holistic benchmarks for characterizing entire quantum processors. Instead of characterizing individual gates or processes, these "volumetric" benchmarks attempt to measure the capabilities of a quantum device for performing generic tasks, without necessarily specifying what that task is. For example, IBM's Quantum Volume (QV) [162, 163] attempts to capture the global performance of a quantum processor with a single number which quantifies the largest square shaped quantum circuit (where circuit depth = circuit width) that a quantum processor can successfully implement with reasonable accuracy. This method captures how well a device can decompose Haar random SU(4) gates; therefore, fully connected devices are expected to have a larger QV than sparsely connected devices. QV is not a scalable method, as it requires the ideal circuit results to be simulated on a classical computer. Instead, scalable methods based on mirror circuit benchmarking (MCB) [164, 165] can similarly measure the capabilities of entire quantum processors [142], but do not restrict the circuits to be square in shape. MCB can be used to estimate how well random circuits or structured circuits will perform on average for any given circuit depth and circuit width. Finally, Google's cross-entropy benchmark (XEB) [166] was developed to compute the fidelity of random circuit sampling experiments performed on entire quantum processors. XEB is – by design – also not scalable, as it requires the ideal results to be computed classically, specifically in order to demonstrate quantum supremacy by showing that a classical supercomputer could not sample the output distributions of random quantum circuits in any reasonable amount of time [7].

While many of the above benchmarking methods involve generating and decomposing random circuits on different hardware platforms, it has been shown that these are poor predictors of the global perform of structured quantum circuits [142]. Prototypical quantum algorithms have much more structure than random quantum circuits, therefore, it is unclear how well benchmarked error rates predict the performance of these circuits in any given instance. Instead, *application* benchmarks have been developed [167] for the purpose of directly measuring the performance of algorithms such as GHZ state preparation, VQE, QAOA, and Hamiltonian simulation. These have the advantage of providing a direct, unbiased cross-platform comparison of the performance of different quantum algorithms.

5.7 QCVV for Improved System Design

Beyond informing about error rates, what types of errors impact gates the most, etc., QCVV protocols can play an important role in improving the performance of current gates on contemporary quantum processors, as well as guiding improved designs for future quantum processors. For existing quantum processors, error rates measured via RB or CB represent the "cost" of implementing quantum gates. If the gates are coherence limited, then this cost cannot be reduced without reducing their operational time. However, if there are

[9]Note that GST can be made more scalable by restricting the error model of quantum gates, as depicted in Fig. 5.15, and therefore reducing the number of parameters that need to be fit.

calibration errors or unwanted crosstalk, then the cost can be minimized further by improved calibration design (e.g. more fine-tuned pulse parameters, including compensation pulses to null crosstalk, etc.). In this case, RB and CB error rates can be used to improve the calibration of gate parameters using closed-loop optimization [168], with the infidelity acting as the cost that must be minimized by an optimizer.

Using protocols such as CER or GST, one can learn more precise information about the dominant errors effecting entire registers of qubits. This information can be used to address the most pernicious errors in a closed-loop fashion (see Fig. 5.10). For example, as discussed above, unconditional and conditional phase errors on spectator qubits during a two-qubit gate can be mitigated by virtual Z gates or slow Rabi oscillations, respectively. Without such detailed benchmarking tools, knowledge about which error channels dominate cycles of gates would remain obscured.

When it comes to improving the design of future quantum processors, all of the aforementioned QCVV protocols can aid in the design of such processors. For example, being able to benchmark the coherence limit of quantum gates via purity benchmarking is a measure of how well qubits are isolated from the surrounding environment, providing quantitative information about the fundamental noise floor of the device. Furthermore, measuring leakage rates in qubits is important for improved qubit design, since the anharmonicity of the qubits can be fabricated to suppress these transitions. Finally, detailed knowledge of what types of correlated errors impact entire quantum processors can be used to improve the engineering design of future quantum processors. For example, measurement of the always in ZZ coupling term can be used to design improved qubits which suppress this term [169, 170], or help improved the design of the qubit-to-qubit coupling element [171, 172, 173]. The suppression of unwanted correlated errors will be important for fault-tolerant quantum error correction.

Chapter 6

Randomized Compiling

The accuracy of quantum algorithms is limited by different types of errors. Interactions between qubits and the surrounding environment result in incoherent (i.e. non-unitary or irreversible) errors, leading to purity-decreasing processes such as the decoherence of a quantum state. In contrast, systematic imperfections in qubit control (e.g. detuning and calibration errors) and crosstalk on multi-qubit processors result in coherent (i.e. unitary or reversible) errors, which are purity-preserving and thus do not result in decoherence. The existence of coherent errors leads to a disconnect between benchmarked error rates and algorithm performance, namely, while many benchmarking methods use twirling to simplify the structure of a quantum process $\mathcal{E}$ in order to measure error rates, quantum algorithms are not implemented with twirling, and therefore $\mathcal{E}$ includes complex error dynamics due to coherent errors and crosstalk. This disconnect typically results in an observed failure rate of quantum algorithms which exceeds the failure rates predicted from benchmarked error rates [142]. Therefore, the global impact of coherent errors is hard to predict for structured circuits, due to both their quadratically-worse impact on gate infidelities relative to average error rates, and the potential for interference over the course of an algorithm.

In recent years, there has been growing theoretical interest in randomization methods to mitigate the problem of coherent errors in quantum computations [174, 175, 176, 177, 178, 179, 180, 181]. Experimentally, it has been shown that methods such as Pauli-frame randomization [175, 176] (PFR) and Pauli twirling can reduce coherent errors in Clifford circuits [157] and the two-qubit CPHASE gate [182], respectively, as measured by GST. Randomized compiling (RC) [145, 174] is a related protocol for twirling errors in quantum algorithms *in situ* that is more scalable and generalizable than PFR and simple Pauli twirling, and does not require *a priori* knowledge of the specific error model. RC uses Pauli twirling in the same manner as cycle benchmarking (CB); therefore, the effective noise of any cycle under CB is equal to the tailored noise under RC, enabling accurate predictions of algorithmic performance under RC via CB process infidelities. In this chapter, we introduce the RC protocol and demonstrate how it can be used to improve the performance of structured and random circuits. We further discuss the importance of RC as it relates to fault-tolerant quantum error correction (QEC).

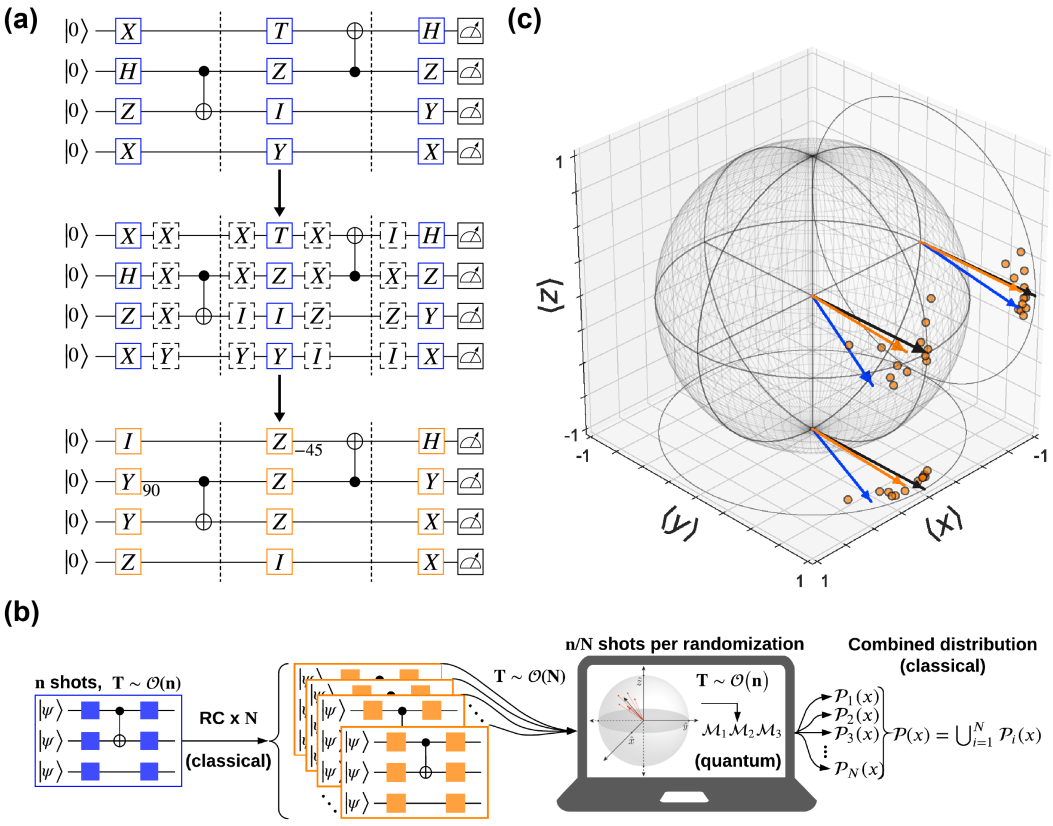


Figure 6.1: **Randomized compiling protocol.** (a) Randomization protocol. The bare circuit, (top) split into K cycles of easy/hard gates (separated by dashed lines), (middle) is converted into a logically-equivalent circuit by inserting random single-qubit twirling gates between each easy and hard cycle, inverting them in the following cycle, and (bottom) then compiling the twirling gates into new easy gate cycles. (b) Measurements using RC. Use the randomization protocol to generate N randomizations of a bare circuit. Measure each randomization n/N times, then combine all the results to obtain an equivalent statistical distribution for a circuit measured n times. The time it takes to convert N randomizations into experimental pulse sequences and upload them to the quantum computer scales linearly in N, but the actual measurement time is unchanged since the total number of shots n remains the same as the bare circuit. (c) Single-qubit state-tomography for visualizing RC: the black vector is the ideal (noiseless) final state of the qubit, but coherent errors cause an over-rotation in the measured state (blue vector). The orange vector represents the final state of the combined distribution of $N = 12$ randomizations (orange data points). The rotation error in the bare result has been tailored into stochastic noise under RC, as the RC vector has a lower purity than the bare vector.

6.1 Randomized Compiling Protocol

RC tailors Markovian errors into stochastic Pauli noise by combining the results of many logically-equivalent circuits. By inserting and compiling random single-qubit (virtual) twirling gates into a circuit in a way that preserves the overall unitary operation, RC creates a set of "randomized" circuits that are logically equivalent to the original "bare" circuit, without increasing circuit depth. In Protocol 4, we outline the basic procedure for performing RC. Here, we consider a bare circuit composed of K cycles of interleaved single-qubit "easy" gates and two-qubit "hard" gates, like the one shown in Fig. 6.1a. Step 1 outlines how the single-qubit gates in a quantum circuit can be randomized by insertion of random twirling gates from some set $\mathcal{T}$. In step 1a, it is necessary to consider how the tensor product of inverting gates is commuted through two-qubit gates in hard-gate cycles. Typically, $\mathcal{T}$ is chosen to be the set of tensor products of single-qubit Paulis, with the edge terms T_0^c and T_K set to the identity gate, so that if the two-qubit gates are all Clifford gates (or locally-equivalent to Cliffords), then the correction gates will also lie in $\mathcal{T}$, and we need place no restriction on the types of allowed easy gates. Therefore, RC is efficiently compatible with universal quantum computation. However, if the two-qubit gates in hard-gate cycles are not locally-equivalent to Clifford gates, then refocusing pulses will be needed, which can result in additional arbitrary two-qubit correction gates at the end of the circuit. No such restrictions are placed on single-qubit gates in hard-gate cycles, which can be arbitrary. In step 1b, the new randomized circuit is logically equivalent to the original bare circuit and has the same number of elementary gates.

Because the correction gates are computed locally for each cycle, the classical resource requirements for each randomization scale linearly in the number of qubits and circuit depth. Therefore, generating many (N) logically-equivalent randomizations of a bare circuit requires very low classical overhead and can be efficiently done before runtime. While certain hardware platforms (e.g. superconducting circuits) may be better equipped to measure large N than others (e.g. trapped ions) due to faster gate times, quantum hardware capable of modifying pulse phases on the fly will enable the utilization of a new randomization per shot [157] on platforms whose single-qubit gates only differ by a change in virtual phases [55]. By measuring each randomization n/N times and computing the union of all N results, we obtain an equivalent statistical distribution for a circuit measured n times in which coherent errors in each computational cycle (except the last) have been averaged into Pauli channels (Eq. 3.36).

To visualize the RC protocol, we performed state tomography on a single qubit (Q7) after 50 random single-qubit gates, as shown in Fig. 6.1c. We find that coherent errors cause a net over-rotation in the measured state compared to the ideal (noiseless) final state. When RC is applied, each randomization results in a different net coherent error; however, the combined result is more aligned with the ideal vector. The state fidelity $\mathcal{F} = 0.862$ and purity $\gamma = 0.938$ for the bare result, and $\mathcal{F} = 0.879$ and $\gamma = 0.881$ for the RC result. The fidelities are comparable, since fidelity is unchanged by twirling (see Sec. 5.1), but the rotation error in the bare result has been tailored into stochastic noise under RC, as the

Protocol 4 Randomized Compiling

1. Randomization protocol for a bare circuit:
   a) Conjugate each round of easy gates C_k by a twirling gate T_k randomly sampled from a set $\mathcal{T}$ and an inverting operator T^c_{k-1}: $C_k \rightarrow T_k C_k T^c_{k-1}$, where T^c_{k-1} is chosen to undo the twirling gate that was inserted in the previous cycle when commuted through the hard gate cycle G_k: $T^c_k = G_k T^\dagger_k G^\dagger_k$.
   b) Compile the original single-qubit gates and twirling gates into new easy gate cycles: $C'_k = T_k C_k T^c_{k-1}$.
2. Use the randomization protocol to generate N logically-equivalent randomizations of the bare circuit.
3. For each randomization, convert the abstract gates into experimental pulse sequences and measure n/N times.
4. Compute the union of all N results, such that the total number of shots over N randomizations is n.

fidelity and purity of the RC result are approximately equal in magnitude.

6.2 Noise Tailoring via Randomized Compiling

RC can benefit circuit performance in two different regimes:

1. Single-randomization limit: a single randomization under RC interrupts the coherent accumulation of unitary errors between cycles of gates, similar to dynamical decoupling sequences [183].
2. Many-randomization limit: averaging over many randomizations under RC tailors errors into stochastic Pauli channels, completely eliminating off-diagonal terms in the error process (in the limit of perfectly-implemented Pauli twirling).

To understand (1), consider the simple example of applying many $R_x(2\pi)$ rotations to a qubit in the ground state, but each time the qubit over-rotates by a small angle θ. The resulting state of the qubit after M rotations is

$$|\psi\rangle = \prod^{M} e^{-i\theta\sigma_x} |0\rangle = \cos(M\theta)|0\rangle - i\sin(M\theta)|1\rangle . \tag{6.1}$$

The fidelity of this state with respect to $|0\rangle$ is $F = |\langle 0|\psi\rangle|^2 = \cos^2(M\theta) \approx 1 - (M\theta)^2$, thus the infidelity $r = 1 - F \approx (M\theta)^2$. Therefore, the infidelity scales quadratically in both the over-rotation angle θ and the number of rotations M. Under RC, the trajectory from the initial state to the final state is randomized, thus ensuring that coherent errors will not grow quadratically between gates. While exact quadratic growth is highly unlikely for longer-depth multi-qubit circuits due to the complex dynamics of crosstalk, and because coherent errors can act in any direction (not just along the axis of rotation), coherent errors can still accumulate in an adversarial fashion and grow faster than average error rates, especially in structured quantum circuits [142].

Regardless of the rate at which coherent errors accumulate between cycles of gates, they can still impact each computational gate G in a circuit. We can model this process as an ideal gate G_0 followed by an unwanted unitary operator $G = U(\hat{\mathbf{n}}, \theta)G_0$, where $U(\hat{\mathbf{n}}, \theta) = e^{-i\theta\hat{\mathbf{n}}\cdot\boldsymbol{\sigma}/2}$, $\hat{\mathbf{n}}$ is the axis of rotation, $\boldsymbol{\sigma}$ the Pauli vector, and θ is the rotation angle relative to the intended target state. For simplicity, consider a unitary error about the x-axis for a single qubit,

$$\begin{aligned} U(x,\theta) &= \exp\left(-i\frac{\theta}{2}\sigma_x\right) \\ &= \begin{pmatrix} \cos(\theta/2) & -i\sin(\theta/2) \\ i\sin(\theta/2) & \cos(\theta/2) \end{pmatrix}. \end{aligned} \tag{6.2}$$

In the PTM representation, this coherent error takes the following form,

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos(\theta) & -\sin(\theta) \\ 0 & 0 & \sin(\theta) & \cos(\theta) \end{pmatrix}. \tag{6.3}$$

For small θ, the diagonal components of Λ scale as $\cos(\theta) \approx 1-\theta^2$, and the off-diagonal terms scale as $\sin(\theta) \approx \theta$. While the infidelity of the diagonal terms is $e_F \approx \theta^2$, we see that the off-diagonal terms are quadratically larger, with $\sqrt{e_F} \approx \theta$. While not all error metrics are sensitive to the off-diagonal terms in an error process (e.g. fidelity-based measures), norm-based error metrics such as the diamond distance (Eq. 5.33) and total variation distance (Eq. 5.27) are generally sensitive to such terms, setting the $\sim \sqrt{e_F}$ upper bound of the diamond norm (Eq. 5.36).

Under RC in the many-randomization limit, all off-diagonal terms in the error process are completely suppressed in the limit that $N \longrightarrow \infty$. This is referred to as the "noise tailoring" property of RC. To understand how this occurs, consider Pauli twirling Λ, i.e. $P\Lambda P^\dagger$ for any $P \in \{I, X, Y, Z\}$, where P represents the Pauli superoperator. Under Pauli conjugation, the signs of the off-diagonal terms remain the same (are reversed) if P commutes (does not commute) with Λ. Thus, the off-diagonal terms change sign with a 50% probability upon

conjugation with a randomly-selected Pauli. When averaging over N randomizations, the magnitude of the off-diagonal terms scale as $\theta/\sqrt{N}$, reminiscent of a random walk, and thus vanish as $N \longrightarrow \infty$ or if by luck the correct Paulis were sampled which average to zero. While the noise tailoring property of RC rests on assumption that the noise impacting the easy gates is gate-independent, Wallman *et. al.* [174] prove that RC is robust to small gate-dependent errors, which are inevitable in modern-day experiments. As we can see, RC can provide both suppression (i.e. reduction in magnitude) and mitigation (i.e. averaging away) of coherent errors.

Single-qubit State Tomography

Single-qubit dynamics provide an intuitive picture for understanding how noise is tailored under RC: consider the ideal final state of a qubit after performing a sequence of gates. The ensemble measurements of this final state will yield a probability distribution of 0s and 1s, which depends on the location of the final state on the Bloch sphere and the measurement basis. In a circuit dominated by coherent errors, the actual final state of the bare circuit will have some angle error relative to the ideal state due to the interference of coherent errors during the sequence of gates. Under RC, the impact of coherent errors is different for each randomization, since each circuit represents a slightly different trajectory to the same (ideal) final state. Thus, the combined distribution of all the randomizations averages out the impact of coherent errors, resulting in a stochastic Pauli channel.

To verify this intuitive picture of noise tailoring via RC, we performed state tomography on all four qubits independently, with and without RC. Random circuits were generated by randomly sampling $K = 100$ interleaved cycles of "easy" and "hard" single-qubit gates, as defined by the following gate sets: the Clifford set, $C_{easy} = \{\mathbf{C_1}\}$, and common non-Clifford gates, $G_{hard} = \{X45,\ Y45,\ T = Z45\}$, respectively. Thus, the total number of single-qubit gates for each qubit was $2K + 1 = 201$. State tomography results were constructed by performing ensemble measurements in the X, Y, and Z bases at five different points during each random circuit, as defined by the circuit depth: $K = 5$, $K = 25$, $K = 50$, $K = 75$, and $K = 100$. These results are plotted in Fig. 6.2; 2d projections of the results are also included for visual clarity. At each circuit depth, measurements were performed on the bare circuit (blue vector) and on 12 different randomizations (orange points). The ideal result is plotted as a black vector, and the combined RC result is represented by the orange vector. 6,000 shots were taken in each measurement basis for the bare circuits, and 500 shots were taken for each randomization. Focusing on Q5 as an example, at the beginning of the circuit ($K = 5$) all three vectors are more or less co-aligned at the south pole. However, at later circuit depths, coherent errors manifest as a separation between the ideal and bare vectors. At all circuit depths, the RC vector is approximately co-linear with the ideal state.

Since each tomography result corresponds to measuring same final state in X, Y, and Z, the TVD can be calculated with respect to each basis. These results are summarized in Table 6.1. When the target state is approximately aligned with the measurement basis, as is the case for all four qubits at $K = 5$ with respect to Z, the bare and RC TVD performance

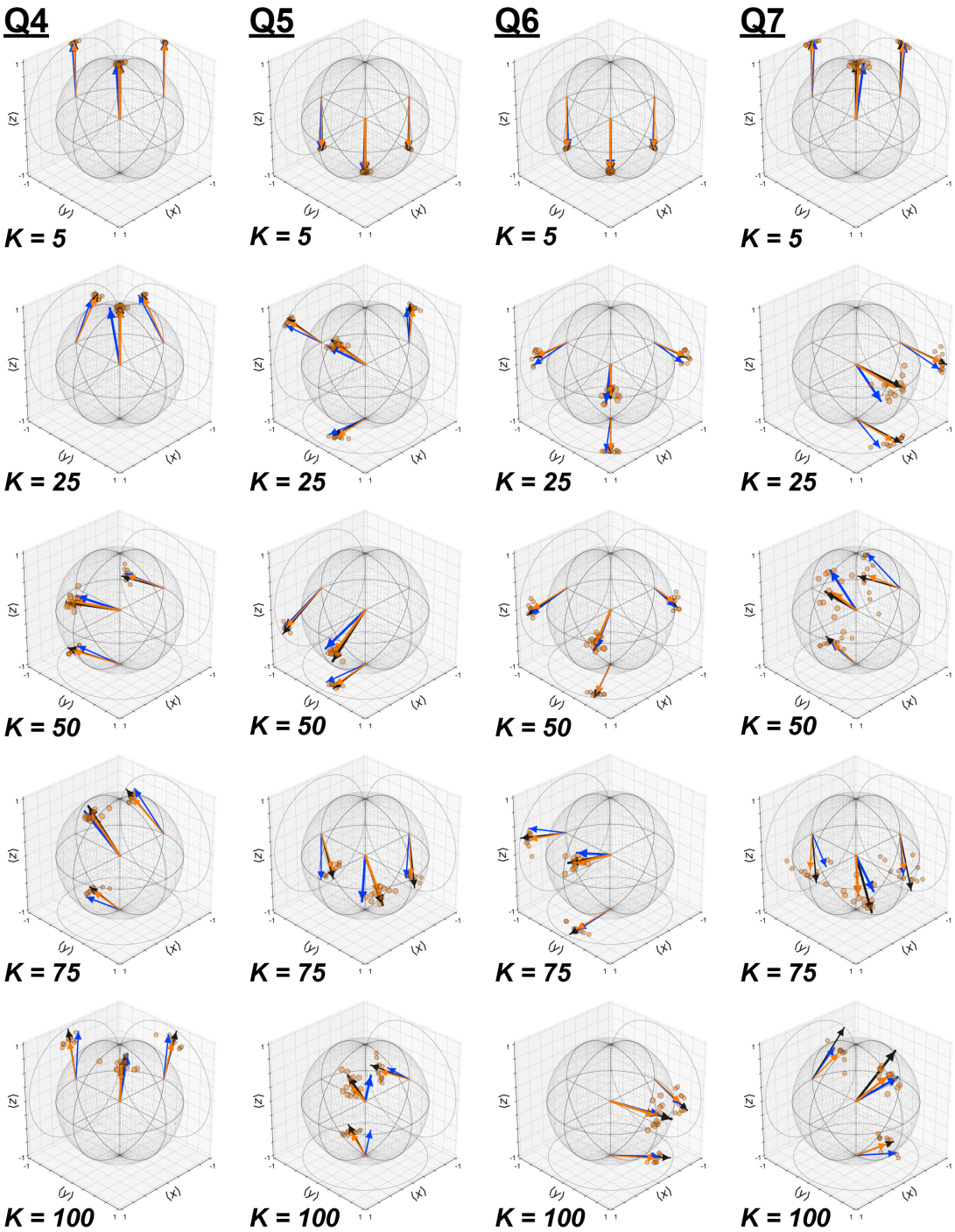


Figure 6.2: **Single-qubit state tomography results.** K is the circuit depth (number of cycles of easy/hard gates). The black vector is the ideal result, the blue vector is the bare result, the orange points are the individual RC results, and the orange vector is the combined RC result.

Qubit (Depth)	TVD_x		TVD_y		TVD_z		Purity		Fidelity	
	Bare	RC	Bare	RC	Bare	RC	Bare	RC	Bare	RC
Q4 ($K = 5$)	0.031(8)	0.006(2)	0.009(8)	0.002(2)	0.009(2)	0.0089(4)	0.985	0.982	0.983	0.982
Q4 ($K = 25$)	0.074(8)	0.014(2)	0.037(7)	0.020(2)	0.029(6)	0.024(2)	0.935	0.932	0.922	0.932
Q4 ($K = 50$)	0.077(8)	0.004(2)	0.037(7)	0.062(1)	0.042(8)	0.035(2)	0.891	0.870	0.879	0.869
Q4 ($K = 75$)	0.130(8)	0.024(2)	0.095(6)	0.072(2)	0.038(7)	0.049(2)	0.887	0.821	0.838	0.820
Q4 ($K = 100$)	0.111(8)	0.010(2)	0.081(8)	0.038(2)	0.060(4)	0.107(2)	0.845	0.773	0.812	0.772
Q5 ($K = 5$)	0.029(8)	0.008(2)	0.011(8)	0.006(2)	0.020(2)	0.021(1)	0.963	0.958	0.961	0.957
Q5 ($K = 25$)	0.014(6)	0.013(2)	0.034(8)	0.033(2)	0.059(7)	0.030(2)	0.945	0.941	0.937	0.939
Q5 ($K = 50$))	0.025(5)	0.042(2)	0.104(8)	0.023(2)	0.060(8)	0.046(2)	0.892	0.876	0.868	0.875
Q5 ($K = 75$)	0.138(8)	0.003(2)	0.097(8)	0.035(2)	0.054(4)	0.090(1)	0.852	0.818	0.800	0.815
Q5 ($K = 100$)	0.067(6)	0.073(2)	0.195(7)	0.106(2)	0.050(8)	0.015(2)	0.782	0.748	0.716	0.746
Q6 ($K = 5$)	0.009(8)	0.009(2)	0.010(8)	0.003(2)	0.023(2)	0.023(1)	0.955	0.955	0.955	0.955
Q6 ($K = 25$)	0.010(6)	0.024(2)	0.041(6)	0.020(2)	0.062(8)	0.009(2)	0.967	0.939	0.957	0.939
Q6 ($K = 50$)	0.038(5)	0.046(2)	0.020(7)	0.027(2)	0.009(8)	0.029(2)	0.918	0.887	0.918	0.886
Q6 ($K = 75$)	0.088(5)	0.087(1)	0.003(8)	0.015(2)	0.052(8)	0.045(2)	0.874	0.807	0.861	0.806
Q6 ($K = 100$)	0.045(7)	0.091(2)	0.077(7)	0.077(2)	0.020(8)	0.039(2)	0.821	0.752	0.820	0.751
Q7 ($K = 5$)	0.037(8)	0.006(2)	0.044(8)	0.018(2)	0.012(2)	0.017(1)	0.983	0.968	0.977	0.967
Q7 ($K = 25$)	0.170(8)	0.029(2)	0.069(4)	0.060(1)	0.073(8)	0.008(2)	0.938	0.881	0.862	0.879
Q7 ($K = 50$)	0.041(8)	0.031(2)	0.105(5)	0.114(2)	0.251(8)	0.052(2)	0.844	0.754	0.705	0.751
Q7 ($K = 75$)	0.113(8)	0.068(2)	0.090(8)	0.084(2)	0.107(5)	0.149(2)	0.808	0.672	0.762	0.658
Q7 ($K = 100$)	0.112(7)	0.065(2)	0.136(7)	0.017(2)	0.139(8)	0.210(2)	0.761	0.657	0.688	0.619

Table 6.1: **Total variation distance, purity, and fidelity of the single-qubit state tomography results.** The TVD is calculated with respect to each measurement basis for all qubits at each circuit depth. The state fidelity and purity are calculated for the bare results and unioned RC results.

is approximately equal. This is a reflection of the fact that RC is not expected to provide algorithmic improvement when the target state is an eigenstate of the measurement basis (discussed below).

The consequence of tailoring coherent errors into stochastic Pauli noise is that the tailored noise becomes a decoherence channel, resulting in a reduction of the purity of the Bloch vector. The purity and state fidelity of each result are included in Table 6.1, and these results are summarized in Fig. 6.3, in which both are plotted as a function of circuit depth for all four qubits. As seen in this plot, the purities of the RC vectors decrease more rapidly than their bare counterparts. However, the state fidelities of the RC results are approximately equal in magnitude to their respective purities, underscoring that the most dominant error mechanism under RC is stochastic noise. Under pure depolarizing noise, the RC vector

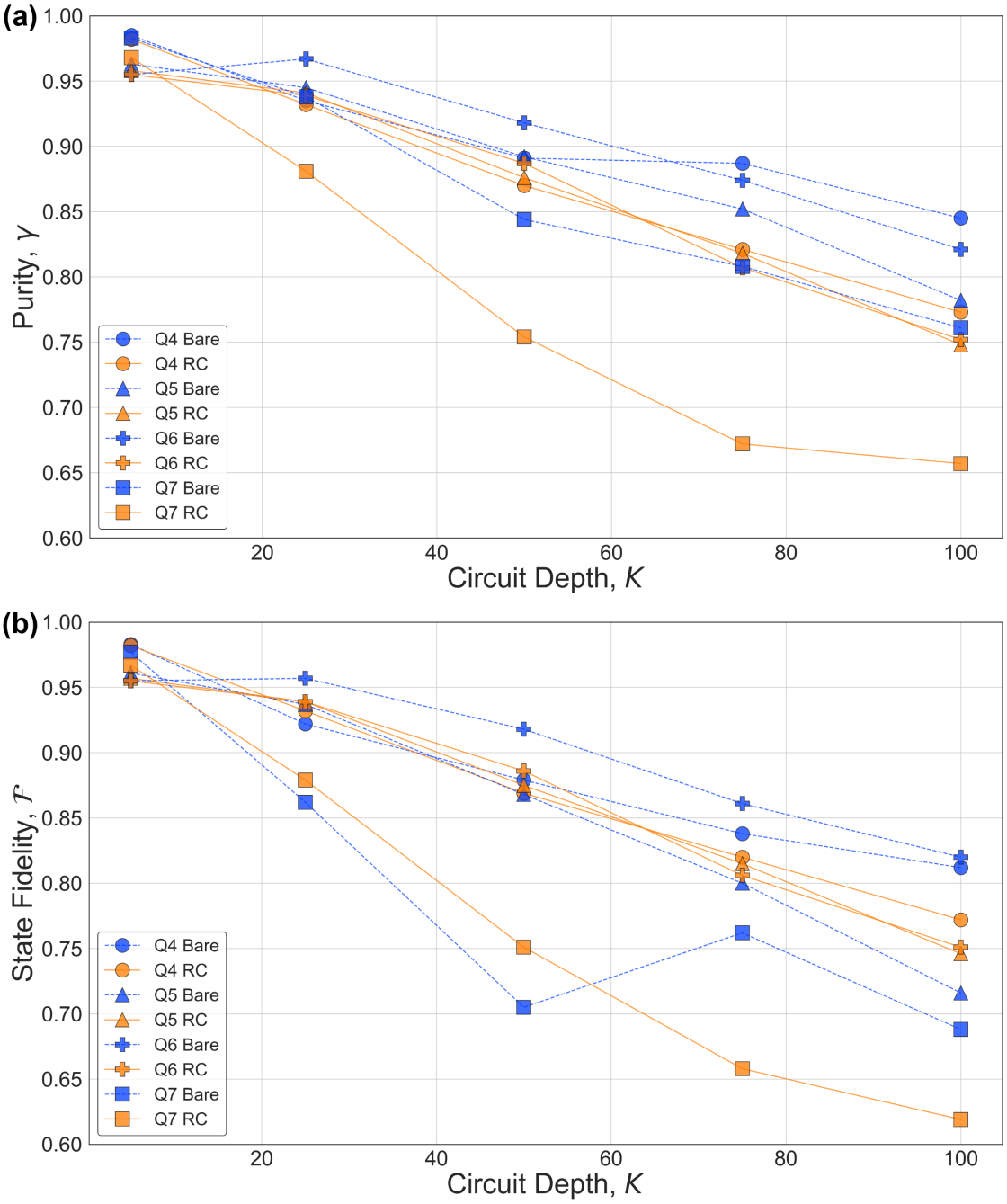


Figure 6.3: **Purity and state fidelity vs. circuit depth for the single-qubit state tomography results.** The purity and fidelity of the bare results can differ significantly due to angle errors in the bare states resulting from coherent errors. In contrast, the purity and fidelity of the RC results are approximately equal in magnitude at each circuit depth, demonstrating that the most dominant error mechanism under RC is stochastic noise.

would decrease linearly with circuit depth; the deviation from perfect linearity in our case shows that the tailored noise under RC is due to Pauli errors with different contributions. While both the purity and state fidelity decrease monotonically for the RC results, the same is not always true for the bare results due to the nature of coherent errors and how they interfere over the course of a circuit. As an example, we see the bare fidelity of Q7 increases from $K = 50$ to $K = 75$, even though the corresponding purity continues to decrease.

Gate Set Tomography

While single-qubit dynamics are useful for visualizing the impact of noise tailoring via RC, we really care more about the impact of RC on quantum gates used in an experiment. To demonstrate the affect of RC on logic gate operations, we perform gate set tomography (GST) with and without RC. Our gate set consists of all possible combinations of I, $X_{\pi/2}$, and $Y_{\pi/2}$ single-qubit gates applied simultaneously to both qubits, as well as the controlled-Z (CZ) gate [60]: $\mathbb{G} = \{G_1 \otimes G_2 : G_1, G_2 \in \{I, X_{\pi/2}, Y_{\pi/2}\}\} \cup \{CZ\}$. We utilized GST up to depth $L = 128$ layers to benchmark the performance of all gates in the gate set. We apply RC to GST circuits using $N = 1$, $N = 10$, and $N = 100$ randomizations to study the transition of RC from the single- to many-randomization limit, and compare the results to GST performed without randomization ($N = 0$).

To study the types of errors present in a gate's process matrix G estimated using GST, we write a noisy quantum gate as

$$G = \Lambda G_0 = e^{\mathcal{L}} G_0, \tag{6.4}$$

where G_0 is the ideal quantum gate, Λ the gate error channel, and $\mathcal{L}$ the gate error generator [78]. $\mathcal{L}$ can be considered roughly equivalent to the Linbladian superoperator that generates all gate errors (coherent, stochastic, and non-unital) in the limit that the noisy gate G estimated by GST is both CP and divisible. However, not all gate errors occurs *after* the gate, and $\mathcal{L}$ is gauge-dependent, therefore this representation has its limitation. It is nonetheless useful to utilize the GST estimates of $\mathcal{L}$ to visualize the error maps under RC.

To visualize the impact of noise tailoring on Markovian gate errors, in Fig. 6.4 we plot the error generator $\mathcal{L}$ in the PTM representation for the CZ gate. We compare the results for $N = 0, 1, 10, 100$ and find that RC efficiently transforms the error generator from dense to sparse as we increase N, twirling $\mathcal{L}$ into a diagonal stochastic Pauli channel. Going from $N = 0$ to $N = 1$ randomizations significantly reduces the magnitude of $\mathcal{L}$'s diagonal and off-diagonal elements, signifying the presence of coherent errors, but $\mathcal{L}$ for $N = 1$ still has significant magnitude in its off-diagonal elements. As N increases from 1 to 10, and from 10 to 100, the magnitude of the off-diagonal terms is greatly reduced.

Because GST enables model fitting, we determine a best-fit model for each dataset ($N = 0, 1, 10, 100$) by computing the evidence ratio γ (Eq. 5.67) and N_σ (Eq. 5.66) for each pair of nested models in Fig. 5.15. Shown in Fig. 6.4e, we plot the model violation and list the γ for each restricted model. To be as rigorous as possible, we set $\gamma = 1$ as our threshold for

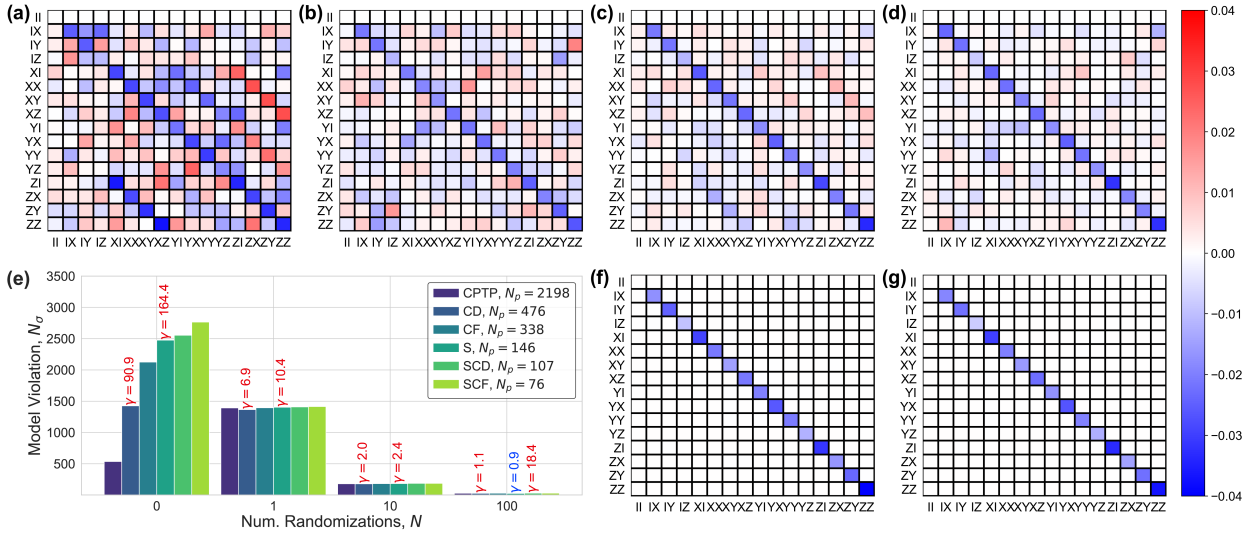


Figure 6.4: **Improving model accuracy via noise tailoring.** The error generator $\mathcal{L}$ in the PTM representation for the CPTP model of the CZ gate is plotted for (a) $N = 0$, (b) $N = 1$, (c) $N = 10$, and (d) $N = 100$ randomizations under RC. If an error channel (i.e. PTM cell) in $\mathcal{L}$ is zero, then this component of the estimated gate matches the corresponding component of the ideal target gate. (e) The model violation N_σ for each GST model, plotted as a function of N. The evidence ratio γ (number of free parameters, N_p) is labeled above each nested model (listed in the legend). For each N, we choose the model with the least number of parameters that satisfies $\gamma \leq 1$, finding that CPTP is the best fit for $N = 0$, $N = 1$, and $N = 10$, and S for $N = 100$; blue (red) text indicates that the model is accepted (rejected). The CZ gate error generator $\mathcal{L}$ is plotted for the S model for (f) $N = 10$ and (g) $N = 100$, showing that the stochastic models capture the dominant error channels for both. Even though it is ultimately rejected by the evidence ratio for $N = 10$, the S model could be reasonably selected for its simplicity. (Figure reprinted with permission from Ref. [184].)

choosing a smaller over a larger model. We find that the CPTP model fits best for $N = 0$, $N = 1$, and $N = 10$, and that the stochastic (S) model fits as well as the CPTP model for $N = 100$. For $N = 10$, none of the nested models satisfy the evidence ratio criteria, but the S or context-dependent (CD) model could also be reasonably selected, as $\gamma = 2.0$ and 2.4, respectively. For $N = 100$, only the S model satisfies the evidence ratio criteria, and we therefore prefer this model over the CPTP model. Strictly speaking, RC only completely tailors the noise in the limit that $N \longrightarrow \infty$; however, because the S model is the best fit model for $N = 100$, and because the S model restricts the error generators to be diagonal in the Pauli basis, any small residual off-diagonal terms do not show up in the model. For the RC GST data presented in the rest of this work, we fix the model for each dataset using the best-fit models outlined above.

The large model violation for $N = 0$ ($N_\sigma \approx 535$) and $N = 1$ ($N_\sigma \approx 1394$) provides

high confidence that there are non-Markovian errors, which cannot be modeled by a PTM. $N = 10$ ($N_\sigma \approx 180$) and $N = 100$ ($N_\sigma \approx 25$) have significantly less model violation, indicating less statistical evidence of non-Markovianity. Finally, we note that while the $N = 1$ CPTP model has much larger model violation than $N = 0$, this is entirely consistent with reconstructing PTMs from pulse-amplified sequences with randomized gates under RC, even if the underlying gates themselves are Markovian (see below for further discussion).

Method for Performing RC on GST sequences

In order to preserve the circuit depth of GST sequences under RC, randomly sampled single-qubit Paulis and their correction gates are inserted between every layer. For circuits only containing single-qubit gates, the random Paulis are compiled into the previous layer and the correction gates are compiled into the subsequent layer. For circuits containing two-qubit gates, the correction gates are commuted through the two-qubit gate before being compiled into the subsequent layer.

To highlight this method, consider a circuit $\mathcal{C}$ containing N layers L of single-qubit gates:

$$\mathcal{C} = L_N L_{N-1} ... L_3 L_2 L_1. \tag{6.5}$$

Under RC, a single randomized circuit takes the following form

$$\mathcal{C} = P_N L_N P_{N-1}^\dagger P_{N-1} L_{N-1} ... L_3 P_2^\dagger P_2 L_2 P_1^\dagger P_1 L_1, \tag{6.6}$$

where $P_N^\dagger$ is omitted in the circuit but taken into account in the final ideal measurement results. The compiled circuit is

$$\mathcal{C} = \tilde{L}_N \tilde{L}_{N-1} ... \tilde{L}_3 \tilde{L}_2 \tilde{L}_1, \tag{6.7}$$

where the kth layer $\tilde{L}_k = P_k L_k P_{k-1}^\dagger$ (except for the first layer, which does not contain a correction gate). For circuits containing two-qubit gates in layer $k-1$, the kth layer becomes $\tilde{L}_k = P_k L_k P_{k-2}^c$, where $P_{k-2}^c = L_{k-1} P_{k-2}^\dagger L_{k-1}^\dagger$. This method therefore randomizes all layers of single-qubit gates, while also maintaining the original circuit depth, and was developed within the `pyGSTi` framework specifically for the purpose of randomizing GST and related benchmarking circuits.

As highlighted above, we observe larger model violation for a single-randomization under RC than for no randomizations. This is expected behavior for $N = 1$ even when the physical gates are Markovian, which we illustrate with the following example: consider two single-qubit circuits, $\mathcal{C}_1 = G_I G_I$ and $\mathcal{C}_2 = G_I G_I G_I G_I$, and consider a physical gate set $\{G_I, G_X, G_Y, G_Z\}$, with $G_X = X_\pi$, etc. Consider a simple error model where each gate is followed by a small coherent X_θ rotation error. When we implement $\mathcal{C}_1$ with a single randomization under RC, we actually implement a circuit that should perform the identity rotation, but the gates have been randomized according to the method outlined above. Let's assume our randomized circuit $\tilde{\mathcal{C}}_1$ is the same as the original circuit, $\tilde{\mathcal{C}}_1 = \mathcal{C}_1 = G_I G_I$. If we initialize

our qubit in the ground state, we will observe a small coherent over-rotation error by 2θ, which will result in some θ-dependent probability of not measuring 0. Similarly, when we run $\mathcal{C}_2$ we will actually actually implement one of the many length-4 combinations of Pauli gates that produces the identity rotation. For example, suppose we sampled $\tilde{\mathcal{C}}_2 = G_Z G_Z G_Z G_Z$; this combination of gates would echo away the X rotation error, and we would measure 0 with probability 1. Therefore, altogether we observe $\Pr(0|\mathcal{C}_1) < 1$ and $\Pr(0|\mathcal{C}_2) = 1$. This is inconsistent with every possible process matrix for G_I (or at least every process matrix that is close to the target identity matrix), because repeating an identity gate amplifies all of its error parameters, but does not echo away errors. Finally, note that this argument is predicated on the assumption that we measure each circuit many times; if we measure each circuit only once, the results will not be inconsistent, but will also not be very informative.

6.3 Quantum Fourier Transform

RC can be applied to any gate-based quantum algorithm, including those at the heart of many quantum applications, like the quantum Fourier transform (QFT). Here, we utilized a state-of-the-art synthesis algorithm [185] that numerically approximates circuit unitaries in order to reduce the CNOT count for a given hardware topology, which generated two, three, and four-qubit QFT circuits consisting of only $K = 3$, $K = 8$, and $K = 13$ CNOTs, respectively, for our linear connectivity. Much like the classical discrete Fourier transform, the QFT maps singular inputs (e.g. $|0000\rangle$) into uniform distributions, and maps superposition states (e.g. $|++++\rangle$) into singular distributions. To measure the performance of RC for different resultant probability distributions, we applied the QFT to various single-qubit product states involving permutations of Pauli basis states $\{|0\rangle, |1\rangle, |+\rangle\}$, as well as random separable input states $(SU(2)^{\otimes n}_{\text{rand}} |0\rangle^{\otimes n})$; see Fig. 6.5a for several examples of the measured distributions for four qubits. Experimental TVD results for the four-qubit QFT can be seen in Fig. 6.5b; two- and three-qubit QFT results can be seen in Fig. 6.6. $N = 50$ randomizations were generated for each bare circuit, and 100 random inputs were generated for each data set. Each bare circuit was measured 10,000 times, and each randomization was measured 200 times.

To evaluate the efficacy of RC, we assess algorithmic performance by the total variation distance (TVD), a standard metric for the statistical distance between two probability distributions and a relevant measure in quantum supremacy experiments [8]:

$$d_{\text{TV}}(\mathcal{P}, \mathcal{P}_{\text{ideal}}) = \frac{1}{2} \sum_{x \in X} |\mathcal{P}(x) - \mathcal{P}_{\text{ideal}}(x)|, \tag{6.8}$$

where $\mathcal{P}_{\text{ideal}}(x)$ is the ideal probability of measuring a bit string x in a set of possible bit strings X, and $\mathcal{P}(x)$ is the experimentally-measured probability. The TVD is a basis-dependent metric which determines the *probability of obtaining an incorrect solution*, with 0 (1) indicating that the correct distribution of bit strings is always (never) measured. Thus, improvements in algorithmic performance equate to lower TVDs, as exemplified by the ob-

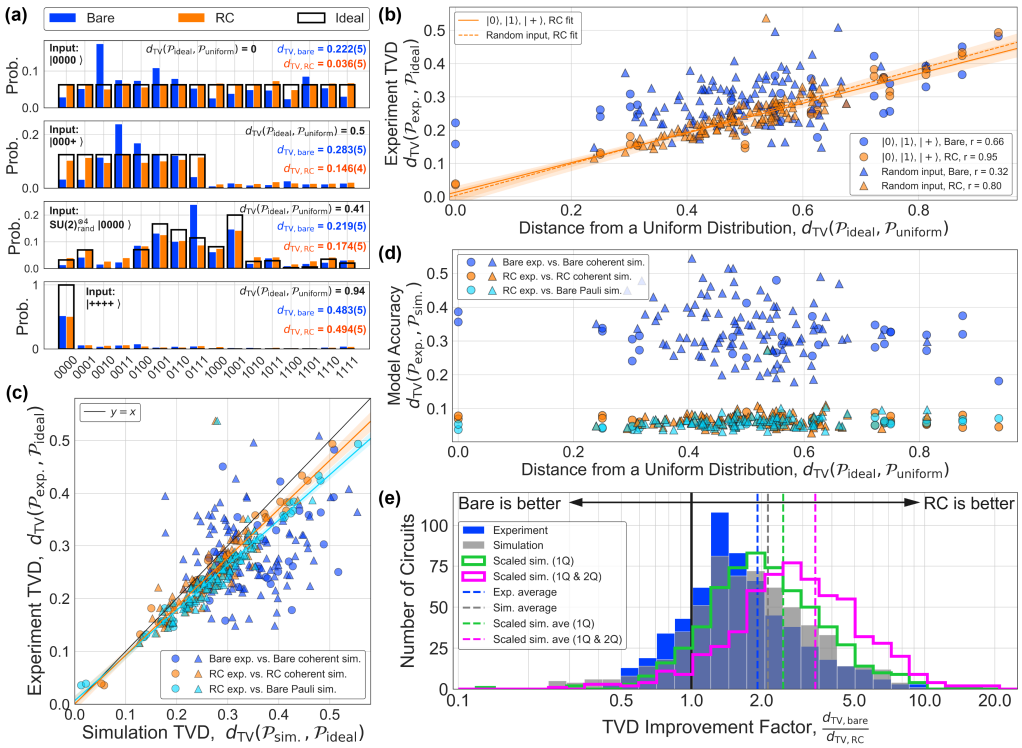


Figure 6.5: **Improving the QFT with RC.** (a) Measured probability distributions for the QFT applied to $|0000\rangle$, $|000+\rangle$, $|++++\rangle$, and a random input state ($SU(2)^{\otimes 4}_{\text{rand}}|0000\rangle$). (b) Bare and RC TVDs for all four-qubit QFT results, as a function of distribution uniformity of the ideal results. RC provides more improvement as $d_{\text{TV}}(\mathcal{P}_{\text{ideal}}, \mathcal{P}_{\text{uniform}}) \longrightarrow 0$. Circles indicate the QFT applied to Pauli basis states ($\{|0\rangle, |1\rangle, |+\rangle\}$), and triangles indicate random inputs states ($SU(2)^{\otimes 4}_{\text{rand}}|0000\rangle$). Pearson r values listed in the legend quantify the correlation strength of each data set, justifying linear fits for the RC data (transparent bands indicate the 95% confidence intervals). (c) Experimental vs. simulated TVDs from two models based on the Pauli error rates in Fig. 5.10b. The blue (orange) markers denote the bare (RC) circuits simulated with the coherent error model, and the cyan markers denote the bare circuits simulated with the Pauli model. (d) Accuracy of the two models compared to experimental results. The bare circuits simulated with the Pauli model are plotted against the experimental RC results in (c), which are also used to compute the model accuracy in (d). (e) Summary of the improvement under RC for all two, three, and four-qubit random input QFT results, showing good agreement between experiment (blue) and theory (grey). Simulations in which single-qubit (green) and two-qubit (pink) error rates have been scaled down by a factor of 10 suggest that RC performance increases as error rates decrease. (Error bars on the TVD [$\mathcal{O}(10^{-3})$] for (b), (c), and (d) are smaller than the markers. This figure has been reproduced with permission from Ref. [145].)

served reduction from $d_{\mathrm{TV,bare}} = 0.073(8)$ to $d_{\mathrm{TV,RC}} = 0.008(2)$ for the single-qubit results in Fig. 6.1c, as measured in the computational basis.

In Figs. 6.5a/b, we show that the relative RC performance is best (equivalent) when the algorithm generates a uniform (singular) distribution across all measurement basis states. This is due to the basis-dependence of the TVD: given a small angle error of θ relative to the ideal final state of a system, the TVD scales as $d_{\mathrm{TV}}(\mathcal{P}, \mathcal{P}_{\mathrm{ideal}}) \simeq \theta^2 \simeq r(\mathcal{E})$ if the target state is in an eigenstate of the measurement basis (i.e., deterministic algorithms), whereas the TVD scales as $d_{\mathrm{TV}}(\mathcal{P}, \mathcal{P}_{\mathrm{ideal}}) \simeq \theta$ if the target state is coherently spread across the measurement basis (i.e., non-deterministic algorithms; see further discussion below). Therefore, if the (ideal) target state is in an eigenstate of the measurement basis, the raw probabilities will not be sensitive to off-diagonal terms in the error process resulting from coherent errors [which scale as $\theta \simeq \sqrt{r(\mathcal{E})}$], so the TVD will not benefit from the general $\sqrt{r(\mathcal{E})} \longrightarrow r(\mathcal{E})$ improvement provided by RC. As such, distribution uniformity is a good proxy for the susceptibility of the target state to coherent errors with respect to the measurement basis, and is thus correlated with improvement under RC. More generally, error assessments that are only sensitive to the diagonal terms in the error process (e.g. process or average gate infidelity, or the TVD of deterministic algorithms) cannot benefit from the suppression of the off-diagonal terms via RC when averaging over many randomizations. In contrast, norm-based error-metrics, such as the TVD of non-deterministic algorithms and the diamond norm, generally will be sensitive to the off-diagonal terms in the error process, and thus generally benefit from RC.

We quantify the *distance from a uniform distribution* by computing the TVD of each ideal probability distribution with the uniform distribution in $d = 2^n$ dimensions for n qubits, $d_{\mathrm{TV}}(\mathcal{P}_{\mathrm{ideal}}, \mathcal{P}_{\mathrm{uniform}})$, which is 0 (maximized) when $\mathcal{P}_{\mathrm{ideal}}$ is uniform (singular). In Fig. 6.5b, the bare and RC TVDs are plotted as a function of $d_{\mathrm{TV}}(\mathcal{P}_{\mathrm{ideal}}, \mathcal{P}_{\mathrm{uniform}})$ for all four-qubit QFT results. For singular input states ($|0000\rangle$ or $|1111\rangle$), RC significantly reduces the TVD, but for a superposition input state ($|++++\rangle$), the bare and RC TVDs are approximately equal. We compute the Pearson correlation coefficient r to quantify the correlation strength between the experiment TVD and $d_{\mathrm{TV}}(\mathcal{P}_{\mathrm{ideal}}, \mathcal{P}_{\mathrm{uniform}})$, where +1 (-1) indicates exact positive (negative) correlation and 0 implies no linear correlation. The RC results are strongly correlated [$r = 0.95$ (0.80) for basis (random) inputs] compared to the bare results [$r = 0.66$ (0.32) for basis (random) inputs], underscoring the stability and predictability of RC compared to non-randomized circuits. A summary of the TVD improvement under RC for all two, three, and four-qubit results can be seen in Fig. 6.7, plotted as the ratio of the bare to RC TVDs, $d_{\mathrm{TV,bare}}/d_{\mathrm{TV,RC}}$, as a function of the distribution uniformity of the ideal results.

In Fig. 6.5c, we predict the TVD performance of the QFT using two models: (1) a Pauli model of our system consisting of the Pauli error rates extracted from the CER results in Fig. 5.10b; (2) a coherent error model which, under simulated CER, produces approximately equal error rates as the experimental results in Fig. 5.10b. The coherent error model is generated by finding a desired CPTP map for a fixed unitarity [139] (i.e. fixed fraction of the total error rate due to coherent errors) which has been minimized with respect to

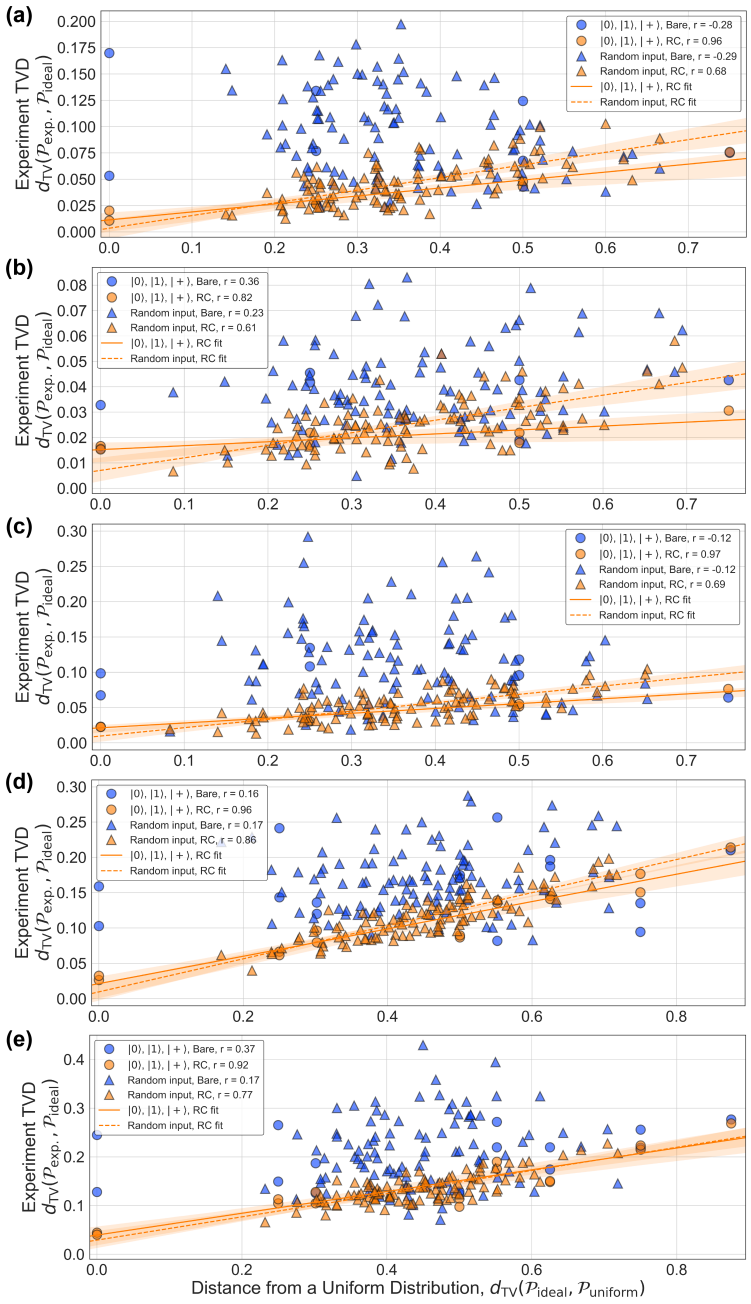


Figure 6.6: **Two- and three-qubit quantum Fourier transform results:** (a) Q4 & Q5, (b) Q5 & Q6, (c) Q6 & Q7, (d) Q4, Q5, & Q6, and (e) Q5, Q6, & Q7. Pearson r values listed in the legend indicate the linear correlation of each data set. Linear fits are plotted for the RC data, with transparent bands indicating the 95% confidence intervals.

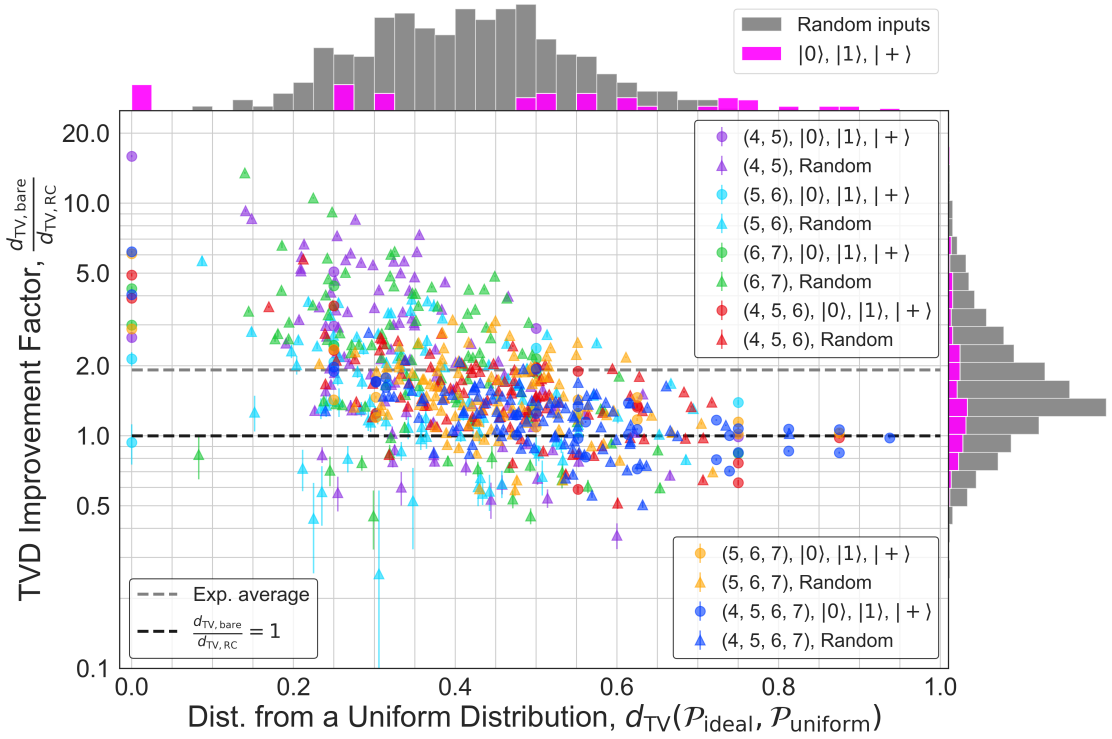


Figure 6.7: **TVD improvement under randomized compiling.** Experimental results of the TVD improvement under RC for the quantum Fourier transform applied to two, three, and four qubits, as a function of the distribution uniformity of the ideal results. The qubit subsets are specified in the legend. Histograms of the Pauli input (pink) and random input (grey) results are included on both axes to show how the results are distributed. RC performs better for results in which $d_{\mathrm{TV,bare}}/d_{\mathrm{TV,RC}} > 1$, but performs worse for results in which $d_{\mathrm{TV,bare}}/d_{\mathrm{TV,RC}} < 1$. For all random input results, the average improvement is $d_{\mathrm{TV,bare}}/d_{\mathrm{TV,RC}} = 1.92$.

the experimental CER results (see "Simulation Model" subsection below). While based on experimental error rates, such a model does not accurately capture detailed information about coherent errors in our system (e.g. context-dependent rotation axes and angles [158]). Therefore, we do not expect the coherent error model to accurately simulate the performance of any individual bare circuit; rather, we expect it to only capture the average performance of the bare circuits, which is possible as long as the magnitude of the unitarity in the model is correct. We note that, in general, measuring an accurate coherent error model for multi-qubit systems with a continuous single-qubit gate set is not experimentally tractable.

Using these models, we simulate the QFT circuits and compare the results to the experimental TVDs in three different ways:

1. Coherent error model applied to the bare circuits vs. experimental bare results

2. Coherent error model applied to the RC circuits vs. experimental RC results

3. Pauli model applied to the bare circuits vs. experimental RC results

For (3), it is unnecessary to simulate the RC circuits with the Pauli model, since the Pauli model already contains the tailored error rates. In Fig. 6.5c, we see excellent agreement between experiment and simulation for (2) and (3), but unreliable predictability (1). This indicates that we can accurately predict the results of an algorithm *a priori* with CER error rates using (a) a Pauli model, as long as the experimental circuit is performed using RC, and (b) a coherent error model, as long as both the simulated and experimental circuits are performed using RC. In Fig. 6.5d, we validate the model accuracy by computing the TVD of the experimental results with the simulated results. Almost all of the simulated results for (2) and (3) are accurate to within 10% with respect to the experimental RC results, but the accuracy of the simulated results for (1) are much worse. By utilizing RC in conjunction with CB/CER, we avoid the difficulty in capturing and modeling the complex interplay of coherent errors, and thus close the gap between the circuit performance predicted from benchmarking diagnostics and experimental results.

In practice, the input states to quantum algorithms will not be known *a priori*, such as when the QFT is used in Shor's algorithm [106]. While unknown inputs are not guaranteed to be random, we use random inputs as a proxy for when the QFT is used as a subroutine in other algorithms. Fig. 6.5b shows that when the QFT is applied to random input states, most of the results are improved under RC. A histogram of the TVD improvement for two, three, and four-qubit random input QFT results can be seen in Fig. 6.5e, showing that the vast majority of circuits ($> 81\%$) are improved under RC by an average of $d_{\mathrm{TV,bare}}/d_{\mathrm{TV,RC}} \approx 1.9$. Here, we include two- and three-qubit results in order to summarize the RC QFT performance using a larger sample size drawn from systems that include differing error rates. In the rare instances in which coherent errors in a circuit benignly cancel, RC can hurt performance ($d_{\mathrm{TV,bare}}/d_{\mathrm{TV,RC}} < 1$); however, in general, this becomes vanishingly unlikely for longer depth circuits.

While the coherent error model may not accurately predict the individual result of any given bare circuit, it does predict the average performance relative to RC. To demonstrate this, we compute the TVD improvement under RC ($d_{\mathrm{TV,bare}}/d_{\mathrm{TV,RC}}$) for all of the two, three, and four-qubit random input QFT circuits simulated using the coherent error model. As seen in Fig. 6.5e, the distribution of improvement under RC predicted by simulation agrees well with experiment, with an overlapping index of 0.94 (out of a maximum of 1), which quantifies the percentage that one normal distribution overlaps with another. The good agreement between experiment and theory in Fig. 6.5e suggests that we can predict the average improvement under RC as error rates decrease. Included in Fig. 6.5e are simulated results in which single-qubit error rates are reduced by a factor of 10, resulting in a modest improvement, and when both single- and two-qubit error rates are reduced by a factor of 10, in which case RC improves $> 94\%$ of the simulated circuits by an average of $d_{\mathrm{TV,bare}}/d_{\mathrm{TV,RC}} \approx 3.4$. In agreement with the predictions made in Ref. [174], these results demonstrate that RC

is expected to provide a larger relative improvement as gate infidelities decrease (for a fixed fraction of the total error rate due to coherent errors). Therefore, as quantum processors improve and error rates decrease, we can expect RC to outperform non-randomized circuits as long as coherent errors persist.

Basis-Dependence of the TVD

As a norm-based error metric, the TVD is generally sensitive to the off-diagonal terms in an error process resulting from coherent errors. Therefore, in the presence of coherent errors the TVD can be as large as

$$d_{\mathrm{TV}}(\mathcal{P}, \mathcal{P}_{\mathrm{ideal}}) \leq \sqrt{r(\mathcal{E})}\sqrt{d(d+1)}, \tag{6.9}$$

but under RC it is instead upper-bounded directly by the average error rate $r(\mathcal{E})$,

$$d_{\mathrm{TV}}(\mathcal{P}_{\mathrm{RC}}, \mathcal{P}_{\mathrm{ideal}}) \leq r(\mathcal{E})\frac{d+1}{d}, \tag{6.10}$$

which is quadratically lower in $r(\mathcal{E})$ and does not scale with the dimension $d = 2^n$ (n qubits). Thus, RC provides a general error reduction from $\sqrt{r(\mathcal{E})} \longrightarrow r(\mathcal{E})$. Because the TVD depends on simulating the ideal results of a quantum circuit, it is not a scalable error metric. Therefore, it is important to benchmark the TVD performance under RC while quantum circuits can still be efficiently classically simulated, which will inform what TVD improvements can be expected under RC in the post-NISQ era.

This quadratic improvement depends on the linearity of an error metric and the degree to which it is sensitive to off-diagonal terms in the error process. Whereas the fidelity is always insensitive to off-diagonal terms in the error process, the TVD is generally sensitive to these terms, except for the case in which the target state is an eigenstate of the measurement basis. To see this, Consider the case in which $\rho_{\mathrm{ideal}} = |0\rangle\langle 0|$, but the actual qubit state has over-rotated by an angle θ due to coherent errors: $|\psi\rangle = \cos(\theta)\,|0\rangle + \sin(\theta)\,|1\rangle$. The fidelity of ρ with ρ_{ideal} is

$$\mathcal{F} = |\,\langle 0|\psi\rangle\,|^2 = \cos^2(\theta) \approx 1 - \theta^2, \tag{6.11}$$

where we have approximated $\cos(\theta)$ in the small angle limit. Here, we can see that to first order the infidelity scales as $r(\mathcal{E}) \simeq \theta^2$. We can similarly compute the TVD between ρ and ρ_{ideal} as measured in the computational basis $|x\rangle \in \{|0\rangle, |1\rangle\}$,

$$d_{\mathrm{TV}}(\mathcal{P}, \mathcal{P}_{\mathrm{ideal}}) = \frac{1}{2}\sum_{x=0,1} \left|\,\mathrm{Tr}[(|\psi\rangle\langle\psi| - |0\rangle\langle 0|)\,|x\rangle\langle x|]\right| \tag{6.12}$$

$$= \frac{1}{2}\sum_{x=0,1} \left|\,\mathrm{Tr}\left[\begin{pmatrix} \cos^2(\theta) - 1 & \cos(\theta)\sin(\theta) \\ \cos(\theta)\sin(\theta) & \sin^2(\theta) \end{pmatrix} |x\rangle\langle x|\right]\right| \tag{6.13}$$

$$= \sin^2(\theta) \approx \theta^2. \tag{6.14}$$

Similar to the fidelity, we can see that $r(\mathcal{E}) \simeq \theta^2$ for the TVD when the target state is an eigenstate of the measurement basis.

Next, consider the case in which the target state is in a superposition state, $\rho_{\text{ideal}} = |+\rangle\langle+|$, and the actual qubit state has over-rotated due to a small rotation about the y-axis,

$$|\psi\rangle = R_y(\theta)\,|+\rangle \tag{6.15}$$

$$= \frac{1}{\sqrt{2}} \begin{pmatrix} \cos(\theta) - \sin(\theta) \\ \sin(\theta) + \cos(\theta) \end{pmatrix}. \tag{6.16}$$

The fidelity of this state is

$$\mathcal{F} = |\,\langle+|\psi\rangle\,|^2 = \cos^2(\theta) \approx 1 - \theta^2, \tag{6.17}$$

where again we see that $\mathcal{F}$ is insensitive to the off-diagonal terms due to $R_y(\theta)$ and that the infidelity scales as $r(\mathcal{E}) \simeq \theta^2$. This is, however, not the case for the TVD, where

$$d_{\text{TV}}(\mathcal{P}, \mathcal{P}_{\text{ideal}}) = \frac{1}{2} \sum_{x=0,1} \left|\, \text{Tr}[(|\psi\rangle\langle\psi| - |+\rangle\langle+|)\,|x\rangle\langle x|] \right| \tag{6.18}$$

$$= \frac{1}{2} \sum_{x=0,1} \left| \text{Tr} \left[\begin{pmatrix} -2\cos(\theta)\sin(\theta) & \cos^2(\theta) - \sin^2(\theta) - 1 \\ \cos^2(\theta) - \sin^2(\theta) - 1 & 2\cos(\theta)\sin(\theta) \end{pmatrix} |x\rangle\langle x| \right] \right| \tag{6.19}$$

$$= \cos(\theta)\sin(\theta) \approx \theta. \tag{6.20}$$

Therefore, we can see that when the target state is coherently spread out among the various basis states, the TVD scales as $\sqrt{r(\mathcal{E})} \simeq \theta$, which is quadratically larger than the TVD for deterministic algorithms. Consequently, the TVD can benefit greatly from the $\sqrt{r(\mathcal{E})} \longrightarrow r(\mathcal{E})$ reduction in the error under RC for non-deterministic algorithms.

It is also possible to reformulate the basis-dependent improvements of RC in terms of whether or not an error measure is linear in the output state ρ. To see this, consider the noisy output state $\mathcal{E}(\rho)$ prepared by a single randomization under RC. Here, the fidelity is linear in ρ:

$$\mathcal{F} = \langle\phi|\,\mathcal{E}(\rho)\,|\phi\rangle\,, \tag{6.21}$$

where the ideal final state is $\rho_{\text{ideal}} = |\phi\rangle\langle\phi|$. When averaging over many randomizations,

$$\mathcal{F} = \langle\phi|\,\frac{1}{N}\sum_i^N \mathcal{E}_i(\rho)\,|\phi\rangle \tag{6.22}$$

$$= \frac{1}{N}\sum_i^N \langle\phi|\,\mathcal{E}_i(\rho)\,|\phi\rangle\,, \tag{6.23}$$

where, in the second line, the sum over N randomizations can be taken outside of the inner product. Thus, if a single randomization provides no benefit over the bare circuit, in the many-randomization limit the fidelity is simply an average of N incorrect results. We note here, however, that a single randomization under RC can still provide an advantage for circuits impacted by structured errors by dynamically-decoupling the qubits from noise and breaking up the coherent accumulation of unitary errors (similar to echoed pulse sequences). Therefore, metrics that are linear in ρ can still benefit from RC if the noisy preparation $\mathcal{E}_{\mathrm{RC}}(\rho)$ is more accurate than $\mathcal{E}(\rho)$ in the single-randomization limit.

In contrast, TVD is generally not linear in ρ, and thus the average over N randomizations cannot be pulled outside of the absolute value. To see this, we first note that the trace distance is convex in its first input (Eq. 5.31), from which it follows that

$$D\left(\frac{1}{N}\sum_i^N \mathcal{E}_i(\rho), \rho_{\mathrm{ideal}}\right) = \frac{1}{2}\,\mathrm{Tr}\left|\frac{1}{N}\sum_i^N \mathcal{E}_i(\rho) - \rho_{\mathrm{ideal}}\right| \tag{6.24}$$

$$\leq \frac{1}{2N}\sum_i^N \mathrm{Tr}\,|\mathcal{E}_i(\rho) - \rho_{\mathrm{ideal}}|. \tag{6.25}$$

The equality in the above equation is saturated only for deterministic algorithms, in which the target state is an eigenstate of the measurement basis. To understand why this is the case, consider measurements made in the computation basis $|x\rangle \in \{|0\rangle, |1\rangle\}$, where $\mathcal{P}(x) = \mathrm{Tr}[\mathcal{E}(\rho)\,|x\rangle\langle x|] = \langle x|\,\mathcal{E}(\rho)\,|x\rangle$ and $\mathcal{P}_{\mathrm{ideal}}(x) = \mathrm{Tr}[\rho_{\mathrm{ideal}}\,|x\rangle\langle x|] = \langle x|\,\rho_{\mathrm{ideal}}\,|x\rangle$. If we take $\rho_{\mathrm{ideal}} = |0\rangle\langle 0|$, then $\mathcal{P}_{\mathrm{ideal}}(0) = 1$ and $\mathcal{P}_{\mathrm{ideal}}(1) = 0$. Therefore, we can write the TVD as

$$d_{\mathrm{TV}}(\mathcal{P}, \mathcal{P}_{\mathrm{ideal}}) = \frac{1}{2}\sum_{x\in\{0,1\}} |\mathcal{P}(x) - \mathcal{P}_{\mathrm{ideal}}(x)| \tag{6.26}$$

$$= \frac{1}{2}\big[|\mathcal{P}(0) - \mathcal{P}_{\mathrm{ideal}}(0)| + |\mathcal{P}(1) - \mathcal{P}_{\mathrm{ideal}}(1)|\big] \tag{6.27}$$

$$= \frac{1}{2}\big[(1 - \mathcal{P}(0)) + \mathcal{P}(1)\big] \tag{6.28}$$

$$= \mathcal{P}(1) \tag{6.29}$$

$$= \langle 1|\,\mathcal{E}(\rho)\,|1\rangle\,. \tag{6.30}$$

In this case, the TVD is linear in ρ, and when $\mathcal{E}(\rho)$ is the average over many randomizations under RC, we can write

$$d_{\mathrm{TV}}(\mathcal{P}, \mathcal{P}_{\mathrm{ideal}}) = \langle 1|\,\frac{1}{N}\sum_i^N \mathcal{E}_i(\rho)\,|1\rangle \tag{6.31}$$

$$= \frac{1}{N}\sum_i^N \langle 1|\,\mathcal{E}_i(\rho)\,|1\rangle\,, \tag{6.32}$$

where the TVD is now simply an average over N independent noisy results, all of which are incorrect.

TVD vs. Expectation Values in Quantum Algorithms

While many NISQ applications (e.g. variational quantum algorithms) depend on measured expectation values as opposed to the TVD, the TVD upper bounds the absolute error of all expectation values measured in the same basis, therefore a small TVD under RC guarantees a small error in any expectation value estimated from the same probability distribution. To see this, consider the expectation value of an operator A acting on a quantum state ρ is given as

$$\langle A \rangle_\rho = \mathrm{Tr}[\rho A] = \sum_x \mathcal{P}(x) \langle x| A |x\rangle = \sum_x \mathcal{P}(x) \langle A \rangle_x , \tag{6.33}$$

where we have written the ρ in terms of its spectral decomposition, and $\langle A \rangle_x$ is the expectation value of A in the $|x\rangle$ basis. We can write the absolute error in the expectation value as

$$| \langle A \rangle_\rho - \langle A \rangle_{\rho,\mathrm{true}} | = \Big| \sum_x (\mathcal{P}(x) - \mathcal{P}_{\mathrm{true}}(x)) \langle A \rangle_x \Big|. \tag{6.34}$$

It is then possible to bound the absolute error of this expectation value with a simple triangle inequality,

$$| \langle A \rangle_\rho - \langle A \rangle_{\rho,\mathrm{true}} | \leq \sum_x \big|(\mathcal{P}(x) - \mathcal{P}_{\mathrm{true}}(x))\big| \langle A \rangle_x | \tag{6.35}$$

$$\leq 2 d_{\mathrm{TV}}(\mathcal{P}, \mathcal{P}_{\mathrm{true}}) \|A\|, \tag{6.36}$$

where $\|A\|$ is the operator norm of A, which is defined to be the maximal absolute value of all eigenvalues of A. Therefore, we can say that the absolute error in the expectation value of an arbitrary observable is upper-bounded by the TVD computed from the same probability distribution. While a reduction in the TVD under RC does not guarantee a reduction in the error of an expectation value measured in the same basis, since the error in the expectation value is only at most given by the TVD, it does tighten the bound within which one can accurately estimate the true expectation value. Since multiple expectation values can be estimated from any given probability distribution, the TVD effectively condenses information about the errors across all the relevant expectation values. In other words, if $d_{\mathrm{TV}}(\mathcal{P}, \mathcal{P}_{\mathrm{true}}) < \epsilon$, then the expectation value of all local observables will be accurate to within ϵ. We therefore generally expect RC to improve the accuracy of expectation values in scenarios in which it improves the TVD.

Simulation Model

Simulated QFT results were generated using two models: (1) a Pauli model, using the Pauli error rates directly extracted from the CER results in Fig. 5.10b; (2) a coherent error model of our system that includes coherent errors, which produces simulated CER results that are approximately equal to the error rates in Fig. 5.10b. The coherent model of our four-qubit system is generated by finding a desired CPTP map for a given unitarity by minimizing simulated CER results with respect to the experimental results; we only considered single-body

errors in the construction of this model, as these are the most dominant error syndromes in our system. Our simulator takes into account realistic values for SPAM errors and implements coherent errors and stochastic noise in each cycle of a circuit. Stochastic noise acts on each qubit per cycle with a finite probability (Eq. 3.31), and coherent errors are implemented by adding an over-rotation ϵ to each qubit (Eq. 3.14), where ϵ is set according to the process infidelity due to coherent errors. A similar model is used for simulating coherent errors on two-qubit gates.

The circuits we are interested in simulating, as with those performed experimentally, have been structured to alternate between cycles containing CNOTs (hard cycles) and cycles of single qubit gates (easy cycles). Simulation comprises of two steps: first, we define a noise channel for every unique hard cycle in the collection of circuits to be simulated. Next, given a circuit, we propagate the initial state by simulating each easy cycle ideally, and each hard cycle ideally but followed by its corresponding noise channel. We do not add any noise to the easy cycles not because we are assuming they are perfect in practice, but rather because our noisy channels will be constructed using CER data, which measures the composition of the errors due to a cycle of random Pauli gates along with a particular hard cycle (see Eq. 5.61). We do, however, assume no correlations between the errors of each gate in a given cycle: we take the total noise channel of a cycle as the tensor product of noise channels of each gate or idling qubit. This assumption allows for noise induced by classical crosstalk (since our noise model is entirely cycle-dependent, we are explicitly allowing noise on one gate to be conditional on the existence of another gate in the cycle), but not noise induced by quantum crosstalk. Our justification is that our CER data shows no such significant correlations (see Figs. 5.10 and 5.12), so it does not appear necessary to include two-body terms in our model as the most dominant errors are single-body. Therefore, limiting ourselves to only single-body errors is sufficient for producing an approximate error model for our system, and simplifies the process of finding the model. In general, measuring all n-body errors is manifestly not scalable as the number of such errors grows exponentially with n. Moreover, the probabilities of two-body errors are the sum of the probabilities of all errors that act non-trivially on the corresponding two bodies, irrespective of their action on other bodies. Therefore, the fact that two-body errors are negligible shows that three or more body errors are also negligible. In what follows, we describe the procedure to generate a noise channel for a specific gate or idling qubit in a given cycle.

First, given a probability simplex $q \in \mathbb{R}^N$ and a real vector $h \in \mathbb{R}^{N-1}$ we can define a quantum channel

$$S(q, h) = \mathcal{U}_h \mathcal{K}_q, \tag{6.37}$$

where $\mathcal{U}_h$ is the unitary superoperator corresponding to the unitary matrix

$$U_h = \exp\left(-i \sum_{i=1}^{N-1} h_i P_i\right), \tag{6.38}$$

and where $\mathcal{K}_q$ is the Pauli Kraus channel

$$\mathcal{K}_q(\rho) = \sum_{i=0}^{N} \sqrt{q_i} P_i \rho P_i^\dagger. \tag{6.39}$$

Here, $P_0, ..., P_{N-1}$ with $N = 4^n$ is some enumeration of the n-qubit Paulis such that $P_0 = I$. $S(q, h)$ defines a somewhat arbitrary but large class of CPTP channels which is sufficient but not necessary to suit our needs: we require a parameterized class of CPTP channels over which to perform a numerical search.

Next, given any pair (q, h) we can compute the Pauli transfer matrix (PTM) of $S(q, h)$, whose diagonal vector we denote as $d(q, h)$. We can likewise compute the unitarity of the channel $S(q, h)$ by taking the 2-norm of the lower PTM block [139], denoting it by $u(q, h)$.

Finally, suppose that for some subset of qubits we experimentally measure the PTM diagonal to have a value of $f \in \mathbb{R}^N$. This is naturally done with CER (see, for example the "(5,4): CX" block of the first cycle in Fig. 5.10), where the error Kraus probabilities are reported for the qubits (4, 5), from which the PTM diagonal for these qubits can be constructed via the inverse Walsh-Hadamard transform [146]. We wish to define a channel for simulation whose PTM diagonal matches f up to a user-defined scaling, and such that the channel has a user-defined unitarity. Therefore, we perform a numerical optimization to find (q, h) such that

$$d(q, h) = 1 - s_0(1 - f) \tag{6.40}$$

$$u(q, h) = 1 - (1 - s_1)(1 - \overline{f}^2) \tag{6.41}$$

where $s_0 \in [0, 1]$ defines the factor with which to decrease the process infidelity (recall the process infidelity corresponding to f is $1 - \overline{f}$ where $\overline{f}$ is the mean value of f), and where $s_1 \in [0, 1]$ defines the unitarity fraction, where a value of 1 results in $S(q, h)$ being unitary, and 0 results in $S(q, h)$ being as stochastic as possible given the constraints. This minimization is performed with SciPy's BGFS solver for all non-overlapping subsets of qubits of the device, which is valid because the experimental data show no significant correlated error between gate-bodies. Minima are consistently found to within numerical precision, though minimum values are not unique.

The tensor product of the resulting channels, one for each gate or idling qubit in the cycle, defines the noisy channel for the given hard cycle. We allow a different value of s for single- and two-qubit subsets. Guided by our CB results, RB, and PB measurements, s_1 was set to 0.7 (0.9) for single-qubit (two-qubit) gates for all simulated results. $s_0 = 1.0$ for the results in Fig. 6.8 to simulate error rates that are equivalent to experimental values. However, $s_0 = 0.1(1.0)$ for single-qubit (two-qubit) gates for the results in Fig. 6.9, and $s_0 = 0.1$ for both single- and two-qubit gates for the results in Fig. 6.10. The histograms of the TVD improvement for random input states under RC for the simulated results (using the coherent error model) in Figs. 6.8, 6.9, and 6.10 are included in Fig. 6.5e.

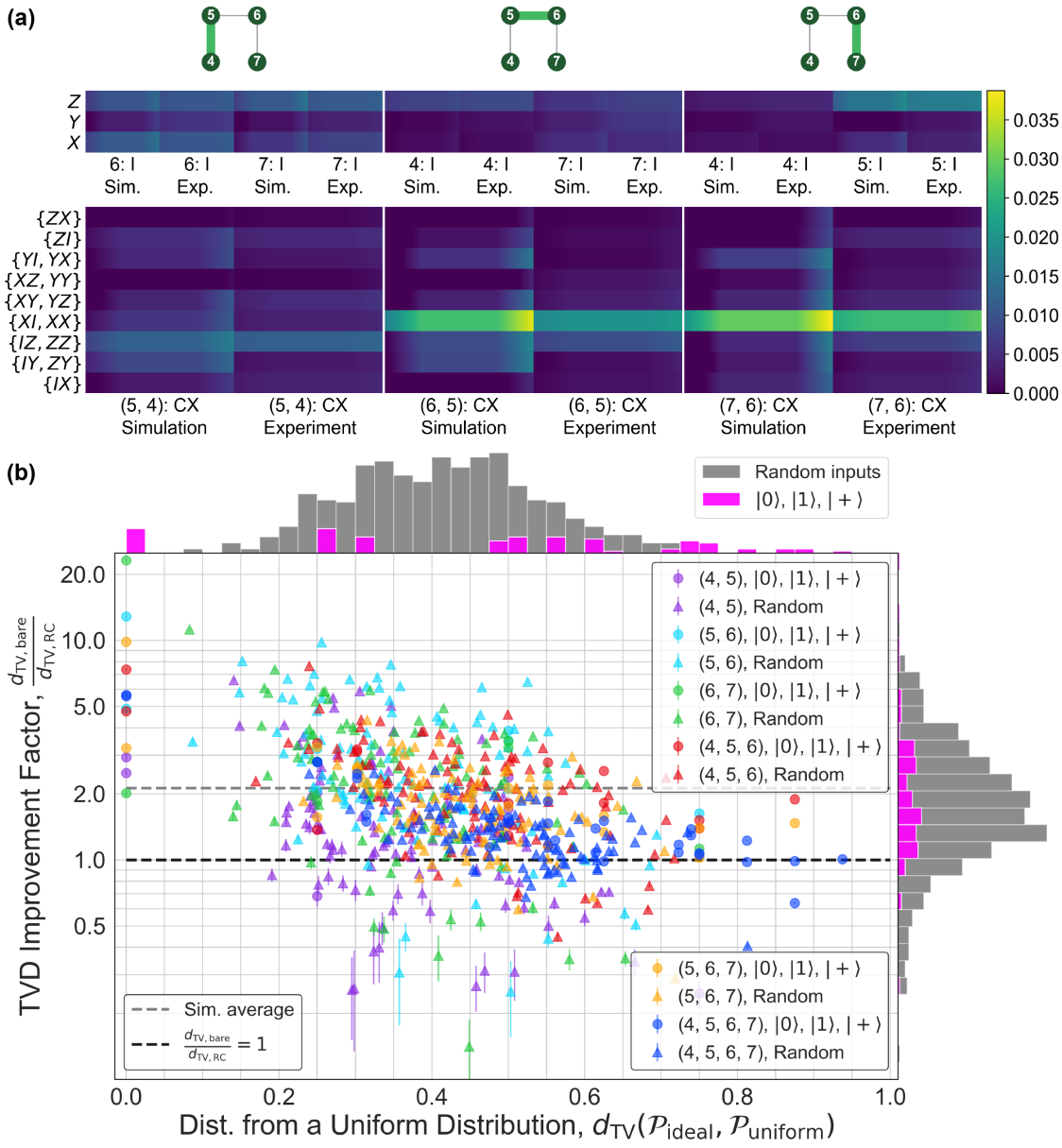


Figure 6.8: **Simulated QFT results using the coherent error model with equivalent error rates.** (a) Experimental vs. simulated CER results based on a model of our system with equivalent single-body error rates in which the fraction of the total error rate due to coherent errors was set to 0.7 (0.9) for single-qubit (two-qubit) gates. (b) Simulated results for data in Fig. 6.7 using the model presented in (a). The average improvement is $d_{\mathrm{TV,bare}}/d_{\mathrm{TV,RC}} = 2.13$ for all random input results, showing good agreement with the experimental results in Fig. 6.7.

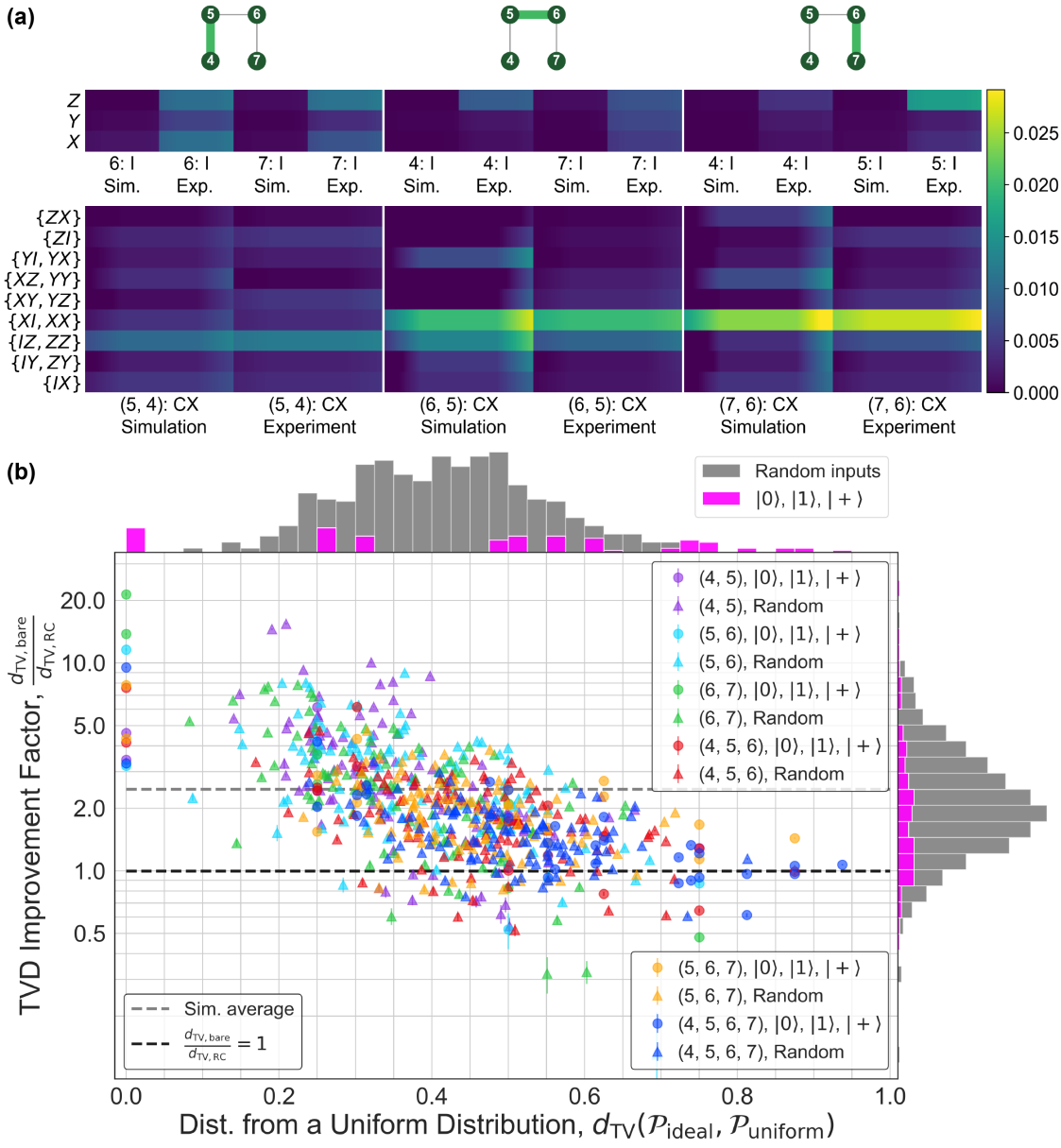


Figure 6.9: **Simulated QFT results using a complete model with improved single-qubit error rates.** (a) Experimental vs. simulated CER results, with the single-qubit error rates reduced by a factor of 10 compared to the model presented in Fig. 6.8. (b) Simulated results for data in Fig. 6.7 using the model presented in (a). The average improvement is $d_{\mathrm{TV,bare}}/d_{\mathrm{TV,RC}} = 2.47$ for all random input results.

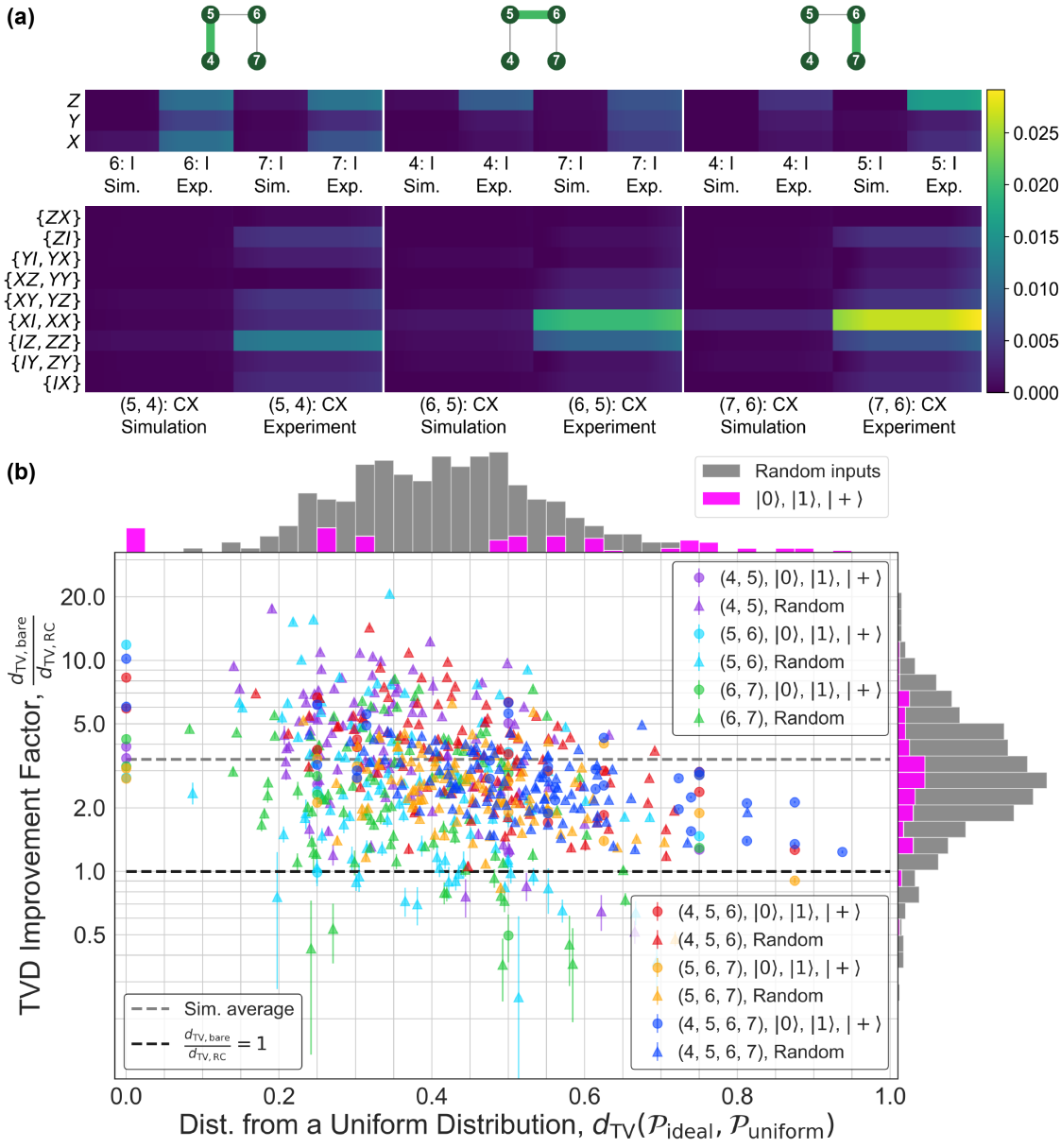


Figure 6.10: **Simulated QFT results using a complete model with improved single- and two-qubit error rates.** (a) Experimental vs. simulated CER results, with both the single- and two-qubit error rates reduced by a factor of 10 compared to the model presented in Fig. 6.8. (b) Simulated results for data in Fig. 6.7 using the model presented in (a). The average improvement is $d_{\mathrm{TV,bare}}/d_{\mathrm{TV,RC}} = 3.39$ for all random input results.

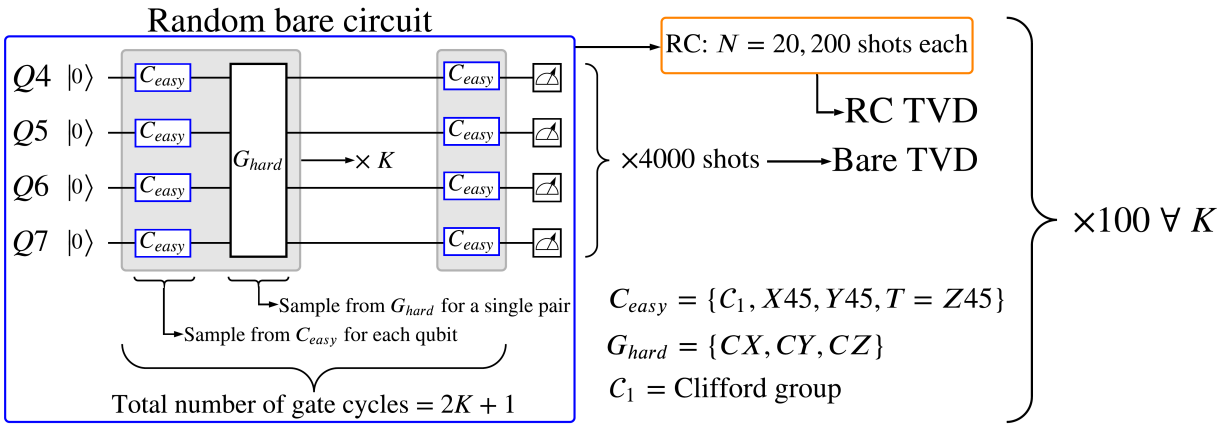


Figure 6.11: **Random circuit sampling.** The process by which K cycles of interleaved easy/hard gates are randomly sampled from a universal gate set. All random bare circuits were measured 4,000 times. $N = 20$ randomizations were generated for each bare circuit, each of which was measured 200 times. 100 random bare circuits were generated for each circuit depth K.

6.4 Random Circuits of Variable Depth

To illustrate the broad applicability and generic benefits of RC for universal circuits, we demonstrate achievable performance gains for RC applied to four-qubit circuits of variable depth composed of K interleaved cycles of easy/hard gates randomly sampled from a universal (Clifford + T) gate set (see Fig. 6.11). Random bare circuits were generated by randomly sampling K interleaved cycles of easy C_{easy} and hard G_{hard} gates from a universal gate set. For the (isolated and simultaneous) single-qubit circuits, $C_{easy} = \{\mathbf{C_1}\}$ and $G_{hard} = \{X45,\ Y45,\ T = Z45\}$, where $\mathbf{C_1}$ is the single-qubit Clifford set. For multi-qubit circuits involving entangling operations, $C_{easy} = \{\mathbf{C_1},\ X45,\ Y45,\ T\}$ and $G_{hard} = \{CX = CNOT,\ CY,\ CZ\}$. For easy gate cycles, single-qubit gates are randomly sampled from C_{easy} independently for each qubit. For hard gate cycles in single-qubit circuits, single-qubit gates are randomly sampled from G_{hard} independently for each qubit. For hard gate cycles in multi-qubit circuits, a two-qubit gate is sampled from G_{hard} for a single pair of nearest-neighbor qubits, and identity gates are applied to the remaining spectator qubits. Because the $\{H,\ S,\ T,\ CNOT\}$ set of gates can be used for universal quantum computation, and H, S, and $CNOT$ are all Clifford gates, this is typically referred to as the universal "Clifford + T" set.

As shown in Fig. 6.12a, RC reduces the average TVD at all circuit depths tested (with $N = 20$ randomizations for each bare circuit), demonstrating how longer-depth quantum circuits can be performed under RC given a fixed error budget in the TVD. These results highlight an important distinction in how RC can improve algorithm performance: because

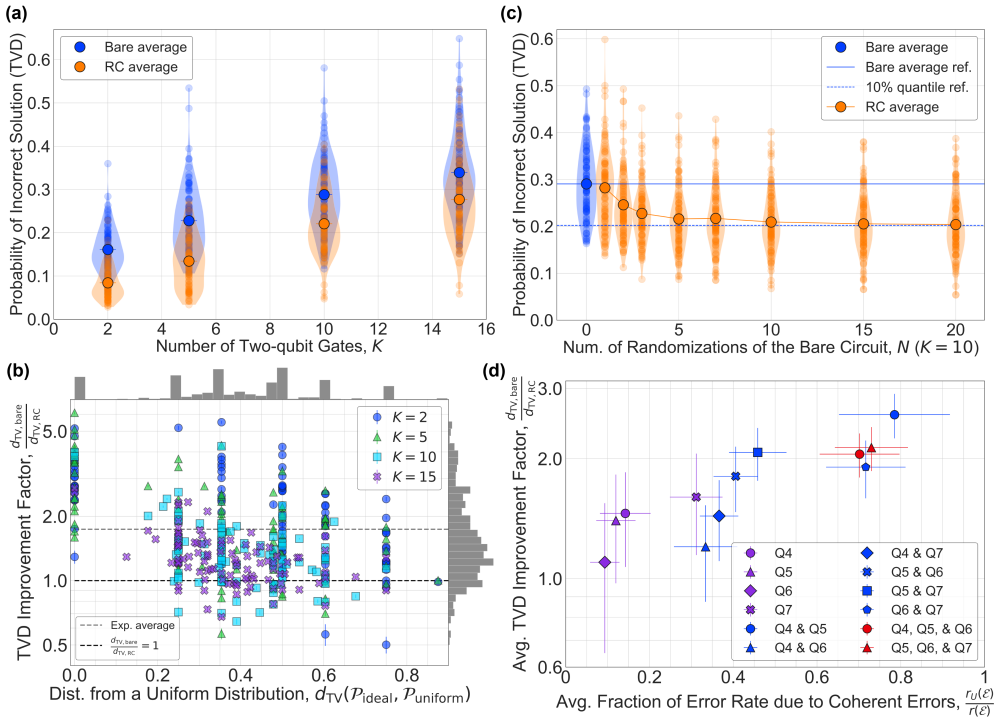


Figure 6.12: **RC extends the computational reach with respect to circuit depth.** (a) Bare and RC TVDs as a function of circuit depth K. RC reduces the TVD on average for all circuit depths tested, allowing one to perform longer gate sequences under a fixed TVD error budget. The semi-transparent blue (orange) points indicate the TVDs of the individual random circuits (the unioned data over all $N = 20$ randomizations of the corresponding bare circuits). Violin plots depict the distribution of results. The TVD error grows approximately linearly with K for both bare and RC results, suggesting that the dominant reason for an improvement under RC is an overall reduction in the error for each gate cycle K, rather than a suppression of the adversarial accumulation of coherent errors (although RC can additionally provide this benefit). (b) RC TVD improvement factor for the random circuits in (a) as a function of distribution uniformity. The average improvement is $d_{\mathrm{TV,bare}}/d_{\mathrm{TV,RC}} \approx 1.7$. (c) TVD as a function of number of randomizations, with $K = 10$ fixed. The average TVD under RC converges to a value close to the 10% quantile level (dashed line) of the non-randomized circuits for $N = 20$. However, only $N = 10$ randomizations are needed to converge to within 2.7% of the $N = 20$ level. (d) For a fixed total error rate, RC provides a larger TVD improvement for systems with a higher fraction of coherent errors. The colored subsets listed in the legend highlight random single-qubit circuits that were performed in isolation (purple) or in parallel (blue and red). The average fraction of the total error rate due to coherent errors was quantified using measurements of RB and PB under isolated or simultaneous operation. (Figure reproduced with permission from Ref. [145].)

Qubit(s)	Q4		Q5		Q6		Q7	
	e_F	e_U	e_F	e_U	e_F	e_U	e_F	e_U
Q4	1.2 (0.57)	0.16 (0.70)						
Q5			1.1 (0.46)	0.14 (0.54)				
Q6					1.4 (0.35)	0.13 (0.48)		
Q7							1.9 (1.1)	0.59 (1.1)
Q4 & Q5	4.8 (6.9)	3.7 (6.9)	6.0 (9.0)	4.9 (9.1)				
Q4 & Q6	1.1 (4.8)	0.19 (0.71)			2.0 (2.2)	0.97 (2.4)		
Q4 & Q7	0.96 (0.46)	0.23 (0.54)					1.7 (1.0)	0.85 (1.1)
Q5 & Q6			2.2 (1.6)	1.2 (1.6)	1.5 (0.87)	0.40 (0.95)		
Q5 & Q7			1.5 (1.0)	0.57 (1.1)			2.4 (2.3)	1.3 (2.5)
Q6 & Q7					4.8 (5.7)	3.7 (5.7)	3.9 (3.7)	2.6 (3.7)
Q4, Q5, & Q6	3.9 (5.1)	2.8 (5.1)	8.3 (13.0)	6.7 (13.0)	3.2 (3.1)	1.9 (3.4)		
Q5, Q6, & Q7			3.5 (3.9)	2.5 (3.9)	5.7 (7.8)	4.1 (7.9)	4.4 (4.9)	3.3 (5.0)

Table 6.2: **RB (e_F) and unitary RB (e_U) process infidelities measured before the random single-qubit circuits of variable depth experiments.** These values are used to quantify the fraction of the total error rate due to coherent errors [see Fig. 6.12d]. All process infidelities are $\times 10^{-3}$ and all standard deviations are $\times 10^{-4}$.

random circuits are already robust to the coherent accumulation of unitary errors, we observe a linear (not quadratic) growth in the average TVD as a function of circuit depth for both the bare and RC circuits. Therefore, the reduction in the TVD under RC is not due to the suppression of the adversarial accumulation of coherent errors (although each randomization under RC can additionally provide this benefit for structured circuits that are susceptible to such errors). Rather, it is due to a reduction in the off-diagonal terms in the error process due to coherent errors per computational gate cycle K when averaging over many randomizations. As previously noted, this improvement is not observed for fidelity-based error metrics (see, for example, the state fidelity results in Fig. 6.1c), which underrepresent the global impact of coherent errors on quantum algorithms, as they are only sensitive to the diagonal terms in the error process. The relative improvement under RC is reduced at longer circuit depths, since both the bare and RC results will converge to a uniform distribution (i.e. statistical mixture) due to decoherence in the limit of large K.

The TVD improvement under RC for all of the random circuits in Fig. 6.12a is plotted in Fig. 6.12b as a function of $d_{\mathrm{TV}}(\mathcal{P}_{\mathrm{ideal}}, \mathcal{P}_{\mathrm{uniform}})$, with an average improvement of $d_{\mathrm{TV,bare}}/d_{\mathrm{TV,RC}} \approx 1.7$. Because randomly sampled circuits approach an approximate unitary-2 design as a function of circuit depth [186], we observe that the results are not uniformly spread across $d_{\mathrm{TV}}(\mathcal{P}_{\mathrm{ideal}}, \mathcal{P}_{\mathrm{uniform}})$ for small K; rather, they are highly concentrated at several uniformities, and only begin spreading out for larger K. Given the concentration of results

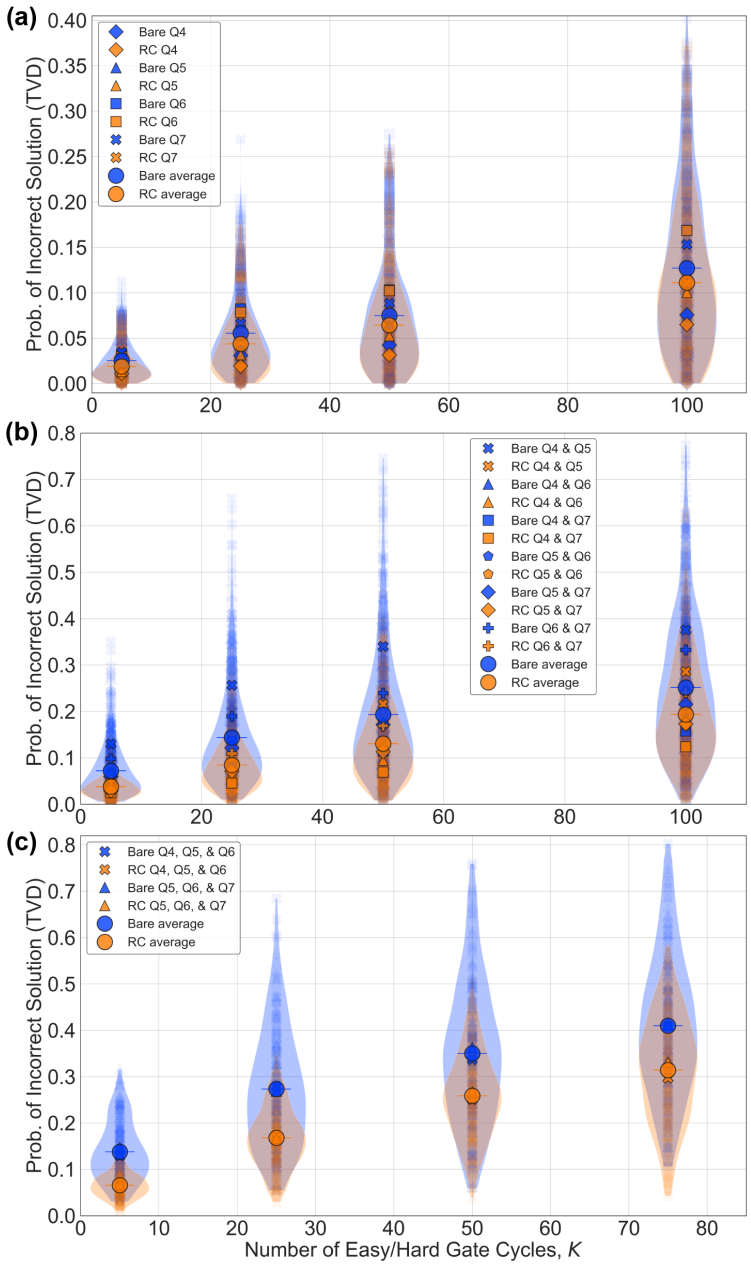


Figure 6.13: **Single-qubit random circuits of variable depth.** RB (e_F) and PB (e_U) process infidelities were measured before each set of experiments to quantify the fraction of the total error rate due to coherent errors (see Table 6.2). (a) Isolated single-qubit circuits. (b) Simultaneous single-qubit circuits on two qubits. (c) Simultaneous single-qubit circuits on three qubits. The $K = 5$ data from each of the above plots was used in Fig. 6.12d.

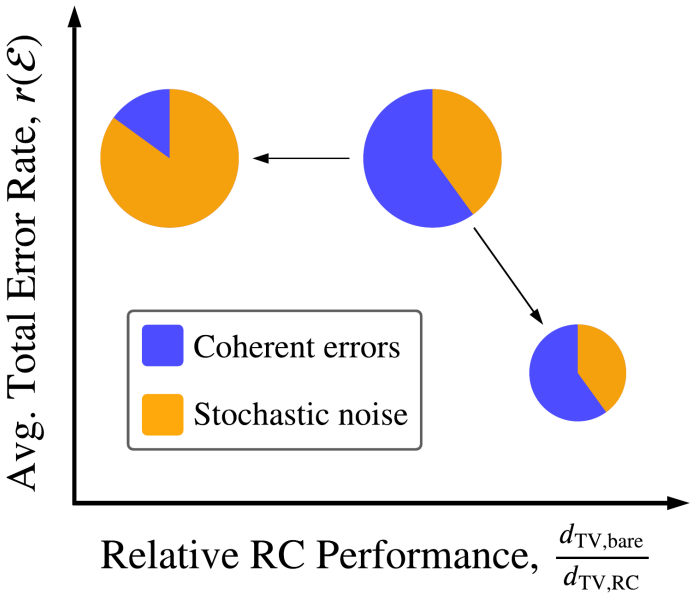


Figure 6.14: **Expected performance scaling of RC.** As the fraction of the total error rate due to coherent errors decreases (for a fixed total error rate, depicted by the area of the circles), the relative RC performance is expected to decrease; in the limit of coherence-limited operations, RC can hurt performance. However, as the total error rate decreases (for a fixed fraction of the total error rate due to coherent errors), RC performance is expected to improve.

in the range between $d_{\text{TV}}(\mathcal{P}_{\text{ideal}}, \mathcal{P}_{\text{uniform}}) \in [0.2, 0.8]$ in both Fig. 6.5b and Fig. 6.12b, we would expect typical algorithms to fall within this range.

Additionally, we show that a small number of randomizations is sufficient to saturate the lowest-possible TVD under RC for a fixed circuit depth ($K = 10$), plotted in Fig. 6.12c. After $N = 20$ randomizations, the average TVD under RC converges to a value that is better than approximately 90% of the non-randomized circuits. However, after only $N = 10$ randomizations, the average RC TVD is already within 2.7% of the $N = 20$ level, highlighting the resource-efficiency of this protocol.

Finally, in Fig. 6.12d we plot the average TVD improvement factor for random single-qubit circuits (at a fixed circuit depth $K = 5$) performed in isolation or in parallel as a function of the average fraction of the total error rate due to coherent errors. From these data, we see that RC provides a larger relative TVD improvement as the fraction of the total error rate due to coherent errors increases (for a fixed total error rate). We measured the average total error rate $r(\mathcal{E})$ using RB and the average error rate due to coherent errors $r_U(\mathcal{E})$ using PB. Even though the average total error rate of any two results are not exactly equal, we group the data by the number of qubits performed in parallel (differentiated by color), since

$r_U(\mathcal{E})/r(\mathcal{E})$ can be more directly compared across these subsets independently. While RC performance decreases as $r_U(\mathcal{E})/r(\mathcal{E}) \longrightarrow 0$, we note that even for single-qubit systems which are close to coherence-limited [$r_U(\mathcal{E})/r(\mathcal{E}] \lesssim 0.1$), RC still provides an average improvement. Therefore, while the trade-off between decreased RC improvement as $r_U(\mathcal{E})/r(\mathcal{E}) \longrightarrow 0$ (for $r(\mathcal{E})$ fixed) and increased RC improvement as $r(\mathcal{E}) \longrightarrow 0$ [for $r_U(\mathcal{E})/r(\mathcal{E})$ fixed] will depend on each system individually, our results suggests that any system with coherent errors can benefit from RC on average, even those which are nearly coherence-limited (see Fig. 6.14).

6.5 Randomized Compiling for Fault Tolerance

Qubits in the NISQ era are short-lived and susceptible to a variety of errors and noise due to imperfect control signals and imperfect isolation from the surrounding environment. Therefore, utilizing quantum computers to solve classically-intractable problems (e.g. integer factoring [106]) will likely require quantum error correction (QEC) [107, 108, 109, 110, 111]. QEC can protect logical qubits from errors, but it is only guaranteed to work if the error rate of each physical qubit is below some fault tolerance (FT) threshold [112, 113, 114, 115, 116, 117]. Analytic lower bounds on FT thresholds for various QEC codes have been derived, ranging from $\sim 10^{-6}$ for generic local noise [117] to $\sim 10^{-5}$–10^{-3} for stochastic and depolarizing noise [187, 188, 189, 190, 191]. More optimistic estimates obtained via numerical simulation are orders of magnitude larger than the lower bounds, ranging from $\sim 10^{-3}$–10^{-1} [192, 193, 194, 195, 196, 197, 198, 199, 200], but often assume stochastic (e.g. Pauli, dephasing, or depolarizing) noise models. While recent claims of quantum gates approaching or surpassing FT thresholds boast impressive gate fidelities [102, 201, 202], there is a discrepancy between these claims and the requisite for FT, which requires that the rate of *all* errors fall below a given threshold, not simply the *average* rate at which errors occur.

As discussed in Sec. 5.1, various error metrics and measures exist for quantifying the "error rate" of a quantum gate. Randomized benchmarks typically define error rates in terms of the average gate fidelity, or, equivalently, the process fidelity

$$F = \langle\psi| \left(\mathbb{I} \otimes \mathcal{E}\right)(\rho) |\psi\rangle \,, \tag{6.42}$$

where $\rho = |\psi\rangle\langle\psi|$ is a maximally-entangled state, $\mathcal{E}$ is the error channel associated with some quantum gate, and $\mathbb{I}$ the identity operation. A gate's process *infidelity* $e_F(\mathcal{E}) = 1 - F$ (or *average error rate*) quantifies the probability that the gate induces an error on a random input state, or, equivalently, the average failure rate of random circuits that contain one instance of this gate but that are otherwise perfect. However, FT thresholds are typically defined via the worst-case error rate (also called the diamond norm error rate) [127],

$$\epsilon_\diamond(\mathcal{E}) = \frac{1}{2}\left|\left|\mathcal{E} - \mathbb{I}\right|\right|_\diamond = \frac{1}{2}\sup_\rho \left|\left| \left[\mathbb{I} \otimes (\mathcal{E} - \mathbb{I})\right](\rho)\right|\right|_1, \tag{6.43}$$

where the supremum is taken over all pure states and $\left|\left|X\right|\right|_1 = \mathrm{Tr}\sqrt{X^\dagger X}$ is the trace norm. Operationally, $\epsilon_\diamond(\mathcal{E})$ represents the worst-case performance of a quantum gate in any circuit,

whereas $e_F(\mathcal{E})$ represents the average-case performance for a single instance of the gate. While the diamond norm is a pessimistic estimate of the error rate of a quantum gate, it provides much more rigorous performance guarantees than average error rates. This is because the diamond norm upper bounds the accumulation of error in any quantum circuit, since the distance between the ideal and actual output probability distributions (measured via total variation distance) for any circuit is bounded above by the sum of the worst-case error rates of all its gates [127].

While $e_F(\mathcal{E})$ can be measured directly via randomized benchmarks, there exists no known scalable method for measuring $\epsilon_\diamond(\mathcal{E})$. While tomographic methods such as GST can be used to estimate $\epsilon_\diamond(\mathcal{E})$ [152] by means of a semi-definite program [203], they are exponentially expensive in the number of qubits. However, $\epsilon_\diamond(\mathcal{E})$ and $e_F(\mathcal{E})$ are related via the following bounds [120, 128, 129, 130]

$$e_F(\mathcal{E}) \leq \epsilon_\diamond(\mathcal{E}) \leq \sqrt{e_F(\mathcal{E})}d, \tag{6.44}$$

where $d = 2^n$ (n qubits). The lower bound of $\epsilon_\diamond(\mathcal{E})$ is saturated when $\mathcal{E}$ is a stochastic Pauli channel, and the upper bound, which is quadratically larger in e_F and scales with the dimension d, is saturated by a unitary channel. While modern experimental platforms routinely report single- and two-qubit infidelities on the order of $e_{F,1Q} \lesssim 10^{-4}$ and $e_{F,2Q} \lesssim 10^{-2}$ [7, 60, 201, 202, 204, 205, 206], respectively, if coherent errors account for as little as $\sim$ 1%–10% of this infidelity, the worst-case error rates can be as large as $\epsilon_{\diamond,1Q} \sim \sqrt{e_{F,1Q}} \lesssim 10^{-2}$ and $\epsilon_{\diamond,2Q} \sim \sqrt{e_{F,2Q}} \lesssim 10^{-1}$ [120, 128, 174]. Therefore, $e_F(\mathcal{E})$ and $\epsilon_\diamond(\mathcal{E})$ can differ by orders of magnitude in the presence of coherent errors. This means that randomized benchmarks are inadequate for testing whether gates have achieved FT thresholds [120]. Refs. [102, 201] report benchmarking single- and two-qubit error rates below the FT threshold for the surface code, but base their claims on average error rates from randomized benchmarking (RB) or GST, not worst-case error rates. Similarly, Ref. [207] demonstrates a universal gate set approaching FT thresholds, but only presents gate fidelities not diamond norms. Notably, Ref. [201] includes estimates of the diamond norm measured via GST which are an order of magnitude larger than their reported process infidelities.

While it is not generally possible to directly compare $e_F(\mathcal{E})$ to a FT threshold for QEC, if it can be guaranteed that $e_F(\mathcal{E}) \approx \epsilon_\diamond(\mathcal{E})$ then randomized benchmarks can be used to efficiently verify that gate error rates are below a FT threshold. One method of ensuring that an error budget is dominated by stochastic noise in a quantum algorithm is via RC, which converts all gate errors into stochastic Pauli channels via Pauli twirling. This ensures that the direct measurement of $e_F(\mathcal{E})$ (e.g. via cycle benchmarking [141]) accurately captures the worst-case error rate, which enables comparison to FT thresholds, as well as bounding the overall failure rate of any quantum circuit or application.

In this section, we use GST to study RC performed on two qubits (Q5 and Q6) on our superconducting transmon processor. GST enables measurements of both the process infidelity and diamond norm for all gates in our gate set, allowing us to study the behavior of RC. In Fig. 6.4, we found that in the limit of many randomizations, RC eliminates signatures of coherent errors, enabling one to accurately fit the data with error models containing

only diagonal stochastic Pauli noise. For the purposes of FT, we further show that RC suppresses spatially-correlated coherent errors impacting both qubits and non-Markovian errors. Additionally, we show that the diamond norm converges to the process infidelity under RC, saturating the lower bound of Eq. 6.44, providing strong experimental evidence that our quantum logic operations are close to or below a threshold for fault tolerance. By combining RC with GST, our results provide a novel framework for verifying that FT-required assumptions are satisfied, demonstrating that FT thresholds can be accurately measured using randomized benchmarks as long as quantum circuits are implemented using RC or related randomization methods [143, 157, 178, 179, 181].

Impact of Randomized Compiling on Error Budgets

To explore how effective RC is at converting all errors into stochastic Pauli noise, we compute the fraction of the total error due to coherent errors and stochastic noise, measured via GST, for all gates in $\mathbb{G}$ as a function of N. To do so, we divide each gate's error generator $\mathcal{L}$ into stochastic and Hamiltonian components, and compute the total rate of stochastic (ϵ_{agg}) and Hamiltonian (θ_{agg}) errors as done in Ref. [202]. In Fig. 6.15, we plot the fraction of the *total error* ($\epsilon_{\text{tot}} = \epsilon_{\text{agg}} + \theta_{\text{agg}}$), which is closely related to diamond norm error [202], due to stochastic and coherent (Hamiltonian) errors. We find that the error budget of the simultaneous single-qubit gates is dominated by coherent errors for $N = 0$, and that coherent errors account for approximately two-thirds of the process infidelity of the CZ. For $N = 1$, coherent errors still dominate the single-qubit gates, but the contribution from coherent errors and stochastic noise are both approaching 50% for the CZ gate. By $N = 10$, stochastic noise makes up the largest contribution to the total error for the CZ gate, but we observe only a modest change for the single-qubit gates. However, by $N = 100$ the error budget for all gates is entirely due to stochastic noise. We note that for the $N = 100$ data our model selection chose the S model, which enforces the constraint that the error budget is entirely due to stochastic Pauli noise (see the diagram in Fig. 5.15). This is because there is no statistically significant evidence in the data for any coherent errors (if more data were taken, evidence for some residual coherent errors might be found). These data demonstrate the effectiveness of RC in eliminating the impact of coherent errors. However, this does not mean that coherent errors are physically not present; rather, each individual randomization under RC is impacted by coherent errors in a different manner, in such a way that the aggregate effect of the coherent error is stochastic noise.

We additionally compare our $N = 0$ GST results to results obtained via RB and PB for the CZ gate and simultaneous single-qubit gates. Here, the average error rate from RB represents the total error, and purity RB can be used to measure the unitarity [139]; together, these can be used to compute the fraction of the total error rate due to coherent and stochastic contributions. We find good agreement between the two methods for the CZ gate, but poor agreement for the simultaneous single-qubit gates. These results indicate a discrepancy in different methods for benchmarking the purity of a quantum gate, and that

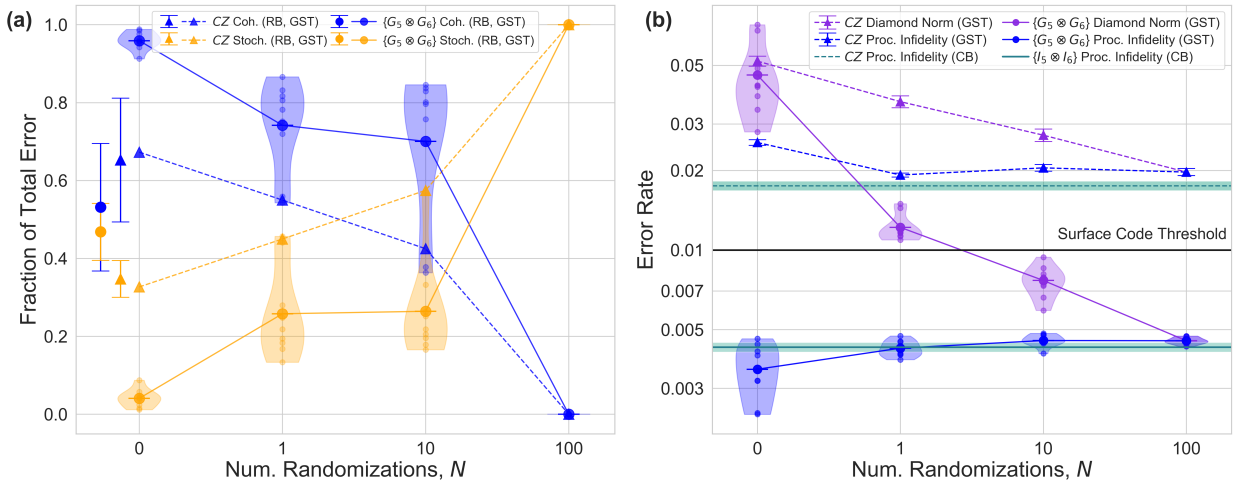


Figure 6.15: **RC for improving estimates of the diamond norm.** (a) Fraction of the total error due to coherent errors [blue] and stochastic noise [orange] as a function of the number of randomizations N under RC. As N increases, stochastic noise makes up a larger fraction of the total error, which is dominated by coherent errors for small N. For (a) and (b), triangular markers denote the CZ gate, and circular markers denote the single-qubit gates $\{G_5 \otimes G_6\}$ (small transparent markers depict the individual gates, large markers depict the averages, and violin plots outline the distribution). We compare the $N = 0$ GST results to estimates obtained by taking the ratio of PB process infidelities to RB process infidelities (markers with error bars to the left of $N = 0$) and find good agreement for the CZ gate, but poor agreement for the simultaneous single-qubit gates. (b) Saturation of the lower bound of the diamond norm under randomized compiling. We plot the GST process infidelity e_F [blue] and diamond norm $\epsilon_\diamond$ [purple] as a function of the number of randomizations N. We observe that $\epsilon_\diamond > e_F$ for all gates for $N = 0, 1, 10$, but that the two are equal for $N = 100$. For visual clarity, we omit the idle cycle for $N = 0$, as $e_{F,II} = 0.0001$ and $\epsilon_{\diamond,II} = 0.001$ fall well below their respective averages. We compare the GST estimates with the process infidelity measured via cycle benchmarking (CB) for the CZ gate [dashed green line] and idle cycle $\{I_5 \otimes I_6\}$ [solid green line]. The GST error bars and CB transparent bands indicate the 95% confidence intervals. Additionally, we compare the benchmarked error rates with the 1% FT threshold for the surface code [black line] and find that the single qubit gates are well below the threshold value, whereas the CZ gate is approaching, but does not surpass the threshold. (Figure reprinted with permission from Ref. [184].)

this discrepancy can depend on the magnitude of coherent errors in the system, and the type of error measure employed.

Process Infidelity vs. Diamond Norm

Error rates measured via randomized benchmarks are generally inadequate for benchmarking FT thresholds because they do not provide any rigorous guarantees on the performance of the gate relative to any given QEC threshold. The diamond norm (Eq. 5.33) is a better metric for comparing error rates to FT thresholds, as it captures the infidelity below which all gate errors can be guaranteed, but it is not trivial to measure experimentally in a scalable manner. If the error model of a quantum processor is dominated by stochastic noise, the average error rate and diamond norm will be equivalent, in which case randomized benchmarks provide the necessary performance guarantees relative to FT thresholds.

In Fig. 6.15, we plot the process infidelity and the diamond norm as a function of N for all gates in our gate set. We find that $\epsilon_\diamond(\mathcal{E})$ converges to $e_F(\mathcal{E})$ as N increases. This is strong experimental evidence that $\epsilon_\diamond(\mathcal{E}) = e_F(\mathcal{E})$ in the many-randomization limit for $N = 100$ (saturating the lower bound of Eq. 5.36). We also compare these results with process infidelities measured independently via CB and find good agreement between the two. These results demonstrate that randomized benchmarks are sufficient for benchmarking FT thresholds if — and only if — a quantum application is impacted only by stochastic noise, which can only be guaranteed if implemented using methods which tailor noise. Similar results were previously reported using PFR for single-qubit gates [157]. We note that since these results were obtained via ensemble measurements, it remains an open question as to whether such results provide any rigorous threshold guarantees for performing QEC on a single-shot basis.

The largest diamond norm error over a gate set is typically the most relevant quantity to compare to a FT threshold. The surface code [208] is a popular QEC code due to its high FT threshold, which is estimated to be between $\sim 0.75\% - 3\%$ [195, 209, 210, 211, 212, 213], with 1% being the threshold that is typically quoted in the literature [102, 201]. In Fig. 6.15, we find that the diamond norm of simultaneous single-qubit gates is below the surface code threshold for $N = 10$ and $N = 100$, and that the error rate of our CZ gate is approaching — but does not surpass — the surface code threshold.

Correlated Errors Under Pauli Twirling

A major requirement for reliable fault-tolerant QEC is the absence of correlated errors, which can occur temporally [214] or spatially [92]. Many-qubit correlated errors cannot be corrected by QEC (each QEC scheme has a maximal weight of error that it can correct), causing logical failures. Furthermore, large-scale correlated errors among many qubits breaks the assumption that only a single error occurs in each block of a quantum error-correcting code, leading to a failure of QEC to prevent catastrophic error propagation. Therefore, the rate of correlated errors must be low to achieve reliable FT quantum computation. For the purposes of FT, we define a correlated error as one which satisfies $m < k \leq n$ for a weight-k operator impacting n qubits during an m-qubit gate. For example, for cycles of single-qubit gates, weight-2+ errors pose a problem. For a two-qubit gate, weight-1 and

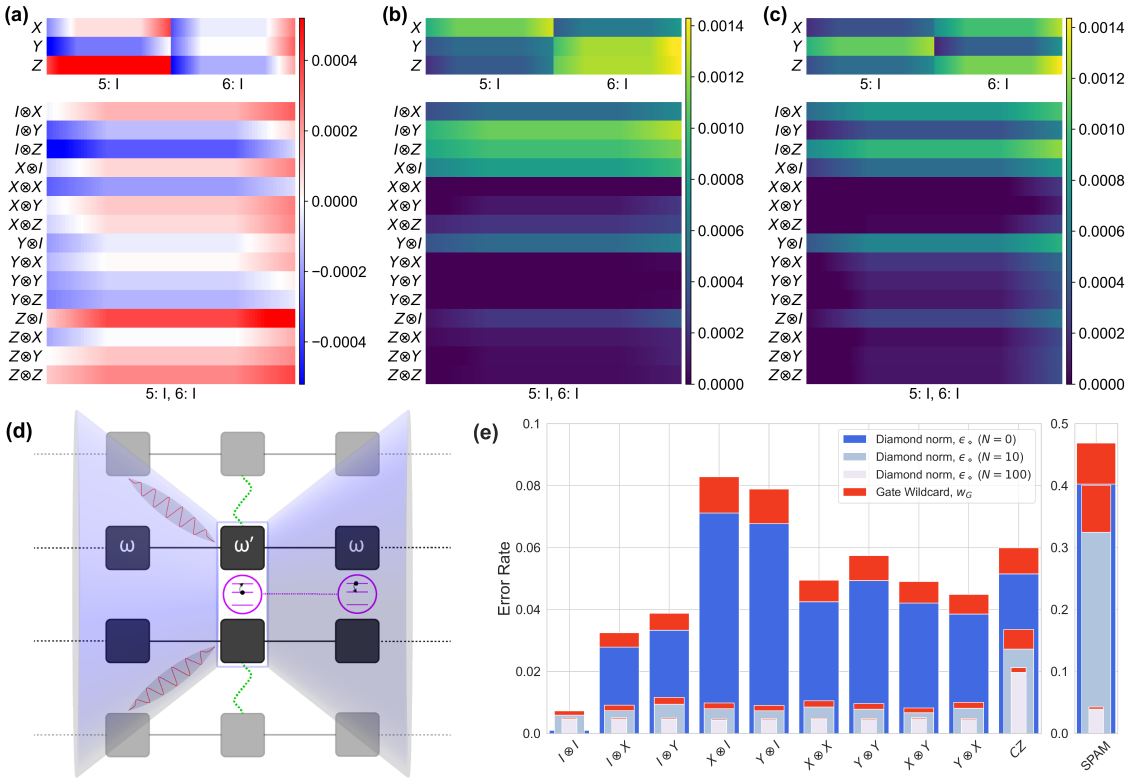


Figure 6.16: **Suppressing correlated and non-Markovian errors.** Heat maps of the weight-1 and weight-2 errors acting on Q5 and Q6 during the idle cycle $\{I_5 \otimes I_6\}$ reconstructed using (a) the Hamiltonian projection of the GST error generator $\mathcal{L}$ for $N = 0$, (b) error rates calculated from the stochastic projection of $\mathcal{L}$ for $N = 100$, and (c) CER. The x-axis labels the target gate, the y-axis labels the Hamiltonian [Pauli Kraus] error for (a) [(b) and (c)], and the cell color [gradient] denotes the over-rotation angle for (a), and error rate for (b) and (c) [95% confidence interval]; the first [second] row of subplots shows marginalized weight-1 [correlated weight-1 and weight-2] errors. While weight-2 errors are dominant for $N = 0$, (b) and (c) show that Pauli twirling suppresses weight-2 errors to negligible levels. (d) Non-Markovian errors in gate-based quantum computing (see Fig. 3.5). (e) Unmodeled error versus the diamond norm. The per-gate wildcard w_G is plotted on top of the diamond norm for each gate in $\mathbb{G}$ (plus SPAM) as a function of N. As N increases, w_G becomes negligible, indicating that non-Markovian errors are suppressed under RC. The x-axis label X (Y) denotes the $X_{\pi/2}$ ($Y_{\pi/2}$) gate. (Figure reprinted with permission from Ref. [184].)

weight-2 errors cannot be distinguished because a single gate-body acting on two qubits can transform one into the other; therefore, correlated errors involving a two-qubit gate must satisfy weight-$k > 2$.

To characterize the extent to which correlated errors are present in our system with and

without Pauli twirling, we extract the weight-1 and weight-2 coherent (stochastic) errors for the idle cycle $\{I_5 \otimes I_6\}$ for the CPTP (S) model for $N = 0$ ($N = 100$), shown in Fig. 6.16; we focus on coherent (stochastic) errors for $N = 0$ ($N = 100$), which is justified by the error budget results in Fig. 6.15a. We observe significant coherent weight-2 errors for $N = 0$, which corresponds to unintended entanglement (i.e. quantum crosstalk), such as static ZZ coupling [170, 171, 173], which can be seen in Fig. 6.16a. In contrast, we observe that weight-1 Pauli errors dominate the error model for $N = 100$, and that weight-2 error are largely suppressed in comparison.

Additionally, we compare the Pauli error rates for $N = 100$ to the Pauli error rates measured via CB and reconstructed using CER, shown in Fig. 6.16c. We find good agreement between the two error maps with respect to correlated errors: both demonstrate that the dominant errors in our system are weight-1 errors, and that weight-2 errors are negligible in comparison, showing that Pauli twirling can suppress correlated errors due to entangling crosstalk and static ZZ coupling between superconducting qubits. (Any discrepancy between GST and CER regarding which weight-1 Pauli errors are most dominant is possibly due to the gauge-dependence of either benchmarking procedure).

In general, one can expect RC to provide a quadratic suppression of correlated coherent errors — RC converts a coherent error that contributes $\mathcal{O}(\theta)$ to the diamond norm (and the total error) into a stochastic error that contributes $\mathcal{O}(\theta^2)$ to the diamond norm. Even though correlated errors might not be *entirely* suppressed by Pauli twirling, as long as there is moderate suppression of large-scale and/or long-range correlated errors, QEC can still be achieved with a reasonable overhead [215].

Non-Markovian Errors

In the context of quantum computing, in particular for characterization and benchmarking, an error process is typically considered to be non-Markovian if it cannot be modeled by a process matrix; in other words, a non-Markovian error is one that occurs from outside of the defined system of qubits and/or outside of the timescale defined by a cycle of gates. We use a cycle as the characteristic timescale for Markovian processes, as this term is derived from the notion of "clock cycles" on a quantum computer, which defines the time within which discrete operations act on qubits (i.e. quantum gates). More precisely, non-Markovian errors are present if each n-qubit cycle (or layer) of gates cannot be modeled by a fixed, context-independent n-qubit process matrix. Various common sources of non-Markovianity in NISQ systems include leakage out of the computational basis states [84, 85, 86, 87, 88, 89], unwanted entangling interactions (e.g. static ZZ coupling) with qubits outside of the studied system, drift in qubit parameters [68] (e.g. stochastic fluctuations in transition frequencies), unwanted coupling to environmental systems with memory beyond the timescale of a cycle (e.g. two-level fluctuators and nonequilibrium quasiparticles [90, 91, 92, 93]), and qubit heating [94]; see Fig. 6.16d. Studying and suppressing non-Markovian errors is important for at least two reasons: they interfere with the quantification of Markovian errors, and their impacts on QEC are less well-understood.

GST is designed to reconstruct all possible Markovian errors on any cycle of quantum gates. So, when GST cannot fit the data, this implies that there are non-Markovian errors. While model violation N_σ is useful for providing evidence of the existence of non-Markovian errors, to quantify the magnitude of such errors, we add a *wildcard error model* [161] to each of our GST models. These wildcard error models assign a wildcard error rate $w_G \in [0, 1]$ to each gate G, and a wildcard error rate to the SPAM $w_{\text{SPAM}} \in [0, 1]$, which quantifies how much additional error on each operation is missing from the model (i.e. is required to make the model consistent with the data). By comparing w_G to $\epsilon_\diamond$ for each gate, we are able to quantify whether Markovian or unmodeled non-Markovian errors dominate the error model [161, 158]. If $w_G \ll \epsilon_\diamond$ for all gates, then Markovian errors dominate and non-Markovian errors are negligible. In this case, the model captures the majority of the errors in the gate (or, more precisely, all those errors that were revealed in this experiment), despite the fact that there is evidence that the model is incomplete. On the other hand, if $w_G \geq \epsilon_\diamond$, the non-Markovian errors dominate the Markovian errors, in which case the GST estimate is unreliable.

Fig. 6.16 shows the wildcard error and the diamond norm error for all of the gates in our gate set, and SPAM, for $N = 0, 10, 100$ (we omit the $N = 1$ data due to its systematic inconsistency with a Markovian error model; see the section describing RC performed on GST sequences). Without RC ($N = 0$), we observe diamond norm errors as large as $\epsilon_\diamond \approx 0.07$, with up to 0.01 additional non-Markovian error. For $N = 10$ randomizations, the wildcard errors are still significant, but they are much smaller in magnitude than with $N = 0$ randomizations *and* they are a small fraction of the diamond norm error rates. By $N = 100$ randomizations, the wildcard errors are negligible — in absolute terms and as a fraction of $\epsilon_\diamond$ — contributing at most 0.0012 additional error per gate. This indicates that the S model accurately captures almost all of the $N = 100$ data (note that in this case the wildcard error quantifies the combined contribution of both non-Markovian errors and all non-S Markovian errors). Because $w_G \ll \epsilon_\diamond \; \forall \; G \in \mathbb{G}$ for all N, we consider all of our models trustworthy, even for $N = 0$. We additionally note that RC significantly improves the both the wildcard error and worst-case error rate for SPAM.

The $N = 100$ data was gathered over a period of over 40 hours, and no re-calibration of gates was performed during the experiment. Therefore, the negligible amount of unmodeled error speaks to the robustness of RC to the inevitable drift in gate and qubit parameters during this time period. While RC was not designed to specifically target non-Markovian errors, its apparent robustness to both non-Markovian errors and correlated errors is promising for future large-scale fault-tolerant applications.

Chapter 7

Error Mitigation for NISQ Applications

The potential exists for noisy intermediate-scale quantum (NISQ) computers to solve complex problems even without quantum error correction. However, using NISQ devices to achieve some kind of quantum advantage will likely require efficient strategies for improving the performance of noisy computations, either via randomization methods for circumventing the impact of different types of errors (e.g. randomized compiling), or via post-processing methods which attempt to generate "noiseless" results from a set of noisy measurements. These methods are broadly termed "error mitigation" (EM) strategies, and in recent years various EM protocols have been proposed and/or developed in an attempt to improve the performance of NISQ computations [149, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228]. EM protocols often require a large number of samples and are therefore not generally scalable in terms of circuit size [229]. However, their relatively small overhead in terms of qubits and gates makes them suitable for NISQ applications, which are already limited in circuit depth to begin with.

In this chapter, we introduce several different error mitigation protocols that are uniquely enabled by randomized compiling (RC), and show how these can improve the performance of various different quantum applications. We additionally comment on their efficiency and how they scale to larger system sizes. Finally, we introduce a new randomization method — equivalent circuit averaging — that is inspired by randomized compiling, but is fundamentally different in its execution.

7.1 Leveraging Randomized Compiling for the QITE Algorithm

Many common NISQ applications fall under the category of variational quantum algorithms (VQAs) [27], by which some variational parameter(s) in a quantum circuit is updated in a quantum-classical hybrid loop by means of classical optimization. Many VQAs, such as the

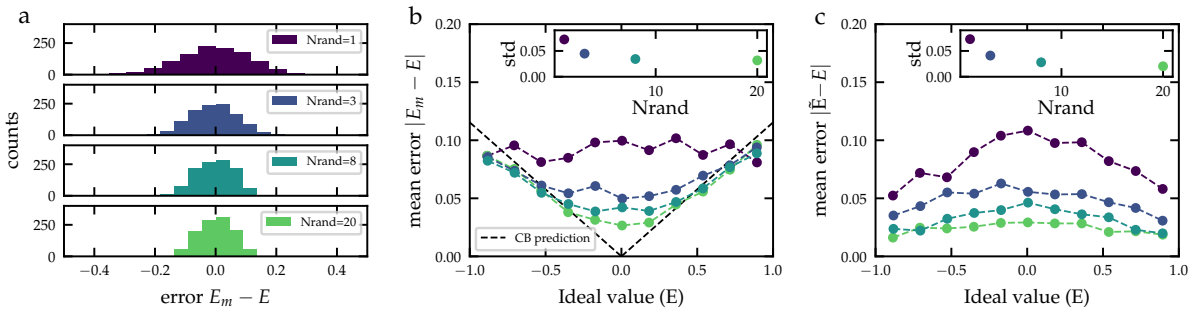


Figure 7.1: **Effect of RC on measured expectation values.** (a) The variance of the error in measured expectation values is reduced as the number of randomizations N_{rand} under RC is increased. (b) Mean error in measured expectation values as a function of the ideal expectation value E for all random circuits. The mean error approximates a depolarizing noise channel as N_{rand} increases. The dashed line is an estimate of the depolarizing parameter $f = 0.98$ measured via CB. (c) Mean error in the renormalized expectation values using the purification formula in Eq. 7.1. For (b) and (c), standard deviations for the data for different N_{rand} are plotted in the inset. (Figure reprinted with permission from Ref. [236].)

variational quantum eigensolver (VQE) [230, 231], are robust to coherent errors due to the fact that the optimization step can "learn" and account for calibration errors in gates or crosstalk-induced coherent errors. One example of a VQA which is *not* robust to coherent errors is the Quantum Imaginary Time Evolution (QITE) [232, 233, 234, 235] algorithm, which is an iterative algorithm that approximates imaginary time evolution with a unitary operation; in the limit of infinite imaginary time, the algorithm converges to the ground state of a given Hamiltonian. One advantage of QITE is that it does not require *a priori* knowledge of an ansatz, as the circuit used in each iteration is different and depends on the results of the previous iteration. However, as previously mentioned, QITE is sensitive to coherent errors, making it an optimal example of a NISQ algorithm whose performance can be boosted by RC. We perform QITE on three qubits (Q4, Q5, and Q6) and solve the ground and first-excited state of transverse-field Ising model (TFIM). We utilize RC to simplify the errors affecting the computed expectation values of the QITE algorithm, and further improve their values by means of an RC-enabled purification method. The work presented here is based on Ref. [236].

Noise Characterization of Random Circuits under RC

In the previous chapter, we demonstrated the impact of RC on random circuits of variable depth in terms of the total variation distance (TVD). Here, we instead explore the impact of RC on the expectation values of random circuits. To do so, we generate and measure random two-qubit circuits of depth 6 CZ gates and compute the expectation values E_m of all the

possible Pauli strings composed of Z and I. In Figure 7.1a, we plot distributions of the error in the measured expectation values ($E_m - E$, where E denotes the true value) as a function of the number of randomizations N under RC. We see that as N increases, the variance in $|E_m - E|$ decreases. In effect, by sampling from a greater number of randomizations of each bare circuit (but keeping the total number of shots the same), we are purposefully undersampling the bias of each circuit in order to reconstruct more precise, bias-free expectation values.

In Figure 7.1a, we plot the mean error in the measured expectation values as a function of the ideal value E. We find that with a single randomization, the mean error is essentially flat across E. However, as N is increased, $|E_m - E|$ obtains a linear dependence on E. We observe that for $N = 20$, the mean error appears to approximate a 'V' shape. Under a fully depolarizing noise model, we can write $E_m = fE$, where f is the depolarizing constant (Eq. 4.9), or, more generally, the effective length of the generalized Bloch vector. Therefore, it stands to reason that under RC, the errors impacting our circuits approximate a fully depolarizing noise model. This is rigorously not correctly, since RC only tailors errors into stochastic Pauli channels, but we will see that this approximation is sufficient to perform error mitigation on QITE circuits.

Estimating the Depolarizing Parameter f

Errors under a fully depolarizing noise model can be mitigated by simple purification. Knowledge of the depolarizing parameter f is sufficient to recover ideal expectation values from noisy estimations by re-scaling all measured expectation values by this same coefficient f [237, 238]. A direct way of estimating f is to perform full state tomography at the end of a sequence to compute the purity of the final state, and thus extract the length of the generalized Bloch vector. In this case, the purification is given by

$$\tilde{E} = \frac{E_m}{f} \quad \text{with} \quad f^2 = \frac{1}{2^n - 1} \sum_{P \in \mathbb{P}_n \setminus \{I^{\otimes n}\}} E_P^2, \tag{7.1}$$

where E_P denotes the expectation value of the Pauli operator P and $\mathbb{P}_n$ is the n-qubit Pauli group, but we omit the n-qubit identity when computing f. In Fig. 7.1c, we re-normalize all expectation values by their respective depolarizing parameters and compute the mean error as a function of the ideal expectation value. We see that as the number of randomizations N is increased, the mean error in the renormalized expectation values becomes flat across E, indicating that RC better approximates a depolarizing error model in the limit of large N.

Limitations of Purification

Full state tomography is, however, inefficient for determining f, as it is exponentially expensive in the number of qubits. An alternate, more scalable strategy for estimating f is via cycle benchmarking (CB), which measures the Pauli fidelities f_P for some subset of all

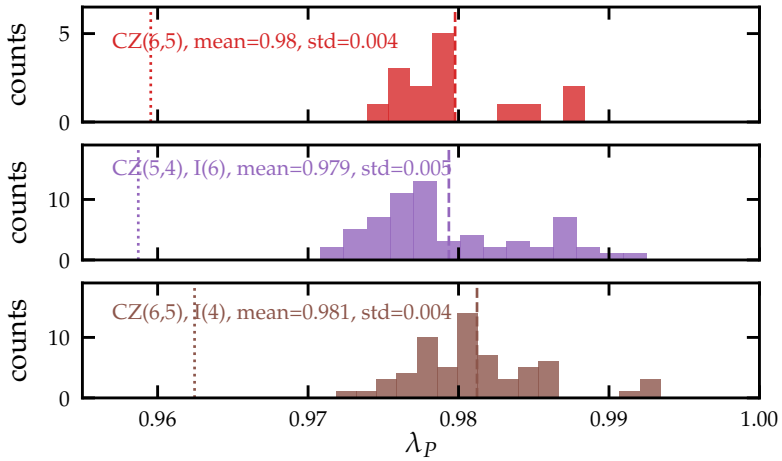


Figure 7.2: **Distribution of Pauli fidelities under cycle benchmarking.** Histograms of Pauli fidelities f_P measured via CB for different two-qubit gate cycles (note that λ_P used in Ref. [236] is equivalent to f_P used here): (top) CZ between Q5 and Q6, (middle) CZ between Q4 and Q5 with Q6 idling, and (bottom) CZ between Q5 and Q6 with Q4 idling. The dashed line is the mean value and the dotted line is the lower bound of all possible values (see Supplement of Ref. [141]). (Figure reprinted with permission from Ref. [236].)

of the Paulis $P \in \mathbb{P}_n$ (see Eq. 5.57). From this, the process fidelity (Eq. 5.60) can be computed, from which the depolarizing parameter can be derived by multiplication of a simple dimensionality factor (see Table 5.1). In Fig. 7.1b, we plot f estimated via CB performed on the native two-qubit CZ gate used in the random two-qubit circuits, which shows good agreement with the data as N increases, approximating a depolarizing noise model. Note that because CB measured the process fidelity of a dressed cycle, the depolarizing parameter f extracted via CB also estimates the depolarization of the Bloch vector due to single-qubit gates; therefore, it is not necessary to measure this separately. However, in general it *is* necessary to benchmark each distinct two-qubit gate cycle under CB separately. For circuits containing K distinct two-qubit gate cycles, the effective depolarizing parameter for the entire circuit can be estimated as

$$f_{\text{eff}} = \prod_{i=1}^{K} f_i^{N_i}, \tag{7.2}$$

where N_i denotes the number of times the ith cycle appears in the circuit.

The utility of using CB for estimating f is that the variance in the fidelities of the various Pauli channels gives a quantitative measure of the uncertainty in our estimate. In Fig. 7.2 we plot the various Pauli fidelities for different two-qubit gate cycles measure via CB. We see that the performance of the cycles is different under different Pauli channels, and that

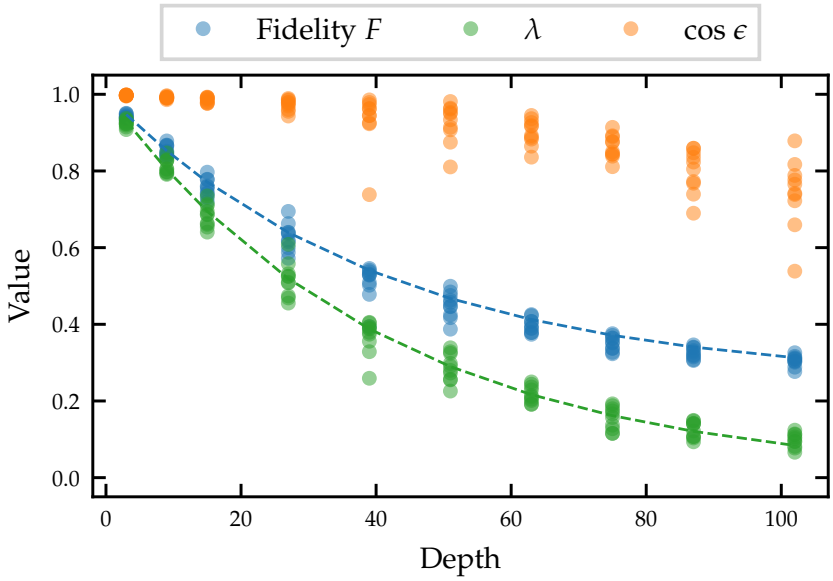


Figure 7.3: **Contributions to fidelity as a function of circuit depth.** Fidelity F (blue points), Bloch vector length f (green points; λ used in Ref. [236] is equivalent to f used here), and angle error $\cos\epsilon$ (orange points) versus circuit depth. The length error agrees well with the fidelity for shorter circuit depths, indicating that depolarizing noise is the dominant source of error. The angle error contributes a larger fraction of the total error at longer circuit depths, suggesting that the assumption of depolarizing noise breaks down for large m. Estimates of F and f measured via CB (dashed lines) agree well with the state tomography results. (Figure reprinted with permission from Ref. [236].)

the spread gets larger as the number of qubits in the cycle increases from two to three. This poses a problem for purification using CB, as it is not clear which Pauli channel(s) will be dominant in any given circuit, even if the circuit is performed with RC. Therefore, in some instances, purification of an expectation value using the mean CB value can produce non-physical results, in which $\tilde{E} > 1$; we discuss this further below in the context of the QITE algorithm.

Another limitation of using CB for purification is that the measured process fidelity of any given cycle will include contributions from both coherent errors and stochastic noise. To see this, we can express the fidelity of a measured state ρ with a pure state σ in terms of

their Pauli expectation values:

$$F(\rho,\sigma) = \mathrm{tr}(\rho\sigma) \tag{7.3}$$

$$= \frac{1}{2^n}\Big(1 + \sum_{P\neq I^{\otimes n}} \rho_P \sigma_P\Big) \tag{7.4}$$

$$= \frac{1}{2^n} + \left(1 - \frac{1}{2^n}\right) f\cos\epsilon \tag{7.5}$$

where f is the length of the Bloch vector given by Eq. 7.1 and ϵ is the angle between the two generalized Bloch vectors. Equation 7.5 indicates that two mechanisms can decrease the fidelity: a reduction of the Bloch vector length, given by f, and an angle error between the two vectors, contributing $\cos\epsilon$. Purification or re-scaling is intended to correct the first type of error, but does not correct the angle error. Assuming that depolarization is the limiting error, the fidelity can be approximated as

$$F(\rho,\sigma) \simeq \frac{1}{2^n} + \left(1 - \frac{1}{2^n}\right) f. \tag{7.6}$$

For a circuit with the same two-qubit gate cycle repeated m times and a depolarizing parameter f from CB, we can simplify the equation for the fidelity:

$$F(\rho,\sigma,m) \simeq \frac{1}{2^n} + \left(1 - \frac{1}{2^n}\right) f^m. \tag{7.7}$$

The residual coherent error contribution is given as

$$\cos\epsilon = \frac{1}{2^n - 1} \sum_{P\neq I^{\otimes n}} \frac{\rho_P}{f}\sigma_P, \tag{7.8}$$

which is assumed to be small compared to the length error.

To test whether the angle errors are negligible compared to the length errors under RC, we generated 10 different random circuits at 10 different circuit depths m, and computed depolarizing parameter f, fidelity F (Eq. 7.6), and residual angle error $\cos\epsilon$ (Eq. 7.8) using state tomography, and compare these values with the equivalent estimates obtained via CB; see Fig. 7.3. We observe that the reduction in fidelity at shorter circuit depths is mostly due to a reduction in the length of the Bloch vector, indicating that the assumption of depolarizing noise is valid for small m. However, we observe that there is a large deviation in F and f for longer circuit depths due to the angle error contributing a larger fraction of the total error, suggesting that the assumption of depolarizing noise breaks down in the limit of large m. This is likely due to the fact that there is incomplete cancellation of coherent errors from RC for longer circuits depth, with coherent errors slowly building up with m. This is coherent build up will likely be worse for structured quantum circuits, since random circuits themselves are already somewhat robust to the coherent accumulation of unitary errors.

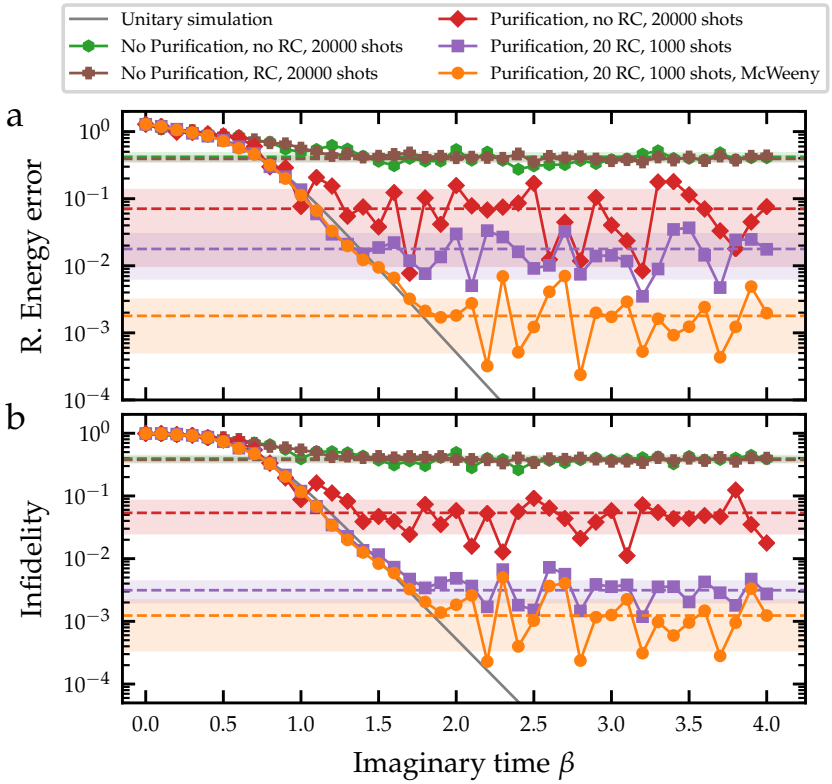


Figure 7.4: **Quantum Imaginary Time Evolution.** Convergence of QITE for solving the ground state energy of the three-site TFIM (with $J = h = 1$) for various different levels of error mitigation, outlined in the legend. The dashed line (colored region) denotes the mean (standard deviation) of the last 10 points. Both the relative energy error and the ground state infidelity converge to a value approaching $\mathcal{O}(10^{-3})$. (Figure reprinted with permission from Ref. [236].)

Application to the QITE Algorithm

Imaginary time evolution is a well-known classical method in special relativity, statistical mechanics, and quantum field theory. The key ingredient of this algorithm is that the imaginary time propagator $U(\beta) = \exp(-\beta\mathcal{H})$ — which is non-unitary — will converge in the limit of large imaginary time, where $\beta = it$. For solving energy levels in quantum systems, the key convergence criteria for imaginary time evolution is that the initial state of the system must overlap with the ground state of the Hamiltonian [239, 240]. In Ref. [234], the authors describe how to use a quantum computer to perform the imaginary time evolution on NISQ hardware. The core idea is to normalize the evolution operator in Trotterized time steps, making it a unitary operator that can then be decomposed in terms of quantum gates. This can be done efficiently by solving a linear system using classical computers, and thus

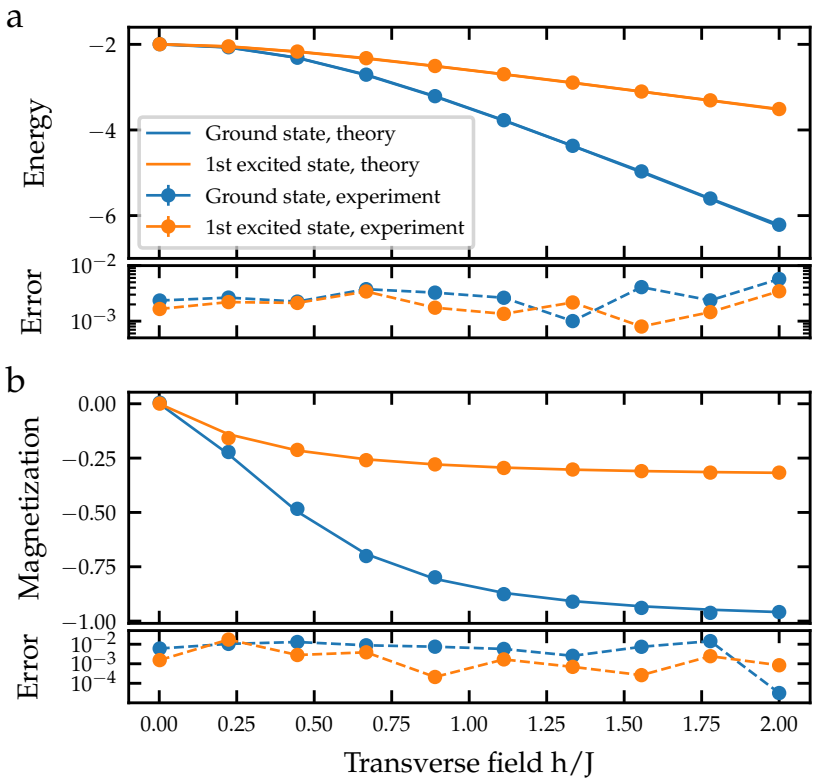


Figure 7.5: **TFIM results.** Phase diagram for the three-site TFIM as we vary h relative to J. (a) Energy in units of E/J for the ground state and first excited state. The reported energy corresponds to the mean of the five last points of the QITE evolution. Error bars for these data are smaller than the markers. The error shown for each point corresponds to the relative error. (b) Measured magnetization of the same data, with the corresponding absolute error. (Figure reprinted with permission from Ref. [236].)

QITE is free of the complex optimizations that arise in VQE [241]. The cost of QITE is that it is not a fixed-depth algorithm, and the number of gates needed in the quantum circuit increases for each time step.

Here, we focus on using QITE to solve the Transverse Field Ising Model (TFIM), given by the following Hamiltonian:

$$\mathcal{H} = J \sum_{\langle ij \rangle} X_i X_j + h \sum_i Z_i, \tag{7.9}$$

where J is the interaction exchange between the nearest neighbors, h is the transverse field applied to the chain, and $\langle ij \rangle$ indicates that the sum is over nearest neighbors. The TFIM is a well-known model which has been previously solved using QITE and other algorithms [234, 235, 242, 243]. The state and the evolution operator at a given step can be expanded

in the Pauli basis:

$$\rho = \sum_{P \in \mathbb{P}_n} \rho_P P \quad \text{and} \quad U = \exp\left(-i \sum_{P \in \mathbb{P}_n} a_P P\right), \tag{7.10}$$

where ρ_P are the expansion coefficients of ρ in each Pauli basis P, and a_P are the generators of the unitary U. Ref. [234] describes how to solve for the generators a_P for each time step using the results of the previous time step.

To demonstrate the error mitigation strategy presented here, we solve for the ground state of the three-site TFIM for $J = h = 1$, and calculate the relative energy error and infidelity of the measured ground state as a function of imaginary time β, shown in Fig. 7.4. We compare the performance of QITE with and without RC, and with and without purification. We empirically find that when using RC + purification, our estimate of the ground state energy level will sometimes overshoot the true value, which is non-physical with a variational ansatz, since the true ground state provides a lower bound on the estimated ground state for any trial wavefunction that is not the true ground state wavefunction. Therefore, to enforce physicality in our estimates, we perform McWeeny purification on our reconstructed density matrix ρ, which projects ρ to the closest state in which the purity $\gamma = 1$. (More formally, McWeeny purification enforces ρ to be idempotent, such that $\rho = \rho^2$; see Sec. 3.1.) We find that the combination of RC + Purification + McWeeny converges most rapidly to the ground state, producing a relatively energy error and ground state infidelity of $\mathcal{O}(10^{-3})$.

The utility of the QITE algorithm is that it can also solve for higher energy levels, as well as other interesting properties of the system, such as the local magnetization. In Fig. 7.5, we plot a phase diagram of first two energy levels of the three-site TFIM and their magnetization as we sweep over the transverse magnetic field parameter h relative to the coupling parameter J. We consistently get errors below 1% for all of these quantities using the RC + Purification + McWeeny method.

We have demonstrated how iterative VQAs, such as QITE, can benefit from the application of both a noise tailoring via RC, and error mitigation via purification. The application of RC and purification results in an improvement over each technique used separately, but comes with the cost of needing to estimate the depolarizing parameter either via state tomography (which is not scalable) or CB (which is scalable, but necessitates benchmarking each distinct two-qubit gate cycle that appears in a circuit). Such error mitigation strategies can improve the ability of NISQ systems to perform useful computations, even without quantum error correction.

7.2 Noiseless Output Extrapolation and Pauli Error Cancellation

Unlike quantum error correction, error mitigation does not attempt to correct errors, but rather characterize them in a manner in which they can be removed in post-processing.

	PEC	NOX	Unmit.
Runtime	$\frac{(1-n\varepsilon)^{-2m}}{\sigma^2}$	$\frac{m^3}{\sigma^2}$	$\frac{1}{\sigma^2}$
Bias	$\delta_{\text{PEC}} + \delta_{\text{rec}}$, with $\delta_{\text{PEC}} = O(mn^2\varepsilon^2)$	$\delta_{\text{NOX}} + \delta_{\text{rec}}$, with $\delta_{\text{NOX}} = O(m^2n^2\varepsilon^2)$	$O(mn\varepsilon)$

Table 7.1: **Run time and bias for NOX and PEC in comparison to an unmitigated implementation of the target circuit.** σ is the desired standard deviation of the results, m the circuit depth, and $n\varepsilon$ the error rate of each cycle, which for simplicity we take to be the same every cycle, with n the number of qubits and ε a constant (Ref. [149] provides a generalization of this assumption). The quantity δ_{rec} depends on the accuracy of the noise reconstruction (e.g. $\delta_{\text{rec}} = 0$ if the noise is known exactly).

Several different categories of EM protocols include (but are not limited to) methods for intentionally amplifying errors in order to extrapolate to a zero-noise limit [149, 216, 219, 220, 222, 228, 244], methods for characterizing errors and subsequently inverting them via quasi-probabilities [149, 217, 245], and post-selection protocols which eliminate wrong outputs, for example, by checking them against some expected symmetry [246, 247]. Here, we introduce two novel EM strategies — Noiseless Output Extrapolation (NOX) and Pauli Error Cancellation (PEC) — which fall into the first two categories, respectively. Both protocols take as input a quantum circuit and an operator O, and return an estimator of the expectation value of O for a noiseless implementation of the input circuit. As we will see, PEC is a quasi-probabilistic protocol that is made scalable through the use of cycle error reconstruction (CER) and randomized compiling (RC). On the other hand, while NOX does not implicitly require RC, RC does help satisfy certain assumption we make about the noise impacting hard gate cycles, and we empirically find that its performance is greatly improved by RC and therefore build it into the protocol. In what follows, we do not give a formal derivation of the protocols; rather, we focus on a heuristic description of how they work, and showcase recent results obtained on our processor. The work presented here is based on Ref. [149]; formal derivations of the protocols can be found therein.

NOX

NOX relies on the ability to amplify the errors afflicting individual noisy cycles in a circuit $\mathcal{C}$. Specifically, it requires replacing a noisy cycle of quantum gates $\mathcal{G}$ with $\mathcal{G}^k$ for some

integer $k > 1$, where $\mathcal{G} = \tilde{\mathcal{G}}\mathcal{G}_0$, with $\mathcal{G}_0$ the ideal cycle of gates and $\tilde{\mathcal{G}}$ the noise acting on the cycle. Therefore the amplification of $\mathcal{G}$ also amplifies $\tilde{\mathcal{G}}$. We take $\tilde{\mathcal{G}}$ to be a stochastic Pauli channel, as is the case under RC.

The traditional method for amplifying noise is "Identity Insertion" [222], in which a noisy cycle $\mathcal{G}$ is replaced with $\mathcal{G}(\mathcal{G}\mathcal{G}^{-1})^k$. This method is efficient in its construction and straightforward to implement, but it is accurate only if two conditions are satisfied: (1) $\mathcal{G}$ and $\mathcal{G}^{-1}$ must have identical errors, which is only trivially satisfied if $\mathcal{G} = \mathcal{G}^{-1}$ (e.g. CX or CZ gates). And (2), the noise acting on the cycle $\tilde{\mathcal{G}}$ must commute with the ideal cycle $\mathcal{G}_0$: $\tilde{\mathcal{G}}\mathcal{G}_0 = \mathcal{G}_0\tilde{\mathcal{G}}$; this is only trivially satisfied under a depolarizing noise channel. Ref. [149] also introduces a modification of Identity Insertion, which they term "Append Errors," that can be utilized for more precise amplification if either of the two conditions are violated. However, in our experiments we found this to be unnecessary, and therefore do not discuss it further.

The NOX protocol works as follows (formally described in the supplemental material of Ref. [149]): NOX takes as input the circuit $\mathcal{C}$, an n-qubit state ρ_0, an operator O such that the spectral norm $||O||_\infty \sim 1$, a number $\sigma \in (0, 1)$ representing the desired uncertainty of the results, an integer $k > 1$, and a Boolean $\mathsf{id\_insert} \in \{\mathsf{True}, \mathsf{False}\}$. It requires running $m+1$ circuits in total. The first circuit is the original circuit $\mathcal{C}$, and the remaining m circuits are identical to $\mathcal{C}$, except that the jth noisy cycle $\mathcal{G}_j$ has been replaced by $\mathcal{G}_j(\mathcal{G}_j\mathcal{G}_j^{-1})^k$. If $\mathsf{id\_insert} = \mathsf{True}$, the noise amplification is performed with Identity Insertion, otherwise it is performed with Append Errors. Each circuit is measured $m^2/(k-1)^2\sigma^2$ times and yields a noiseless estimation of $E(O)$, where $E(O)$ is the estimator of $\mathrm{Tr}[O\mathcal{C}(\rho_0)]$. We denote $\widetilde{E}_0(O)$ the noisy estimator of the original circuit $\mathcal{C}$, and $\widetilde{E}_{\mathcal{G}_j,k}(O)$ the noisy estimator for the circuit in which the jth noisy cycle has been amplified. After running all the $m+1$ circuits, NOX returns the estimator

$$\widehat{E}_{\mathrm{NOX}}(O) = \widetilde{E}_0(O) + \sum_{j=1}^{m} \frac{\widetilde{E}_0(O) - \widetilde{E}_{\mathcal{G}_j,k}(O)}{k-1}. \tag{7.11}$$

If successful, this quantity is an estimator of $E(O)$ that is more accurate than $E(O)$ itself. In addition to the standard protocol, one can also implement the protocol with RC, in which case each of the m additional circuits is implemented with N different randomizations under RC.

NOX differs from other typical noise-amplification EM protocols in that it amplifies the errors across an entire cycle of quantum gates, rather than the errors affecting individual quantum gates within a circuit, and therefore can address non-trivial processes such as non-local correlated errors. This is important because amplifying errors on individual quantum gates can have unintended consequences on nearby idle qubits, which are not necessarily correctable in the same post-processing step. As outlined in Table 7.1, to achieve an uncertainty of $\mathcal{O}(\sigma)$, NOX requires running $m+1$ circuits and measuring each of them $\mathcal{O}(m^2)$ times. Thus, NOX has runtime $\mathcal{O}(m^3)$ and is efficient in m. Note that the bias of NOX grows quadratically with m.

PEC

PEC is based on quasi-probabilistic error cancellation, in which an unbiased estimator is constructed by sampling from a distribution of biased estimators [216, 217]. The goal of quasi-probabilistic error cancellation is to calculate a set of probabilities p_i, a set of signs $s_i \in \{-1, +1\}$ and a cost $C_{\text{tot}} > 0$, such that for any state ρ_0 and operator O,

$$C_{\text{tot}} \sum_{i=1}^{L} s_i p_i \operatorname{Tr}\left[O\widetilde{\mathcal{U}}_i(\rho_0)\right] = \operatorname{Tr}\left[O\mathcal{U}(\rho_0)\right] + \delta, \tag{7.12}$$

where $\mathcal{U}$ is a noiseless operation, $\{\widetilde{\mathcal{U}}_i\}_{i=1}^{L}$ a set of noisy operations, and δ a residual bias that captures the effectiveness of the EM protocol. All EM protocols based on quasi-probabilistic error cancellation guarantee a negligible bias $\delta \approx 0$, provided that the noisy maps $\widetilde{\mathcal{U}}_i$ can be accurately characterised.

As we can see, the success of PEC relies on the accurate characterization of $\widetilde{\mathcal{U}}_i$. Related works on quasi-probabilistic methods have used Gate set tomography (GST) to reconstruct the noise afflicting $\widetilde{\mathcal{U}}_i$ [217, 218]. However, GST is inefficient in the number of qubits, and therefore the error reconstruction is done on individual single- and two-qubit gates, not cycles containing both. This approach immediately limits the utility of quasi-probabilistic error cancellation for larger circuits, as crosstalk-induced coherent errors typically play a major role in multi-qubit circuits [145, 147, 248]. As a result, EM protocols that depend on GST have only been demonstrated on circuits containing up to two qubits [182, 245]. Instead of GST, we utilize cycle error reconstruction (CER) to characterize all $\widetilde{\mathcal{U}}_i$. CER is both scalable in the number of qubits n[1] and provides an accurate characterization of Pauli-twirled errors under RC [145]. There, our method — PEC — makes quasi-probabilistic protocols uniquely scalable due to RC.

The PEC protocol works as follows (formally described in the supplemental material of Ref. [149]): PEC takes as input the circuit $\mathcal{C}$, the Pauli error rates $\{p^{(\mathcal{G}_j)}\}$ of all the noisy cycles $\mathcal{G}_j$ in $\mathcal{C}$ (which are computed in advance with CER), an n-qubit input state ρ_0, an operator O such that $||O||_\infty \sim 1$, and a number $\sigma \in (0, 1)$ representing the desired uncertainty of the result. It uses quasi-probabilistic error cancellation to suppress the noise afflicting the noisy cycles $\mathcal{G}_j$, and eventually it returns an estimator $\widehat{E}_{\text{PEC}}(O)$ of $\operatorname{Tr}\left[O\mathcal{C}(\rho_0)\right]$. To calculate $\widehat{E}_{\text{PEC}}(O)$, PEC requires running $N = (C_{\text{tot}}/\sigma)^2$ circuits in total, with a cost given by

$$C_{\text{tot}} = \prod_{j=1}^{m} \frac{1}{\left(p_0^{(\mathcal{G}_j)}\right)^2 - \sum_{k=1}^{4^n} \left(p_k^{(\mathcal{G}_j)}\right)^2}. \tag{7.13}$$

[1] As discussed in Sec. 5.4, CER is only scalable in the limit in which one wishes to characterize all k-body errors, where $k < n$, since measuring all n-body errors is exponentially expensive. For $k < n$, CER is only polynomially expensive in n. Typically, $k = 2$ is assumed to be sufficient, which is justified (and subsequently verified) under the assumption that Pauli twirling suppresses spatially-correlated coherent errors.

Each of these circuits is obtained by appending randomly chosen Pauli gates to the noiseless cycles. Specifically, every circuit in PEC implements an operation of the type

$$\mathcal{C}^{(\mathrm{PEC})}(P_1,\ldots,P_m) = \mathcal{R}_{m+1}\big(\circ_{j=1}^{m} P_j\mathcal{G}_j\mathcal{R}_j\big), \tag{7.14}$$

where $P_j \in \{I,X,Y,Z\}^{\otimes n}$ is chosen at random with probability $p_{k_j}^{(\mathcal{G}_j)}$, and $\mathcal{R}_j$ denotes the jth cycle of single-qubit gates. Together with the N circuits, PEC also initializes a list of signs $s_1,\ldots,s_N$, where $s_k = 1$ if circuit k contains an even number of random Pauli cycles P_j that are different from the identity, and $s_k = -1$ otherwise. After initializing the circuits and signs, PEC applies RC to every circuit, runs the circuits and stores the results $r_1,\ldots,r_N$. Finally, it computes the estimator $\widehat{E}_{\mathrm{PEC}}(O)$ as

$$\widehat{E}_{\mathrm{PEC}}(O) = \frac{C_{\mathrm{tot}}}{N}\sum_{k=1}^{N} s_k r_k. \tag{7.15}$$

As outlined in Table 7.1, to achieve an uncertainty of $\mathcal{O}(\sigma)$, PEC requires implementing $C_{\mathrm{tot}}^2/\sigma^2$ circuits, where C_{tot} scales exponentially with m. For example, when all the noisy cycles have the same error rate ε, we have $C_{\mathrm{tot}} = \mathcal{O}[(1-\varepsilon)^{-2m}]$. Thus, in general PEC (as well as all other protocols based on quasi-probabilistic cancellation [217]) is exponentially inefficient with circuit depth. Nevertheless, as long as the depth $m \lesssim \varepsilon^{-2}$, PEC remains efficient, and therefore is a practical EM protocol for near-term algorithms.

W-state Circuits

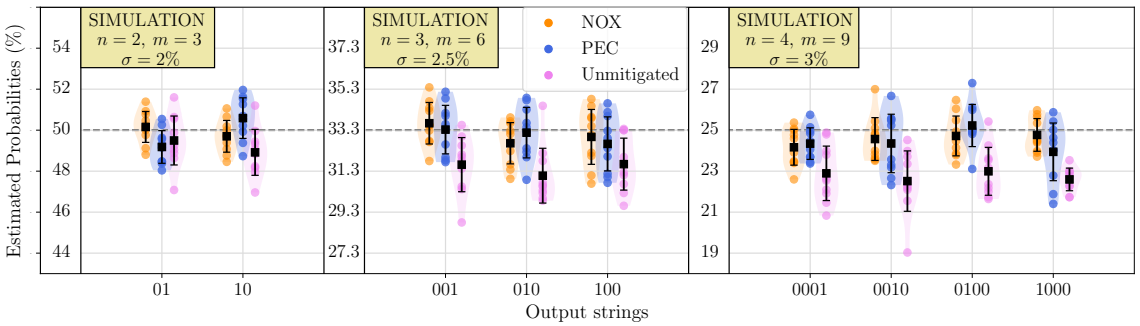
To demonstrate the broad applicability of our EM protocols, we test their performance on several different types of quantum circuits. Our first experiment is on W-state circuits, which generate a special type of multipartite entanglement, which can be written as an equal superposition of all weight-1 basis states. An n-qubit W state is given as

$$|W_n\rangle = \frac{|0\ldots01\rangle + |0\ldots10\rangle + |1\ldots00\rangle}{\sqrt{n}}. \tag{7.16}$$

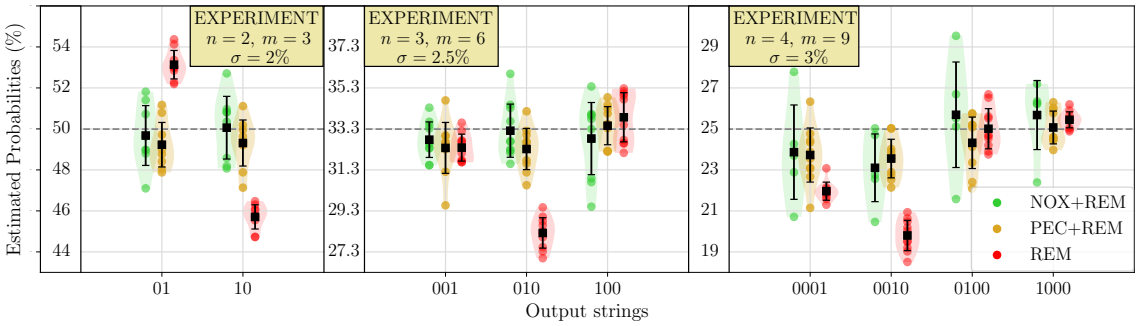
When decomposed into our linear topology containing only two-qubit entangling CZ gates, an n-qubit W-state circuit contains $3(n-1)$ two-qubit gate cycles. We generate and measure W-state circuits on $n = 2,3,4$ qubits, and compare the error mitigated results with the unmitigated results, as well as with simulated results using the measured CER data (see Fig. 5.12). We also apply standard readout error mitigation (REM) [249] to both the EM and unmitigated results to correct basic readout and misclassification errors on our raw outputs.

In Fig. 7.6 we plot the W-state results in terms of the most frequent output bit strings and the total variation distance (TVD)

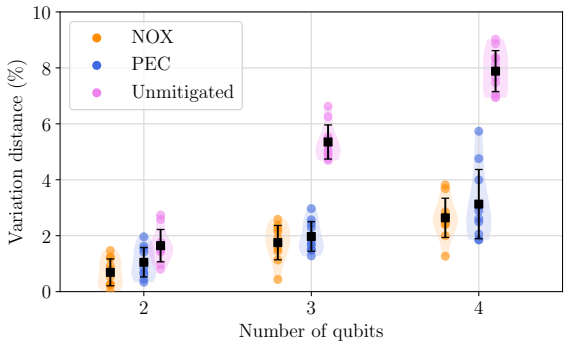
$$D(p,q) = \frac{1}{2}\sum_{x\in X}|p_x - q_x|, \tag{7.17}$$



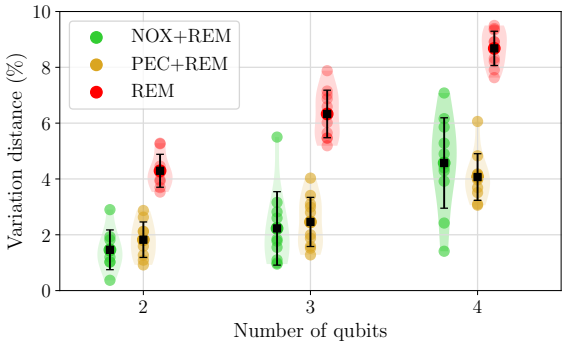
(a) Most frequent bit strings (simulation)



(b) Most frequent bit strings (experiment)



(c) Total variation distance (simulation)



(d) Total variation distance (experiment)

Figure 7.6: **W-state circuit results.** (a) Numerical simulations and (b) experimental results for the most frequent bit strings for $n = 2, 3, 4$ qubits. (c) Numerical simulations and (d) experimental results for the total variation distance between the ideal and noisy outputs. m denotes the circuit depth in terms of two-qubit gates, and σ denotes the intended uncertainty in the results. Data points correspond to the results of the individual circuits, the squares represent their means, the bars represent the standard deviation uncertainty, and the dashed lines highlight the ideal values for (a) and (b). (Figure reprinted with permission from Ref. [149].)

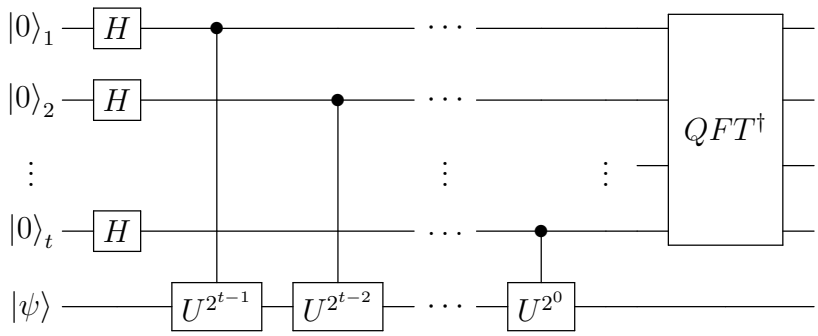


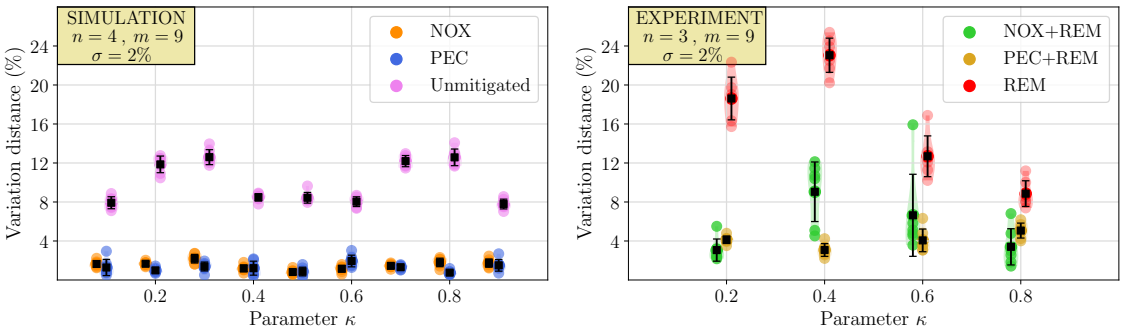
Figure 7.7: **Quantum Phase Estimation algorithm.** We set $U = R_Z(\kappa)$ for different κ, where $R_Z(\kappa)$ is defined in Eq. 7.18.

between the ideal q_x and noisy p_x probability distributions, where x is a bit string in a set X. We find that NOX and PEC (+REM) produce the correct bit string results within a statistical uncertainty for almost all of the ideal outputs, the same of which is not true for the unmitigated results (+REM), and that both EM protocols significantly reduce the TVD compared to REM alone, with average improvements between 47% and 66%. We compare the experimental results with numerical simulations (which assume perfect readout) and find that they generally predict the same performance under PEC or NOX (as should be expected when using RC + CER to simulate the circuits), but that the unmitigated simulated results perform significantly better than the REM experimental results. The difference between these two is likely due to the fact that the simulator does not take into account coherent errors.

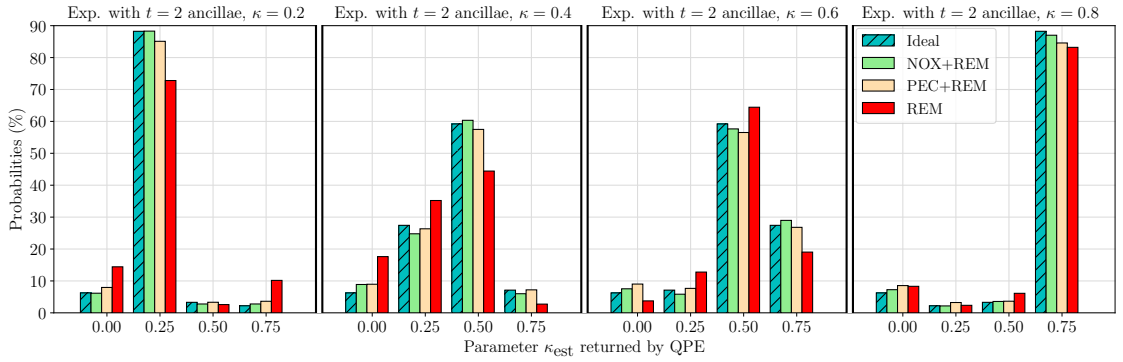
Quantum Phase Estimation

In our second set of experiments, we test the performance of NOX and PEC on the quantum phase estimation (QPE) algorithm, an important primitive in many different quantum algorithms. Given a gate U, a "target" state $|\psi\rangle$ and a (potentially unknown) number $\kappa \in [0, 1)$ such that $U|\psi\rangle = \exp(2i\kappa\pi)|\psi\rangle$, the task of QPE is to produce an estimate $\widehat{\kappa}$ of κ. To do so, QPE requires initialising $t \geq 1$ ancillae, entangling each of them with $|\psi\rangle$ and eventually performing the inverse Quantum Fourier Transform (QFT), as shown in Fig. 7.7. By measuring the ancillae and post-processing the outputs, QPE returns an estimate $\widehat{\kappa} \in \{p/2^t : p \in \{0, \ldots, 2^t - 1\}\}$, such that $|\kappa - \widehat{\kappa}| \leq 2^{-t}$ with high probability. For $t = 2$ ancillae and $|\psi\rangle = |1\rangle$, we estimate the parameter κ for a series of gates that perform rotations of the type

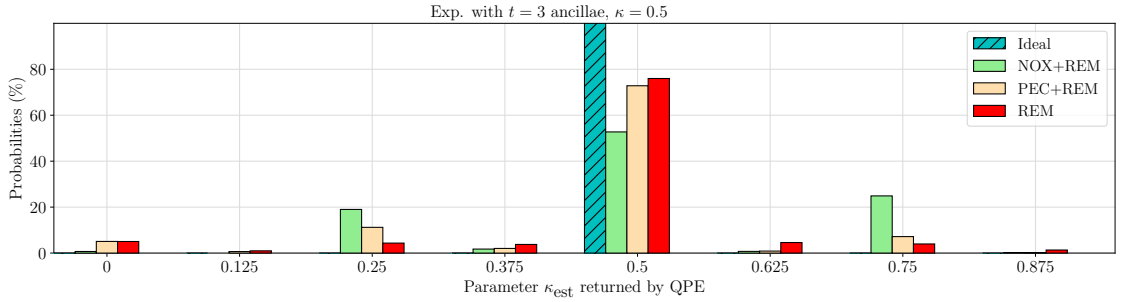
$$R_Z(\kappa) = \mathrm{diag}(1, \exp(2i\kappa\pi)). \tag{7.18}$$



(a) QPE with $t = 2$ ancillae (simulation) (b) QPE with $t = 2$ ancillae (experiment)



(c) Probabilities of κ_{est} for different values of κ with $t = 2$ ancillae.



(d) Probabilities of κ_{est} for $\kappa = 0.5$ with $t = 3$ ancillae.

Figure 7.8: **Quantum Phase Estimation results.** TVDs between the ideal and noisy outputs for (a) simultation and (b) experiment with $t = 2$ ancillae qubits and a target uncertainty of $\sigma = 2\%$; these circuits contain a total of $m = 9$ two-qubit gate cycles. (c) Probability that the QPE algorithm returns $\widehat{\kappa} \in \{0.0, 0.25, 0.50, 0.75\}$ for various different values of κ and $t = 2$ ancillae. The mitigated and unmitigated (+REM) results are compared versus the ideal probabilities. (d) Probability that the QPE algorithm returns $\widehat{\kappa} \in \{0.0, 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875\}$ for $\kappa = 0.5$ and $t = 3$ ancillae. (Figure reprinted with permission from Ref. [149].)

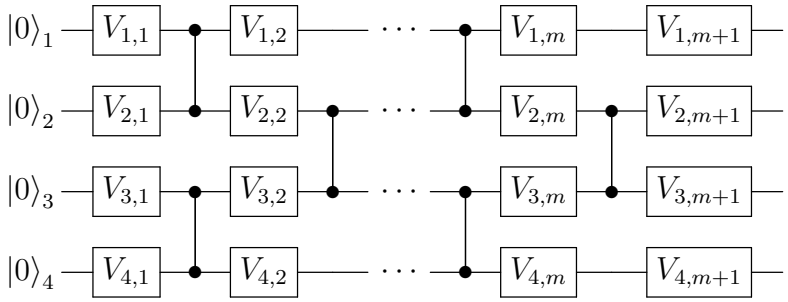


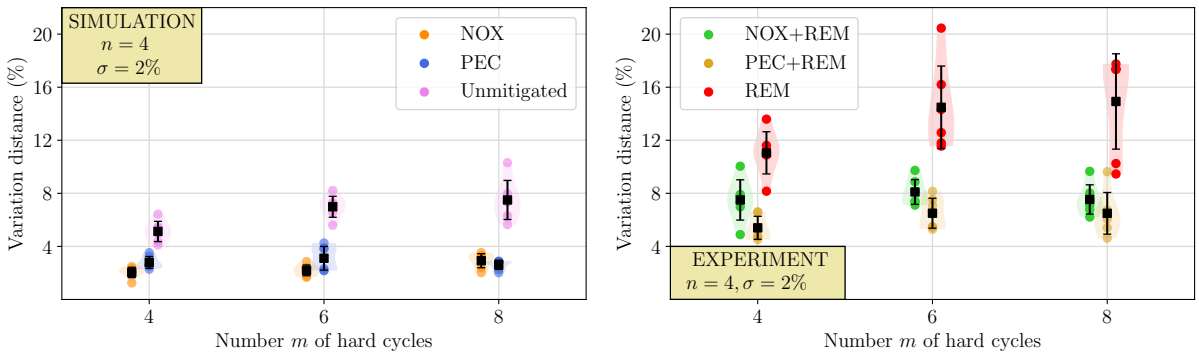
Figure 7.9: **Four-qubit pseudo-random circuits.** Each gate $V_{i,j}$ is a Haar random single-qubit gate.

Decomposed into our native gate set, this QPE circuit contain $n = 3$ qubits and $m = 11$ two-qubit gate cycles.

In Fig. 7.8 we plot the numerical and experimental results for the QPE circuits for various values of κ and find that NOX and PEC reduce the TVD on average between 42% and 86% over the results obtained with REM alone. To understand how NOX and PEC improve the precision of the QPE algorithm, we calculate the probability $p_{\text{est}}(\widehat{\kappa}|\kappa)$ that the QPE algorithm returns for $\widehat{\kappa} \in \{0.00, 0.25, 0.50, 0.75\}$, where κ is the true value. Fig. 7.8c shows the probabilities $p_{\text{est}}(\widehat{\kappa}|\kappa)$ obtained experimentally compared with the probabilities obtained via noiseless simulation. We see that NOX and PEC are generally closer to the ideal ones than those obtained in the experiments with only REM. We repeat the same experiment with $t = 3$ ancillae, setting $\kappa = 0.5$. Decomposed into our native gate set, the resulting QPE circuit contains $n = 4$ qubits and $m = 25$ two-qubit gate cycles. As opposed to the experiments with $t = 2$ ancillae, both NOX and PEC return less accurate outputs than the unmitigated (+REM) circuit and visibly decrease the precision of the QPE algorithm (Fig. 7.8d). We attribute this unsuccessful result primarily due to inaccurate amplification or cancellation of the noise at longer circuit depths. These unmitigated noise processes accumulate along the circuit, resulting in a bias that grows linearly in m, which becomes non-negligible in deep circuits. This failed test demonstrates what happens when assumptions regarding errors and noise break down, which have the potential to negate the performance gains of EM protocols.

Pseudo-random Circuits

Our final test was on pseudo-random circuits of the type plotted in Fig. 7.9. These circuits alternate between cycles containing Haar random single-qubit gates and two-qubit gate cycles. Fig. 7.10 shows the TVDs between ideal and estimated probability distributions of the outputs obtained numerically (Fig. 7.10a) and experimentally (Fig. 7.10b). As in our



(a) Simulation (b) Experiment

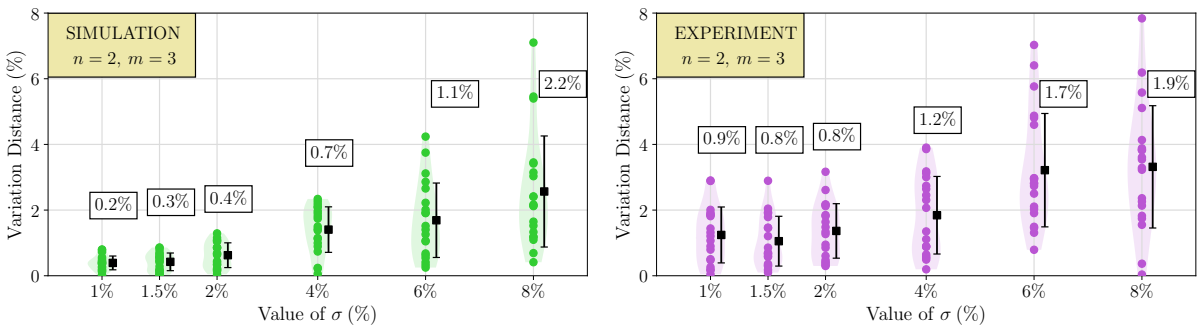
Figure 7.10: **Summary of the results for the pseudo-random circuits.** TVDs between the ideal and noisy probability distributions for (a) simulation and (b) experiments. The dots correspond to the actual data, the squares represent their means, and the bars their standard deviations. (Figure reprinted with permission from Ref. [149].)

previous tests, the mitigated outputs are visibly more accurate than the unmitigated ones, with average improvements between 32% and 56% over the unmitigated (+REM) outputs.

Relation between σ and the uncertainty of the estimators

In addition to suppressing the bias of the final estimators, our protocols provide guarantees about their statistical fluctuations. The input parameter σ represents the indented standard deviation uncertainty of the output estimator, and our EM protocols guarantee $\mathcal{O}(\sigma)$ standard deviation for every estimator, at the cost of running a number $N = \mathcal{O}(\sigma^{-2})$ of circuits. To verify that our protocols are providing the uncertainty guarantees as intended, we numerically and experimentally test the performance of NOX on two-qubit W-state circuits for different values of σ, shown in Fig. 7.11. For the numerical results in Fig. 7.11a, we observe a clear linear correlation between the value of σ and the uncertainty of the simulated results. However, for the experimental results in Fig. 7.11b, we observe a saturation in the measured uncertainty for lower values of σ. We find that for $\sigma \gtrsim 2\%$, the uncertainty of the results decreases with σ as expected, whereas for $\sigma \lesssim 2\%$ the uncertainty remains approximately constant.

The simulation results can be considered"ideal" experimental conditions in which the noise from the "environment" is constant and qubit parameters do not drift. However, in a real-world setting, a number of uncontrollable factors (such as drift in the noise afflicting the device in use) may inevitably prompt fluctuations in the estimators, limiting our ability to attain the desired statistical accuracy. Implementing NOX with $\sigma < 2\%$ requires running more circuits than with $\sigma = 2\%$, but for the experimental results in Fig. 7.11b it does not improve the statistical accuracy of the estimators. In other words, our EM protocols cannot



(a) Simulation

(b) Experiment

Figure 7.11: **NOX applied to two-qubit W-state circuits for different values of σ:** (a) Simulation, and (b) experiment. We list the standard deviation uncertainty of each result above the data points. (Figure reprinted with permission from Ref. [149].)

provide performance guarantees below the noise floor of the device.

7.3 Equivalent Circuit Averaging for QAOA

The Quantum Approximate Optimization Algorithm (QAOA) [250] describes a variational ansatz for solving combinatorial optimization problems described by an objective Hamiltonian H. QAOA is characterized by a hyperparameter p that specifies the depth of the ansatz. Specifically, the ansatz is $e^{i\beta_p B}e^{i\gamma_p H}...e^{i\beta_1 B}e^{i\gamma_1 H}$, where $B = \sum_i X_i$ is a mixing Hamiltonian and $\vec{\gamma}$, $\vec{\beta}$ represent $2p$ classically optimized variational parameters. It is believed that QAOA is hard to approximate even at $p = 1$ and is therefore a leading candidate for demonstrations of quantum advantage [251]. Therefore, it is important to optimize the execution of QAOA and the components that compose their circuits. We generate and optimize QAOA circuits corresponding to Sherrington-Kirkpatrick spin-glass model Hamiltonians with edge weights J_{ij} randomly selected from ± 1 (see Fig. 7.14 for the exact symbolic form of the circuits). Each $e^{i\gamma H}$ is then implemented with a network of ZZ-SWAP gates $\mathcal{F}_{\pm\gamma}$ [143] (see next section). Parameters β_i, γ_i are sampled uniformly from $[0, 2\pi)$.

ZZ-SWAP

The core operation needed in a QAOA SWAP network is the ZZ-SWAP gate, which is equivalent to a ZZ interaction followed by a SWAP operation, and is defined as the unitary

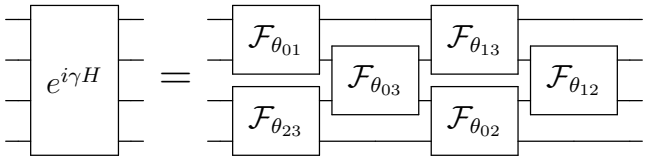
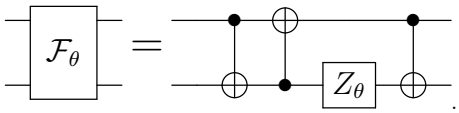


Figure 7.12: **SWAP network implementing the Hamiltonian evolution $e^{i\gamma H}$ for a four-node Sherrington-Kirkpatrick model.** $H = \sum_{i<j<4} J_{ij} Z_i Z_j$ and $\theta_{ij} = \gamma J_{ij}$. Note that the qubit order is reversed after the operation

operation below with input parameter θ:

$$\mathcal{F}_\theta = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & e^{i\theta} & 0 \\ 0 & e^{i\theta} & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{7.19}$$

The standard QASM-decomposed quantum circuit implementation of $\mathcal{F}_\theta$ comprises three CX gates and a single-qubit Z_θ rotation [252]:



(7.20)

It is possible to boost performance beyond this decomposition by leveraging knowledge of the target hardware's underlying native gate set. We do not cover this topic here, however, Ref. [143] demonstrates that the ZZ-SWAP gate can be optimized by having access to an *overcomplete* discrete two-qubit gate set, which contains additional two-qubit gates beyond what is necessary for universal quantum computation. Concretely, they examine the optimized $\mathcal{F}_\theta$ decompositions possible when they have access to both a CZ and $CS = \sqrt{CZ}$ gate, where the CS is shorter in duration than the CZ (see Tables 5.2 and 5.3).

In typical applications, multiple ZZ-SWAP gates are arranged into a nearest-neighbor ZZ-SWAP *network* that carries out $t = 0, \ldots, n-1$ steps. Each step alternates between an odd and even pattern. At steps with odd t, each neighboring qubit pair with indices $(2k, 2k+1)$ for $k \in [0, \frac{n}{2}]$ is entangled in accordance with a target Hamiltonian and then SWAPped. Even-t steps perform this interaction for qubit pairs with indices $(2k+1, 2k+2)$. Note that each step is highly parallel, with $\sim n/2$ operations occurring simultaneously. A prototypical example is shown in Fig. 7.12, which implements the Hamiltonian evolution $e^{i\gamma H}$ corresponding to a Sherrington-Kirkpatrick spin-glass model $H = \sum_{i<j<n} J_{ij} Z_i Z_j$ on $n = 4$ nodes. The utility of a SWAP network is that it efficiently generates an all-to-all interaction

with a linear-depth circuit of nearest-neighbor interactions, where each interaction is a single ZZ-SWAP gate corresponding to one of the commuting weight-2 terms in H.

We can optimize the decomposition of ZZ-SWAP gates in the context of larger SWAP networks (e.g. within QAOA). Various discrete and continuous symmetries in the ZZ-SWAP operation result in degrees of freedom to its optimized decomposition as follows:

1. $\mathcal{F}_\theta = (1 \otimes X)\mathcal{F}_{-\theta}(X \otimes 1)$,
2. $\mathcal{F}_\theta = (Z \otimes Z)\mathcal{F}_{\theta+\pi}$,
3. $\mathcal{F}_\theta(q_0, q_1) = \mathcal{F}_\theta(q_1, q_0)$ (qubit interchange),
4. $\mathcal{F}_\theta = \mathcal{F}_{-\theta}^\dagger$,
5. $\mathcal{F}_\theta = (Z_{-\vartheta} \otimes Z_{-\varphi})\mathcal{F}_\theta(Z_\varphi \otimes Z_\vartheta)\ \forall\ \vartheta, \varphi \in \mathbb{R}$,

where ϑ, φ are continuous parameters. Symmetry 4 is useful only for $\mathcal{F}_\theta$ gates implemented using a CS or $CS^\dagger$ gate, in which case it can be used to reverse the order of entangling gates in the circuit. For the case in which $\mathcal{F}_\theta$ is implemented using 3 CXs (or equivalently 3 CZs), the physical implementations of $\mathcal{F}_\theta$ and $\mathcal{F}_{-\theta}^\dagger$ are identical. Furthermore, any $\mathcal{F}_{\pm\theta}$ admitting a decomposition in terms of CS gates can also be decomposed in terms of $CS^\dagger$ gates, so we are always able to construct both $\mathcal{F}_\theta$ and $\mathcal{F}_{-\theta}^\dagger$ using either one of these gates (see Ref. [143] for more details).

Equivalent Circuit Averaging

QAOA circuits are highly structured, with repeating patterns of SWAP networks for each depth p. In Table 5.3 we list the benchmarking results for the various relevant two-qubit gate cycles appearing our ZZ-SWAP networks. While these results could in theory be used to predict the performance of our QAOA circuits, as outlined in Ref. [142], randomized benchmarks are not accurate predictors of the global performance of structured quantum circuits due to the presence of coherent errors. When averaging over a twirling group, such as the Clifford (Pauli) group for RB (CB), all errors are converted into a global depolarizing (stochastic Pauli) channel. However, in actual quantum algorithms, the physical error mechanisms are more complex than depolarizing or Pauli channels, as coherent errors can interfere constructively or destructively from one cycle to the next.

Being systematic in nature, coherent errors can, in theory, be measured and corrected via recalibration or added compensation pulses. However, the complexity of fully characterizing coherent errors (i.e. context-dependent rotation axes and angles [158]) on multi-qubit processors that arise due to classical and quantum crosstalk is intractable, and no known scalable methods exist for doing so for systems with continuous single-qubit gate sets. Various methods exist for suppressing coherent errors, such as dynamical decoupling [253] and error-correcting composite pulse sequences [254], or randomization methods for "tailoring"

them into stochastic noise, such as Pauli twirling [177, 181, 182, 224], Pauli frame randomization [157, 175, 176], and randomized compiling (RC) [145, 174]. However, these methods generally require the modification of single-qubit gates or the inclusion of more gates (e.g. in the case of dynamical decoupling and composite sequences), or require that the two-qubit gates in circuits are Clifford so that inverting Pauli operators can be efficiently computed and applied (as in the case of RC). Adopting these techniques would therefore require forgoing many potential circuit optimizations — both by necessitating additional single-qubit twirling gates, and precluding the use of non-Clifford CS and $CS^\dagger$ gates.

A similar strategy was proposed for circuit synthesis methods, in which systematic approximation errors are rendered incoherent by averaging over various circuits near a target unitary generated from ensembles of approximate decompositions [178, 179]. We employ this general idea (with systematic errors in the physical gates taking the place of approximation errors) using the space of equivalent ZZ-SWAP decompositions generated by the degrees of freedom outlined in the previous section. By randomly sampling from these decompositions for each ZZ-SWAP gate, we can generate a set of randomized but logically equivalent circuits to average over. We call this strategy equivalent circuit averaging (ECA). The computational overhead of ECA scales linearly with both the number of logically equivalent circuits to be generated and the cost of optimized scheduling for each circuit (proportional to the number of ZZ-SWAP gates in the circuit).

In Fig. 7.13, we measure two-qubit ($p = 1$) and four-qubit ($p = 1$ and $p = 2$) QAOA circuits (see Fig. 7.14 for example circuits) for various angles γ and benchmark the performance using the total variation distance (TVD),

$$D(p, q) = \frac{1}{2} \sum_{x \in X} |p_x - q_x|, \tag{7.21}$$

where p_x is the probability of measuring a bit string x in a set X, and q_x is the ideal (noiseless) probability. Standard (Std.) decompositions of the QAOA circuits are given in terms of the QASM-decomposed $\mathcal{F}_\theta$ (Eq. 7.20), and the optimized (Opt.) circuits outlined in Ref. [143] always have a shorter circuit depth. For the ECA results in Fig. 7.13, we generate $M = 20$ logically equivalent optimized circuits for each angle γ (see Fig. 7.14 for example circuits). To normalize shot statistics, we measure each equivalent circuit $s = S/M$ times and compute the union over all M results to obtain an equivalent statistical distribution for a circuit measured S times; $S = 10{,}000$ and $s = 500$ for the results in Fig. 7.13. We see that ECA dramatically reduces the TVD on average in comparison to both the standard and optimized results for all of the two- and four-qubit QAOA SWAP network results, reducing the average TVD by $\sim$ 60% [26%] from $D_{\text{Std.}} = 0.20(5)$ to $D_{\text{ECA}} = 0.08(2)$ [$D_{\text{Std.}} = 0.23(4)$ to $D_{\text{ECA}} = 0.17(6)$] for the four-qubit $p = 1$ [$p = 2$] QAOA results, and providing the most accurate measured probability distribution in 88% of all of the two- and four-qubit circuits measured.

While the classical overhead of generating and measuring M logically equivalent circuits increases linearly in M, we observe significant improvements in the measured results. These

results demonstrate that ECA is a useful tool for smart compilers which optimize circuit decomposition using various degrees of freedom, and is not limited to circuits only containing two-qubit Clifford gates, adding to the toolbox of randomization methods that can be employed in the NISQ era. We also note that averaging out systematic errors is likely beneficial even at the expense of gate-level optimizations. The simplest version of ECA, which samples from the set of all logically equivalent decompositions of each ZZ-SWAP by randomly applying the symmetries outlined in the previous section without consideration of circuit depth would have negligible computational overhead.

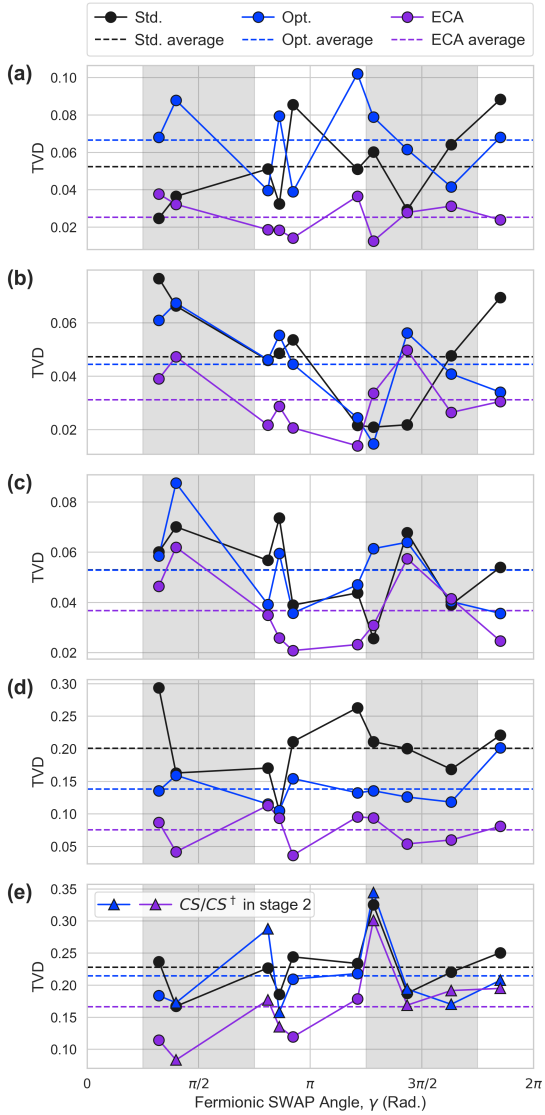


Figure 7.13: **Improved QAOA SWAP networks via gate-based optimizations.** The TVD performance of QAOA networks of angle γ are plotted for qubits (a) (Q4, Q5), (b) (Q5, Q6), and (c) (Q6, Q7), and four-qubit circuits with (d) $p = 1$ and (e) $p = 2$ stages. The mixing parameter β is chosen at random for each network. The ECA (purple) results consistently outperform both the standard and optimized decompositions (see Ref. [143]). The gray shaded regions define the angles for which CS or $CS^\dagger$ gates can be utilized in the circuit decompositions. The results in (e) are plotted against the γ from stage 1 (the triangle markers denote circuits which utilize a CS or $CS^\dagger$ gate in stage 2). (Error bars on the TVD $\sim \mathcal{O}[10^{-3}]$ are smaller than the markers.)

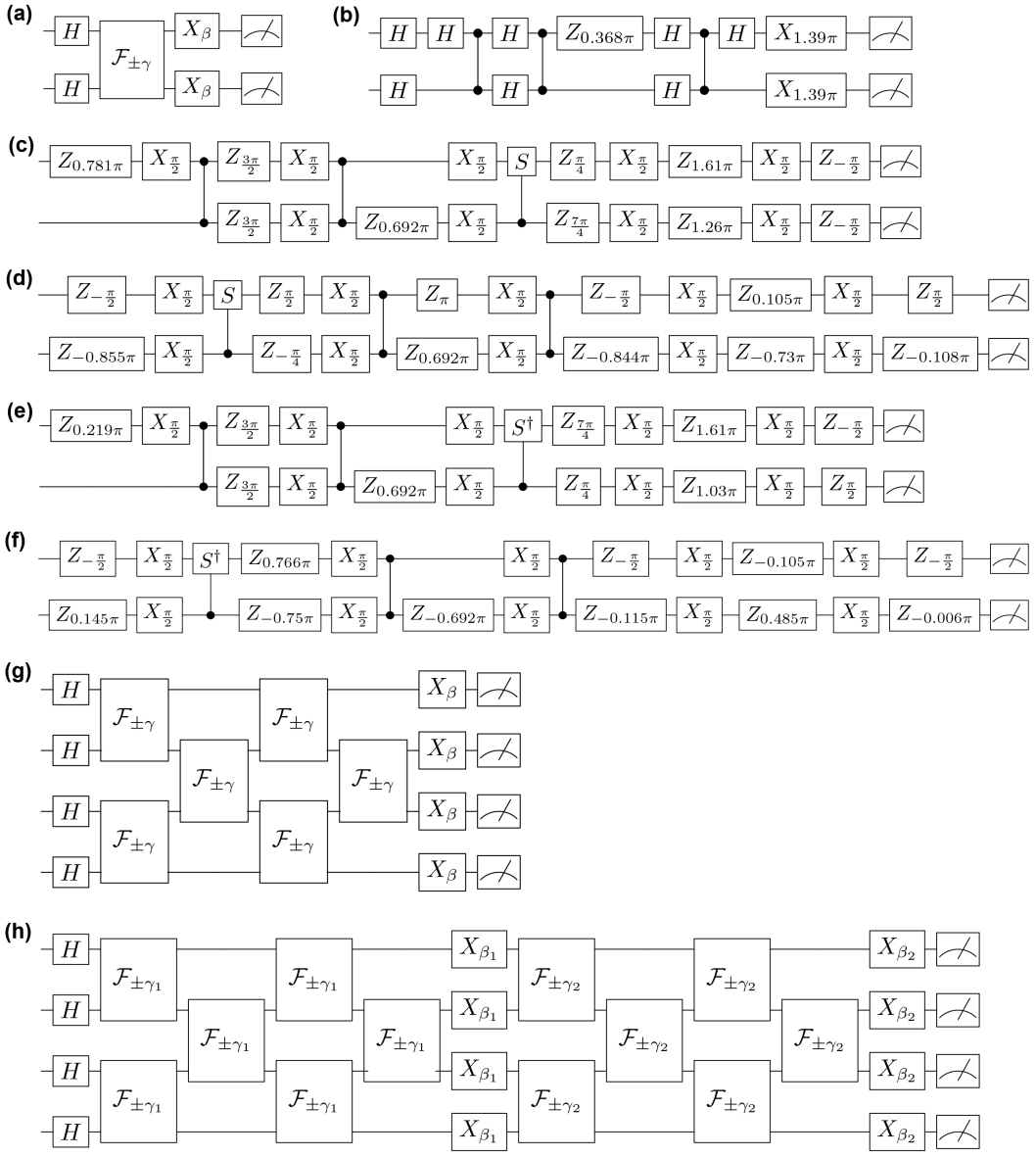


Figure 7.14: **Example SWAP networks for QAOA.** Symbolic circuit representations of QAOA SWAP networks (of depth p) for (a) two qubits ($p = 1$), (g) four qubits ($p = 1$), and (h) four qubits ($p = 2$). (b) Baseline decomposition of a two-qubit QAOA SWAP network for a random choice of γ and β with CZs. (c) Optimized decomposition of the circuit in (b) in terms of the native gate set utilizing a CS instead of a CZ. (d) Logically equivalent decomposition of the circuit in (c). (e) Optimized decomposition of the circuit in (b) in terms of the native gate set utilizing a $CS^\dagger$ instead of a CZ. (f) Logically equivalent decomposition of the circuit in (e).

Chapter 8

Outlook

In this work, we have demonstrated the promising capabilities of randomized compiling (RC), a universal protocol in gate-based quantum computing for tailoring arbitrary Markovian errors that is agnostic to specific error models and hardware platforms. RC provides a strategy for mitigating complex and intractable crosstalk dynamics, extending the computational reach of noisy quantum processors. By twirling noise in the same manner as randomized benchmarks, RC enables accurate predictions of quantum algorithms *a priori*. This improved predictability is essential for building scalable quantum computers. These results have broad relevance across many experimental and theoretical efforts exploring gate-based quantum computing applications

RC is not the only method which can tailor noise for quantum algorithms. Various other randomization methods exists, and to that end we introduced equivalent circuit averaging (ECA) to mitigate the impact of coherent errors in non-Clifford circuits by utilizing various degrees of freedom of quantum compilers to generate many logically equivalent circuits. Given the difficulty in characterizing and predicting the impact of coherent errors on algorithm performance, such a method negates the need for doing so and assumes that the average over many circuits will reduce the impact of coherent errors on the algorithm results. We demonstrate the effectiveness of this approach with application benchmark results, in which we find that ECA improves the accuracy of the measured probability distribution for 88% of randomly-generated QAOA circuits. While ECA was employed by taking advantage of the various degrees of freedom in networks of ZZ-SWAP gates, a more sophisticated search procedure would likely expand the applicability of our methods for scheduling and generating equivalent circuits for more general applications. We further imagine possible "hybrid" strategies in which ECA is combined with other randomization protocols (e.g. RC) for maximizing the ways in which logically equivalent circuits can be expressed, thus minimizing residual coherent errors. The cost of ECA (both classically and in terms of single-qubit gate optimization) in the general case and the degree to which it tailors noise in quantum systems (i.e. in the manner of other randomization methods which twirl over a specific gate set) are open questions.

While fault tolerance remains the long-term goal, improving the performance of NISQ

algorithms is important for intermediate demonstrations of quantum advantage. For this reason, error mitigation (EM) has become an important subfield in QIS. While many different EM strategies exist, each with their own benefits and disadvantages, we have shown that RC uniquely enables several different classes of methods, including purification-based methods, methods which amplify noise in a structured manner, and scalable quasi-probabilistic protocols. These results demonstrate that RC can enhance the performance of quantum circuits implemented on noisy quantum processors with reasonable overhead.

As we look forward toward fault tolerance, it remains an open question to what extent EM will remain necessary, and for how long. Recent works have shown the EM schemes can be built into fault-tolerant quantum computations for correcting errors in non-Clifford T-gates without magic state distillation [255] and improving QEC code distances [256]. Due to their ability to suppress multi-qubit errors, we anticipate that EM can be used to correct large-weight correlated errors that may be beyond the ability QEC codes (depending on the distance of the code), and therefore reduce errors at the logical level.

While full-scale fault-tolerant QEC is still many years away, contemporary quantum gates are rapidly approaching the necessary thresholds for some QEC codes, such as the surface code. Therefore, it is important to be able to accurately benchmark our progress toward this goal. The field of quantum characterization, verification, and validation (QCVV) was developed in large part for this reason — to benchmark our progress toward fault-tolerant QEC. While average error rates measured via randomized benchmarks are useful for tracking progress, they fall short of capturing the information required to determine whether all gate errors fall below a given FT threshold, for which the diamond norm is the relevant metric [120, 130]. However, we demonstrate that FT thresholds can in fact be captured by randomized benchmarks, but only if a quantum application is performed using randomization methods which tailor noise. This noise tailoring may even impact the kinds of errors that manifest at the logical level, as, despite the projective nature of QEC, it has been shown that coherent errors at the physical level can lead to coherent errors at the logical level [257, 258, 259, 260]. An open question remains as to what guarantees these benchmarked values provide for error correction, which operates in the single-shot, single-randomization limit.

While rapid improvements in two-qubit gates on many hardware platforms engender optimism for FT, caution must be taken in the claims inferred from gate fidelities, as results which include estimates of the diamond norm suggest that many contemporary two-qubit gates fall short of FT thresholds [201, 202, 261, 262, 263]. Furthermore, while isolated single- and two-qubit gates may be approaching the necessary requirements for fault-tolerance, any gate(s) performed in parallel with other qubits are likely to be impacted by crosstalk-induced coherent errors, potentially causing the diamond norm to scale with $\sqrt{e_F}$. The figure of merit for determining whether low logical error rates can be achieved via QEC is the error rate of a cycle containing all active qubits in a register, not simply the error rate of isolated gates within the cycle [215]. By utilizing RC, we ensure that fault-tolerant error thresholds are set by stochastic errors, not coherent errors. To this end, we expect that randomized compiling is not just a stopgap measure in the NISQ era, but will remain a powerful technique beyond NISQ.

Bibliography

[1] Jonathan P Dowling and Gerard J Milburn. "Quantum technology: the second quantum revolution". In: *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 361.1809 (2003), pp. 1655–1674.

[2] Ching-Ray Chang et al. "The Second Quantum Revolution with Quantum Computers." In: *AAPPS Bulletin* 30.1 (2020).

[3] Ivan H Deutsch. "Harnessing the power of the second quantum revolution". In: *PRX Quantum* 1.2 (2020), p. 020101.

[4] Michel H Devoret and Robert J Schoelkopf. "Superconducting circuits for quantum information: an outlook". In: *Science* 339.6124 (2013), pp. 1169–1174.

[5] William D Oliver and Paul B Welander. "Materials in superconducting quantum bits". In: *MRS bulletin* 38.10 (2013), pp. 816–825.

[6] Elizabeth Gibney. "The quantum gold rush". In: *Nature* 574.7776 (2019), pp. 22–24.

[7] Frank Arute et al. "Quantum supremacy using a programmable superconducting processor". In: *Nature* 574.7779 (2019), pp. 505–510.

[8] Han-Sen Zhong et al. "Quantum computational advantage using photons". In: *Science* 370.6523 (2020), pp. 1460–1463.

[9] Yulin Wu et al. "Strong quantum computational advantage using a superconducting quantum processor". In: *Physical review letters* 127.18 (2021), p. 180501.

[10] Lars S Madsen et al. "Quantum computational advantage with a programmable photonic processor". In: *Nature* 606.7912 (2022), pp. 75–81.

[11] Andrew J Daley et al. "Practical quantum advantage in quantum simulation". In: *Nature* 607.7920 (2022), pp. 667–676.

[12] Irfan Siddiqi. "Engineering high-coherence superconducting qubits". In: *Nature Reviews Materials* 6.10 (2021), pp. 875–891.

[13] Iulia Buluta, Sahel Ashhab, and Franco Nori. "Natural and artificial atoms for quantum computation". In: *Reports on Progress in Physics* 74.10 (2011), p. 104401.

[14] John Preskill. "Quantum computing in the NISQ era and beyond". In: *Quantum* 2 (2018), p. 79.

[15] Marlan O Scully and M Suhail Zubairy. *Quantum optics*. 1999.

[16] Christopher Gerry, Peter Knight, and Peter L Knight. *Introductory quantum optics*. Cambridge university press, 2005.

[17] D.F. Walls and Gerard J. Milburn. *Quantum Optics*. Springer Berlin, Heidelberg, 2008.

[18] Serge Haroche and J-M Raimond. *Exploring the quantum: atoms, cavities, and photons*. Oxford university press, 2006.

[19] Michael Tinkham. *Introduction to superconductivity*. Courier Corporation, 2004.

[20] Uri Vool and Michel Devoret. "Introduction to quantum electromagnetic circuits". In: *International Journal of Circuit Theory and Applications* 45.7 (2017), pp. 897–934.

[21] SE Rasmussen et al. "Superconducting circuit companion—an introduction with worked examples". In: *PRX Quantum* 2.4 (2021), p. 040204.

[22] Philip Krantz et al. "A quantum engineer's guide to superconducting qubits". In: *Applied Physics Reviews* 6.2 (2019), p. 021318.

[23] Alexandre Blais et al. "Circuit quantum electrodynamics". In: *Reviews of Modern Physics* 93.2 (2021), p. 025005.

[24] David Isaac Schuster. *Circuit quantum electrodynamics*. Yale University, 2007.

[25] Daniel Huber Slichter. *Quantum jumps and measurement backaction in a superconducting qubit*. University of California, Berkeley, 2011.

[26] Michael A Nielsen and Isaac Chuang. *Quantum computation and quantum information*. 2002.

[27] Marco Cerezo et al. "Variational quantum algorithms". In: *Nature Reviews Physics* 3.9 (2021), pp. 625–644.

[28] Aashish A Clerk et al. "Introduction to quantum noise, measurement, and amplification". In: *Reviews of Modern Physics* 82.2 (2010), p. 1155.

[29] Leigh Martin. *Quantum feedback for measurement and control*. University of California, Berkeley, 2019.

[30] Brian David Josephson. "Possible new effects in superconductive tunnelling". In: *Physics letters* 1.7 (1962), pp. 251–253.

[31] Philip W Anderson and John M Rowell. "Probable observation of the Josephson superconducting tunneling effect". In: *Physical Review Letters* 10.6 (1963), p. 230.

[32] Brian D Josephson. "The discovery of tunnelling supercurrents". In: *Reviews of Modern Physics* 46.2 (1974), p. 251.

[33] RC Jaklevic et al. "Quantum interference effects in Josephson tunneling". In: *Physical Review Letters* 12.7 (1964), p. 159.

[34] John Clarke and Alex I Braginski. *The SQUID handbook*. Vol. 1. Wiley Online Library, 2004.

[35] J Clarke. "The use of superconducting junctions in magnetometry". In: *Revue de Physique Appliquée* 5.1 (1970), pp. 32–36.

[36] RH Koch et al. "Three SQUID gradiometer". In: *Applied physics letters* 63.3 (1993), pp. 403–405.

[37] R Pöpel. "The Josephson effect and voltage standards". In: *Metrologia* 29.2 (1992), p. 153.

[38] Tsuyoshi Yamamoto et al. "Flux-driven Josephson parametric amplifier". In: *Applied Physics Letters* 93.4 (2008), p. 042510.

[39] Chris Macklin et al. "A near–quantum-limited Josephson traveling-wave parametric amplifier". In: *Science* 350.6258 (2015), pp. 307–310.

[40] John M Martinis and Kevin Osborne. "Superconducting qubits and the physics of Josephson junctions". In: *arXiv preprint cond-mat/0402415* (2004).

[41] John Clarke and Frank K Wilhelm. "Superconducting quantum bits". In: *Nature* 453.7198 (2008), p. 1031.

[42] Jens Koch et al. "Charge-insensitive qubit design derived from the Cooper pair box". In: *Physical Review A* 76.4 (2007), p. 042319.

[43] Marvin L Cohen and Steven G Louie. *Fundamentals of condensed matter physics*. Cambridge University Press, 2016.

[44] MT Levinsen et al. "An inverse ac Josephson effect voltage standard". In: *Applied Physics Letters* 31.11 (1977), pp. 776–778.

[45] Sidney Shapiro. "Josephson currents in superconducting tunneling: The effect of microwaves and other observations". In: *Physical Review Letters* 11.2 (1963), p. 80.

[46] R.D. Parks. *B.D. Josephson in Superconductivity*. Marcel Dekker, New York, 1969.

[47] M Büttiker. "Zero-current persistent potential drop across small-capacitance Josephson junctions". In: *Physical Review B* 36.7 (1987), p. 3548.

[48] Alexander Shnirman, Gerd Schön, and Ziv Hermon. "Quantum manipulations of small Josephson junctions". In: *Physical Review Letters* 79.12 (1997), p. 2371.

[49] Vincent Bouchiat et al. "Quantum coherence with a single Cooper pair". In: *Physica Scripta* 1998.T76 (1998), p. 165.

[50] Yu Nakamura, Yu A Pashkin, and JS Tsai. "Coherent control of macroscopic quantum states in a single-Cooper-pair box". In: *nature* 398.6730 (1999), p. 786.

[51] Peter Groszkowski and Jens Koch. "Scqubits: a Python package for superconducting qubits". In: *Quantum* 5 (2021), p. 583.

[52] Michel H Devoret and Robert J Schoelkopf. "Amplifying quantum signals with the single-electron transistor". In: *Nature* 406.6799 (2000), p. 1039.

[53] John M Martinis et al. "Rabi oscillations in a large Josephson-junction qubit". In: *Physical review letters* 89.11 (2002), p. 117901.

[54] I Chiorescu et al. "Coherent quantum dynamics of a superconducting flux qubit". In: *Science* 299.5614 (2003), pp. 1869–1871.

[55] David C. McKay et al. "Efficient Z gates for quantum computing". In: *Phys. Rev. A* 96 (2 Aug. 2017), p. 022330. DOI: 10.1103/PhysRevA.96.022330. URL: https://link.aps.org/doi/10.1103/PhysRevA.96.022330.

[56] GS Paraoanu. "Microwave-induced coupling of superconducting qubits". In: *Physical Review B* 74.14 (2006), p. 140504.

[57] Chad Rigetti and Michel Devoret. "Fully microwave-tunable universal gates in superconducting qubits with linear couplings and fixed transition frequencies". In: *Physical Review B* 81.13 (2010), p. 134507.

[58] Jerry M Chow et al. "Simple all-microwave entangling gate for fixed-frequency superconducting qubits". In: *Physical Review Letters* 107.8 (2011), p. 080502.

[59] Sarah Sheldon et al. "Procedure for systematically tuning up cross-talk in the cross-resonance gate". In: *Physical Review A* 93.6 (2016), p. 060302.

[60] Bradley K Mitchell et al. "Hardware-efficient microwave-activated tunable coupling between superconducting qubits". In: *Physical review letters* 127.20 (2021), p. 200502.

[61] Noah Goss et al. "High-Fidelity Qutrit Entangling Gates for Superconducting Circuits". In: *arXiv preprint arXiv:2206.07216* (2022).

[62] Edwin T Jaynes and Frederick W Cummings. "Comparison of quantum and semiclassical radiation theories with application to the beam maser". In: *Proceedings of the IEEE* 51.1 (1963), pp. 89–109.

[63] Bruce W Shore and Peter L Knight. "The jaynes-cummings model". In: *Journal of Modern Optics* 40.7 (1993), pp. 1195–1238.

[64] Alexandre Blais et al. "Cavity quantum electrodynamics for superconducting electrical circuits: An architecture for quantum computation". In: *Physical Review A* 69.6 (2004), p. 062320.

[65] Maxime Boissonneault, Jay M Gambetta, and Alexandre Blais. "Dispersive regime of circuit QED: Photon-dependent qubit dephasing and relaxation rates". In: *Physical Review A* 79.1 (2009), p. 013819.

[66] David P DiVincenzo. "The physical implementation of quantum computation". In: *Fortschritte der Physik: Progress of Physics* 48.9-11 (2000), pp. 771–783.

[67] D Rosenberg et al. "3D integrated superconducting qubits". In: *npj quantum information* 3.1 (2017), p. 42.

[68] Timothy Proctor et al. "Detecting and tracking drift in quantum information processors". In: *Nature communications* 11.1 (2020), pp. 1–9.

[69] Mohan Sarovar et al. "Detecting crosstalk errors in quantum information processors". In: *Quantum* 4 (2020), p. 321.

[70] Adam Winick, Joel J Wallman, and Joseph Emerson. "Simulating and mitigating crosstalk". In: *Physical Review Letters* 126.23 (2021), p. 230502.

[71] Siyuan Niu and Aida Todri-Sanial. "Analyzing crosstalk error in the NISQ era". In: *2021 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*. IEEE. 2021, pp. 428–430.

[72] Peng Zhao et al. "Quantum crosstalk analysis for simultaneous gate operations on superconducting qubits". In: *arXiv preprint arXiv:2110.12570* (2021).

[73] KX Wei et al. "Quantum crosstalk cancellation for fast entangling gates and improved multi-qubit performance". In: *arXiv preprint arXiv:2106.00675* (2021).

[74] Fraidoon Mazda. *Telecommunications engineer's reference book*. Butterworth-Heinemann, 2014.

[75] Sebastian Krinner et al. "Benchmarking coherent errors in controlled-phase gates due to spectator qubits". In: *Physical Review Applied* 14.2 (2020), p. 024042.

[76] Neereja Sundaresan et al. "Reducing unitary and spectator errors in cross resonance with optimized rotary echoes". In: *PRX Quantum* 1.2 (2020), p. 020318.

[77] Alfred G Redfield. "On the theory of relaxation processes". In: *IBM Journal of Research and Development* 1.1 (1957), pp. 19–31.

[78] Robin Blume-Kohout et al. "A taxonomy of small Markovian errors". In: *PRX Quantum* 3.2 (2022), p. 020335.

[79] Daniel Greenbaum. "Introduction to quantum gate set tomography". In: *arXiv preprint arXiv:1509.02921* (2015).

[80] Ángel Rivas, Susana F Huelga, and Martin B Plenio. "Quantum non-Markovianity: characterization, quantification and detection". In: *Reports on Progress in Physics* 77.9 (2014), p. 094001.

[81] Heinz-Peter Breuer et al. "Colloquium: Non-Markovian dynamics in open quantum systems". In: *Reviews of Modern Physics* 88.2 (2016), p. 021002.

[82] Li Li, Michael JW Hall, and Howard M Wiseman. "Concepts of quantum non-Markovianity: A hierarchy". In: *Physics Reports* 759 (2018), pp. 1–51.

[83] C-F Li, G-C Guo, and J Piilo. "Non-Markovian quantum dynamics: What does it mean?" In: *EPL (Europhysics Letters)* 127.5 (2019), p. 50001.

[84] Joydip Ghosh et al. "Understanding the effects of leakage in superconducting quantum-error-detection circuits". In: *Physical Review A* 88.6 (2013), p. 062329.

[85] Joel J Wallman, Marie Barnhill, and Joseph Emerson. "Robust characterization of leakage errors". In: *New Journal of Physics* 18.4 (2016), p. 043021.

[86] Zijun Chen et al. "Measuring and suppressing quantum state leakage in a superconducting qubit". In: *Physical review letters* 116.2 (2016), p. 020501.

[87] Christopher J Wood and Jay M Gambetta. "Quantification and characterization of leakage errors". In: *Physical Review A* 97.3 (2018), p. 032306.

[88] D Hayes et al. "Eliminating leakage errors in hyperfine qubits". In: *Physical Review Letters* 124.17 (2020), p. 170501.

[89] Aravind Plathanam Babu, Jani Tuorila, and Tapio Ala-Nissila. "State leakage during fast decay and control of a superconducting transmon qubit". In: *npj Quantum Information* 7.1 (2021), pp. 1–8.

[90] K Serniak et al. "Hot nonequilibrium quasiparticles in transmon qubits". In: *Physical review letters* 121.15 (2018), p. 157701.

[91] SE de Graaf et al. "Two-level systems in superconducting quantum devices due to trapped quasiparticles". In: *Science advances* 6.51 (2020), eabc5055.

[92] Christopher D Wilen et al. "Correlated charge noise and relaxation errors in superconducting qubits". In: *Nature* 594.7863 (2021), pp. 369–373.

[93] Maya Berlin-Udi et al. "Changes in electric-field noise due to thermal transformation of a surface ion trap". In: *arXiv preprint arXiv:2103.04482* (2021).

[94] Anna E Webb et al. "Resilient entangling gates for trapped ions". In: *Physical review letters* 121.18 (2018), p. 180501.

[95] Marcin Pozniak, Karol Zyczkowski, and Marek Kus. "Composed ensembles of random unitary matrices". In: *Journal of Physics A: Mathematical and General* 31.3 (1998), p. 1059.

[96] Christoph Dankert et al. "Exact and approximate unitary 2-designs and their application to fidelity estimation". In: *Physical Review A* 80.1 (2009), p. 012304.

[97] Daniel Gottesman. "Theory of fault-tolerant quantum computation". In: *Physical Review A* 57.1 (1998), p. 127.

[98] Daniel Gottesman. "The Heisenberg representation of quantum computers". In: *arXiv preprint quant-ph/9807006* (1998).

[99] Eiichi Bannai and Etsuko Bannai. "A survey on spherical designs and algebraic combinatorics on spheres". In: *European Journal of Combinatorics* 30.6 (2009), pp. 1392–1425.

[100] Joseph Emerson et al. "Pseudo-random unitary operators for quantum information processing". In: *science* 302.5653 (2003), pp. 2098–2100.

[101] David Gross, Koenraad Audenaert, and Jens Eisert. "Evenly distributed unitaries: On the structure of unitary designs". In: *Journal of mathematical physics* 48.5 (2007), p. 052104.

[102] Rami Barends et al. "Superconducting quantum circuits at the surface code threshold for fault tolerance". In: *Nature* 508.7497 (2014), pp. 500–503.

[103] Matthew A Graydon, Joshua Skanes-Norman, and Joel J Wallman. "Clifford groups are not always 2-designs". In: *arXiv preprint arXiv:2108.04200* (2021).

[104] Zak Webb. "The Clifford group forms a unitary 3-design". In: *arXiv preprint arXiv:1510.02769* (2015).

[105] Huangjun Zhu. "Multiqubit Clifford groups are unitary 3-designs". In: *Physical Review A* 96.6 (2017), p. 062336.

[106] Peter W Shor. "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer". In: *SIAM review* 41.2 (1999), pp. 303–332.

[107] Peter W Shor. "Scheme for reducing decoherence in quantum computer memory". In: *Physical review A* 52.4 (1995), R2493.

[108] Daniel Gottesman. "Class of quantum error-correcting codes saturating the quantum Hamming bound". In: *Physical Review A* 54.3 (1996), p. 1862.

[109] Andrew Steane. "Multiple-particle interference and quantum error correction". In: *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 452.1954 (1996), pp. 2551–2577.

[110] AM Steane. "Introduction to quantum error correction". In: *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 356.1743 (1998), pp. 1739–1758.

[111] A Robert Calderbank et al. "Quantum error correction via codes over GF (4)". In: *IEEE Transactions on Information Theory* 44.4 (1998), pp. 1369–1387.

[112] Peter W Shor. "Fault-tolerant quantum computation". In: *Proceedings of 37th Conference on Foundations of Computer Science*. IEEE. 1996, pp. 56–65.

[113] Emanuel Knill and Raymond Laflamme. "Concatenated quantum codes". In: *arXiv preprint quant-ph/9608012* (1996).

[114] Emanuel Knill, Raymond Laflamme, and Wojciech H Zurek. "Resilient quantum computation". In: *Science* 279.5349 (1998), pp. 342–345.

[115] John Preskill. "Fault-tolerant quantum computation". In: *Introduction to quantum computation and information*. World Scientific, 1998, pp. 213–269.

[116] A Yu Kitaev. "Fault-tolerant quantum computation by anyons". In: *Annals of Physics* 303.1 (2003), pp. 2–30.

[117] Dorit Aharonov and Michael Ben-Or. "Fault-tolerant quantum computation with constant error rate". In: *SIAM Journal on Computing* (2008).

[118] Alexei Gilchrist, Nathan K Langford, and Michael A Nielsen. "Distance measures to compare real and ideal quantum processes". In: *Physical Review A* 71.6 (2005), p. 062310.

[119] Jennifer L Dodd and Michael A Nielsen. "A simple operational interpretation of the fidelity". In: *arXiv preprint quant-ph/0111053* (2001).

[120] Yuval R Sanders, Joel J Wallman, and Barry C Sanders. "Bounding quantum gate error rate based on reported average fidelity". In: *New Journal of Physics* 18.1 (2015), p. 012002.

[121] Timothy Proctor et al. "What randomized benchmarking actually measures". In: *Physical review letters* 119.13 (2017), p. 130502.

[122] Michał Horodecki, Paweł Horodecki, and Ryszard Horodecki. "General teleportation channel, singlet fraction, and quasidistillation". In: *Physical Review A* 60.3 (1999), p. 1888.

[123] Michael A Nielsen. "A simple formula for the average gate fidelity of a quantum dynamical operation". In: *Physics Letters A* 303.4 (2002), pp. 249–252.

[124] Arnaud Carignan-Dugas. "A walk through quantum noise: a study of error signatures and characterization methods". In: (2019).

[125] Kristine Boone. "Concepts and methods for benchmarking quantum computers". In: (2021).

[126] Christopher A Fuchs and Jeroen Van De Graaf. "Cryptographic distinguishability measures for quantum-mechanical states". In: *IEEE Transactions on Information Theory* 45.4 (1999), pp. 1216–1227.

[127] Aleksei Yur'evich Kitaev. "Quantum computations: algorithms and error correction". In: *Uspekhi Matematicheskikh Nauk* 52.6 (1997), pp. 53–112.

[128] Joel J Wallman. "Bounding experimental quantum error rates relative to fault-tolerant thresholds". In: *arXiv preprint arXiv:1511.00727* (2015).

[129] Joel J Wallman and Steven T Flammia. "Randomized benchmarking with confidence". In: *New Journal of Physics* 16.10 (2014), p. 103032.

[130] Richard Kueng et al. "Comparing experiments to the fault-tolerance threshold". In: *Physical review letters* 117.17 (2016), p. 170502.

[131] Joseph Emerson, Robert Alicki, and Karol Życzkowski. "Scalable noise estimation with random unitary operators". In: *Journal of Optics B: Quantum and Semiclassical Optics* 7.10 (2005), S347.

[132] Emanuel Knill et al. "Randomized benchmarking of quantum gates". In: *Physical Review A* 77.1 (2008), p. 012307.

[133] Easwar Magesan, Jay M Gambetta, and Joseph Emerson. "Scalable and robust randomized benchmarking of quantum processes". In: *Physical Review Letters* 106.18 (2011), p. 180504.

[134] Easwar Magesan, Jay M Gambetta, and Joseph Emerson. "Characterizing quantum gates via randomized benchmarking". In: *Physical Review A* 85.4 (2012), p. 042311.

[135] Isaac L Chuang and Michael A Nielsen. "Prescription for experimental determination of the dynamics of a quantum black box". In: *Journal of Modern Optics* 44.11-12 (1997), pp. 2455–2467.

[136] Robin Harper et al. "Statistical analysis of randomized benchmarking". In: *Physical Review A* 99.5 (2019), p. 052350.

[137] Easwar Magesan et al. "Efficient measurement of quantum gate error by interleaved randomized benchmarking". In: *Physical review letters* 109.8 (2012), p. 080505.

[138] Arnaud Carignan-Dugas, Joel J Wallman, and Joseph Emerson. "Bounding the average gate fidelity of composite channels using the unitarity". In: *New Journal of Physics* 21.5 (2019), p. 053016.

[139] Joel Wallman et al. "Estimating the coherence of noise". In: *New Journal of Physics* 17.11 (2015), p. 113020.

[140] Felix Motzoi et al. "Simple pulses for elimination of leakage in weakly nonlinear qubits". In: *Physical review letters* 103.11 (2009), p. 110501.

[141] Alexander Erhard et al. "Characterizing large-scale quantum computers via cycle benchmarking". In: *Nature communications* 10.1 (2019), pp. 1–7.

[142] Timothy Proctor et al. "Measuring the capabilities of quantum computers". In: *Nature Physics* 18.1 (2022), pp. 75–79.

[143] Akel Hashim et al. "Optimized SWAP networks with equivalent circuit averaging for QAOA". In: *Phys. Rev. Research* 4 (3 July 2022), p. 033028. DOI: 10.1103/PhysRevResearch.4.033028. URL: https://link.aps.org/doi/10.1103/PhysRevResearch.4.033028.

[144] Yosep Kim et al. "High-fidelity three-qubit iToffoli gate for fixed-frequency superconducting qubits". In: *Nature Physics* (2022), pp. 1–6.

[145] Akel Hashim et al. "Randomized Compiling for Scalable Quantum Computing on a Noisy Superconducting Quantum Processor". In: *Phys. Rev. X* 11 (4 Nov. 2021), p. 041039. DOI: 10.1103/PhysRevX.11.041039. URL: https://link.aps.org/doi/10.1103/PhysRevX.11.041039.

[146] Steven T Flammia and Joel J Wallman. "Efficient estimation of Pauli channels". In: *ACM Transactions on Quantum Computing* 1.1 (2020), pp. 1–32.

[147] Robin Harper, Steven T Flammia, and Joel J Wallman. "Efficient learning of quantum noise". In: *Nature Physics* 16.12 (2020), pp. 1184–1188.

[148] Stefanie J. Beale et al. *True-Q*. Version 2.6.13. June 2020. DOI: 10.5281/zenodo.3945250. URL: https://doi.org/10.5281/zenodo.3945250.

[149] Samuele Ferracin et al. "Efficiently improving the performance of noisy quantum computers". In: *arXiv preprint arXiv:2201.10672* (2022).

[150] Seth T Merkel et al. "Self-consistent quantum process tomography". In: *Physical Review A* 87.6 (2013), p. 062119.

[151] Robin Blume-Kohout et al. "Robust, self-consistent, closed-form tomography of quantum logic gates on a trapped ion qubit". In: *arXiv preprint arXiv:1310.4492* (2013).

[152] Robin Blume-Kohout et al. "Demonstration of qubit operations below a rigorous fault tolerance threshold with gate set tomography". In: *Nature communications* 8.1 (2017), pp. 1–13.

[153] Erik Nielsen et al. "Gate set tomography". In: *Quantum* 5 (2021), p. 557.

[154] Erik Nielsen et al. *Python GST Implementation (PyGSTi) v. 0.9*. Tech. rep. Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2019. DOI: 10.11578/dc.20190722.2.

[155] Erik Nielsen et al. "Probing quantum processor performance with pyGSTi". In: *Quantum Science and Technology* 5.4 (2020), p. 044002. DOI: 10.1088/2058-9565/ab8aa4.

[156] Samuel S Wilks. "The large-sample distribution of the likelihood ratio for testing composite hypotheses". In: *The annals of mathematical statistics* 9.1 (1938), pp. 60–62.

[157] Matthew Ware et al. "Experimental Pauli-frame randomization on a superconducting qubit". In: *Physical Review A* 103.4 (2021), p. 042604.

[158] Kenneth Rudinger et al. "Experimental Characterization of Crosstalk Errors with Simultaneous Gate Set Tomography". In: *PRX Quantum* 2 (4 Nov. 2021), p. 040338. DOI: 10.1103/PRXQuantum.2.040338. URL: https://link.aps.org/doi/10.1103/PRXQuantum.2.040338.

[159] Erik Nielsen et al. "Efficient flexible characterization of quantum processors with nested error models". In: *New Journal of Physics* 23.9 (2021), p. 093020.

[160] Hirotugu Akaike. "A new look at the statistical model identification". In: *IEEE transactions on automatic control* 19.6 (1974), pp. 716–723. DOI: 10.1109/TAC.1974.1100705.

[161] Robin Blume-Kohout et al. "Wildcard error: Quantifying unmodeled errors in quantum processors". In: *arXiv preprint arXiv:2012.12231* (2020).

[162] Lev S Bishop et al. "Quantum volume". In: *Quantum Volume. Technical Report* (2017).

[163] Andrew W Cross et al. "Validating quantum computers using randomized model circuits". In: *Physical Review A* 100.3 (2019), p. 032328.

[164] Timothy Proctor et al. "Scalable randomized benchmarking of quantum computers using mirror circuits". In: *arXiv preprint arXiv:2112.09853* (2021).

[165] Karl Mayer et al. "Theory of mirror benchmarking and demonstration on a quantum computer". In: *arXiv preprint arXiv:2108.10431* (2021).

[166] Sergio Boixo et al. "Characterizing quantum supremacy in near-term devices". In: *Nature Physics* 14.6 (2018), pp. 595–600.

[167] Teague Tomesh et al. "Supermarq: A scalable quantum benchmark suite". In: *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE. 2022, pp. 587–603.

[168] Brad Mitchell et al. "Calibration of the Cross-Resonance Gate using Closed-Loop Optimal Control". In: *APS March Meeting Abstracts*. Vol. 2021. 2021, P33–009.

[169] Jaseung Ku et al. "Suppression of Unwanted Z Z Interactions in a Hybrid Two-Qubit System". In: *Physical review letters* 125.20 (2020), p. 200504.

[170] Peng Zhao et al. "High-Contrast ZZ Interaction Using Superconducting Qubits with Opposite-Sign Anharmonicity". In: *Physical Review Letters* 125.20 (2020), p. 200503.

[171] Pranav Mundada et al. "Suppression of qubit crosstalk in a tunable coupling superconducting circuit". In: *Physical Review Applied* 12.5 (2019), p. 054023.

[172] X Li et al. "Tunable coupler for realizing a controlled-phase gate with dynamically decoupled regime in a superconducting circuit". In: *Physical Review Applied* 14.2 (2020), p. 024070.

[173] Zhongchu Ni et al. "Scalable method for eliminating residual ZZ interaction between superconducting qubits". In: *arXiv preprint arXiv:2111.13292* (2021).

[174] Joel J. Wallman and Joseph Emerson. "Noise tailoring for scalable quantum computation via randomized compiling". In: *Phys. Rev. A* 94 (5 Nov. 2016), p. 052325. DOI: 10.1103/PhysRevA.94.052325. URL: https://link.aps.org/doi/10.1103/PhysRevA.94.052325.

[175] E Knill. "Fault-tolerant postselected quantum computation: Threshold analysis". In: *arXiv preprint quant-ph/0404104* (2004).

[176] Oliver Kern, Gernot Alber, and Dima L Shepelyansky. "Quantum error correction of coherent errors by randomization". In: *The European Physical Journal D-Atomic, Molecular, Optical and Plasma Physics* 32.1 (2005), pp. 153–156.

[177] Michael R Geller and Zhongyuan Zhou. "Efficient error models for fault-tolerant architectures and the Pauli twirling approximation". In: *Physical Review A* 88.1 (2013), p. 012314.

[178] Matthew B Hastings. "Turning gate synthesis errors into incoherent errors". In: *arXiv preprint arXiv:1612.01011* (2016).

[179] Earl Campbell. "Shorter gate sequences for quantum computing by mixing unitaries". In: *Physical Review A* 95.4 (2017), p. 042306.

[180] Earl Campbell. "Random compiler for fast Hamiltonian simulation". In: *Physical Review Letters* 123.7 (2019), p. 070503.

[181] Zhenyu Cai, Xiaosi Xu, and Simon C Benjamin. "Mitigating coherent noise using Pauli conjugation". In: *npj Quantum Information* 6.1 (2020), pp. 1–9.

[182] Chao Song et al. "Quantum computation with universal error mitigation on a superconducting quantum processor". In: *Science Advances* 5.9 (2019), eaaw5686.

[183] Lorenza Viola, Emanuel Knill, and Seth Lloyd. "Dynamical decoupling of open quantum systems". In: *Physical Review Letters* 82.12 (1999), p. 2417.

[184] Akel Hashim et al. "Benchmarking verified logic operations for fault tolerance". In: *arXiv preprint arXiv:2207.08786* (2022).

[185] Marc Grau Davis et al. "Heuristics for Quantum Compiling with a Continuous Gate Set". In: *arXiv preprint arXiv:1912.02727* (2019).

[186] Aram W Harrow and Richard A Low. "Random quantum circuits are approximate 2-designs". In: *Communications in Mathematical Physics* 291.1 (2009), pp. 257–302.

[187] Panos Aliferis, Daniel Gottesman, and John Preskill. "Quantum accuracy threshold for concatenated distance-3 codes". In: *arXiv preprint quant-ph/0504218* (2005).

[188] Panos Aliferis and Andrew W Cross. "Subsystem fault tolerance with the Bacon-Shor code". In: *Physical review letters* 98.22 (2007), p. 220502.

[189] Panos Aliferis, Daniel Gottesman, and John Preskill. "Accuracy threshold for postselected quantum computation". In: *arXiv preprint quant-ph/0703264* (2007).

[190] Panos Aliferis and John Preskill. "Fibonacci scheme for fault-tolerant quantum computation". In: *Physical Review A* 79.1 (2009), p. 012332.

[191] Christopher Chamberland, Tomas Jochym-O'Connor, and Raymond Laflamme. "Thresholds for universal concatenated quantum codes". In: *Physical review letters* 117.1 (2016), p. 010501.

[192] Emanuel Knill. "Quantum computing with realistically noisy devices". In: *Nature* 434.7029 (2005), pp. 39–44.

[193] Panos Aliferis et al. "Fault-tolerant computing with biased-noise superconducting qubits: a case study". In: *New Journal of Physics* 11.1 (2009), p. 013061.

[194] Guillaume Duclos-Cianci and David Poulin. "Fast decoders for topological quantum codes". In: *Physical review letters* 104.5 (2010), p. 050504.

[195] David S Wang, Austin G Fowler, and Lloyd CL Hollenberg. "Surface code quantum computing with error rates over 1%". In: *Physical Review A* 83.2 (2011), p. 020302.

[196] Héctor Bombin et al. "Strong resilience of topological codes to depolarization". In: *Physical Review X* 2.2 (2012), p. 021004.

[197] James R Wootton and Daniel Loss. "High threshold error correction for the surface code". In: *Physical review letters* 109.16 (2012), p. 160503.

[198] Ashley M Stephens. "Fault-tolerant thresholds for quantum error correction with the surface code". In: *Physical Review A* 89.2 (2014), p. 022321.

[199] James M Auger et al. "Fault-tolerance thresholds for the surface code with fabrication errors". In: *Physical Review A* 96.4 (2017), p. 042316.

[200] David K Tuckett et al. "Fault-tolerant thresholds for the surface code in excess of 5% under biased noise". In: *Physical review letters* 124.13 (2020), p. 130501.

[201] Xiao Xue et al. "Quantum logic with spin qubits crossing the surface code threshold". In: *Nature* 601 (2022), pp. 343–347. DOI: https://doi.org/10.1038/s41586-021-04273-w.

[202] Mateusz T Madzik et al. "Precision tomography of a three-qubit electron-nuclear quantum processor in silicon". In: *Nature* 601 (2022), pp. 348–353.

[203] J Watrous. "Simpler semidefinite programs for completely bounded norms, Chicago J. of Th". In: *Comp. Sci* 8 (2013), p. 2013.

[204] Sarah Sheldon et al. "Characterizing errors on qubit operations via iterative randomized benchmarking". In: *Physical Review A* 93.1 (2016), p. 012301.

[205] Ye Wang et al. "High-Fidelity Two-Qubit Gates Using a Microelectromechanical-System-Based Beam Steering System for Individual Qubit Addressing". In: *Physical Review Letters* 125.15 (2020), p. 150505.

[206] Juan M Pino et al. "Demonstration of the trapped-ion quantum CCD computer architecture". In: *Nature* 592.7853 (2021), pp. 209–213.

[207] Jerry M Chow et al. "Universal quantum gate set approaching fault-tolerant thresholds with superconducting qubits". In: *Physical review letters* 109.6 (2012), p. 060501.

[208] Austin G Fowler et al. "Surface codes: Towards practical large-scale quantum computation". In: *Physical Review A* 86.3 (2012), p. 032324.

[209] Chenyang Wang, Jim Harrington, and John Preskill. "Confinement-Higgs transition in a disordered gauge theory and the accuracy threshold for quantum memory". In: *Annals of Physics* 303.1 (2003), pp. 31–58.

[210] Robert Raussendorf, Jim Harrington, and Kovid Goyal. "A fault-tolerant one-way quantum computer". In: *Annals of physics* 321.9 (2006), pp. 2242–2270.

[211] Robert Raussendorf, Jim Harrington, and Kovid Goyal. "Topological fault-tolerance in cluster state quantum computation". In: *New Journal of Physics* 9.6 (2007), p. 199.

[212] Robert Raussendorf and Jim Harrington. "Fault-tolerant quantum computation with high threshold in two dimensions". In: *Physical review letters* 98.19 (2007), p. 190504.

[213] Austin G Fowler, Adam C Whiteside, and Lloyd CL Hollenberg. "Towards practical classical processing for the surface code". In: *Physical review letters* 108.18 (2012), p. 180501.

[214] Feng Lu and Dan C Marinescu. "Quantum error correction of time-correlated errors". In: *Quantum Information Processing* 6.4 (2007), pp. 273–293.

[215] Austin G Fowler and John M Martinis. "Quantifying the effects of local many-qubit errors and nonlocal two-qubit errors on the surface code". In: *Physical Review A* 89.3 (2014), p. 032316.

[216] Ying Li and Simon C Benjamin. "Efficient variational quantum simulator incorporating active error minimization". In: *Physical Review X* 7.2 (2017), p. 021050.

[217] Kristan Temme, Sergey Bravyi, and Jay M Gambetta. "Error mitigation for short-depth quantum circuits". In: *Physical review letters* 119.18 (2017), p. 180509.

[218] Suguru Endo, Simon C Benjamin, and Ying Li. "Practical quantum error mitigation for near-future applications". In: *Physical Review X* 8.3 (2018), p. 031027.

[219] Abhinav Kandala et al. "Error mitigation extends the computational reach of a noisy quantum processor". In: *Nature* 567.7749 (2019), pp. 491–495.

[220] Tudor Giurgica-Tiron et al. "Digital zero noise extrapolation for quantum error mitigation". In: *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)*. IEEE. 2020, pp. 306–316.

[221] Ryan LaRose et al. "Mitiq: A software package for error mitigation on noisy quantum computers". In: *arXiv preprint arXiv:2009.04417* (2020).

[222] Andre He et al. "Zero-noise extrapolation for quantum-gate error mitigation with identity insertions". In: *Physical Review A* 102.1 (2020), p. 012426.

[223] Suguru Endo et al. "Hybrid quantum-classical algorithms and quantum error mitigation". In: *Journal of the Physical Society of Japan* 90.3 (2021), p. 032001.

[224] Youngseok Kim et al. "Scalable error mitigation for noisy quantum circuits produces competitive expectation values". In: *arXiv preprint arXiv:2108.09197* (2021).

[225] Bálint Koczor. "Exponential error suppression for near-term quantum devices". In: *Physical Review X* 11.3 (2021), p. 031057.

[226] Armands Strikis et al. "Learning-based quantum error mitigation". In: *PRX Quantum* 2.4 (2021), p. 040330.

[227] William J Huggins et al. "Virtual distillation for quantum error mitigation". In: *Physical Review X* 11.4 (2021), p. 041036.

[228] Vincent R Pascuzzi et al. "Computationally efficient zero-noise extrapolation for quantum-gate-error mitigation". In: *Physical Review A* 105.4 (2022), p. 042406.

[229] Ryuji Takagi et al. "Fundamental limitations of quantum error mitigation". In: *arXiv e-prints* (2021), arXiv–2109.

[230] Alberto Peruzzo et al. "A variational eigenvalue solver on a photonic quantum processor". In: *Nature Communications* 5.1 (July 2014). ISSN: 2041-1723. URL: http://dx.doi.org/10.1038/ncomms5213.

[231] J. I. Colless et al. "Computation of Molecular Spectra on a Quantum Processor with an Error-Resilient Algorithm". In: *Physical Review X* 8.1 (Feb. 2018). ISSN: 2160-3308. URL: http://dx.doi.org/10.1103/PhysRevX.8.011021.

[232] Sam McArdle et al. "Variational ansatz-based quantum simulation of imaginary time evolution". In: *npj Quantum Information* 5.1 (Sept. 2019). ISSN: 2056-6387. URL: http://dx.doi.org/10.1038/s41534-019-0187-2.

[233] Kübra Yeter-Aydeniz, Raphael C Pooser, and George Siopsis. "Practical quantum computation of chemical and nuclear energy levels using quantum imaginary time evolution and Lanczos algorithms". In: *npj Quantum Information* 6.1 (2020), pp. 1–8. URL: https://doi.org/10.1038/s41534-020-00290-1.

[234] Mario Motta et al. "Determining eigenstates and thermal states on a quantum computer using quantum imaginary time evolution". In: *Nature Physics* 16.2 (2020), pp. 205–210.

[235] Kubra Yeter Aydeniz, George Siopsis, and Raphael C Pooser. "Scattering in the Ising Model with the Quantum Lanczos Algorithm". In: *New Journal of Physics* 23 (2021), p. 043033. URL: https://iopscience.iop.org/article/10.1088/1367-2630/abe63d.

[236] Jean-Loup Ville et al. "Leveraging randomized compiling for the QITE algorithm". In: *arXiv preprint arXiv:2104.08785* (2021).

[237] Miroslav Urbanek et al. "Mitigating depolarizing noise on quantum computers with noise-estimation circuits". In: *Physical Review Letters* 127.27 (2021), p. 270502.

[238] Joseph Vovrosh et al. "Simple mitigation of global depolarizing errors in quantum simulations". In: *Physical Review E* 104.3 (2021), p. 035309.

[239] Sadhan K. Adhikari. "Numerical solution of the two-dimensional Gross–Pitaevskii equation for trapped interacting atoms". In: *Physics Letters A* 265.1-2 (Jan. 2000), pp. 91–96. ISSN: 0375-9601. URL: http://dx.doi.org/10.1016/S0375-9601(99)00878-6.

[240] J Auer, E Krotscheck, and Siu A Chin. "A fourth-order real-space algorithm for solving local Schrödinger equations". In: *The Journal of Chemical Physics* 115.15 (2001), pp. 6841–6846. URL: https://doi.org/10.1142/S0217979203020570.

[241] Wim Lavrijsen et al. "Classical optimizers for noisy intermediate-scale quantum devices". In: *2020 IEEE international conference on quantum computing and engineering (QCE)*. IEEE. 2020, pp. 267–277.

[242] Alba Cervera-Lierta. "Exact Ising model simulation on a quantum computer". In: *Quantum* 2 (2018), p. 114. URL: https://doi.org/10.22331/q-2018-12-21-114.

[243] Shi-Ning Sun et al. "Quantum Computation of Finite-Temperature Static and Dynamical Properties of Spin Systems Using Quantum Imaginary Time Evolution". In: *PRX Quantum* 2.1 (Feb. 2021). ISSN: 2691-3399. URL: http://dx.doi.org/10.1103/PRXQuantum.2.010317.

[244] E. F. Dumitrescu et al. "Cloud Quantum Computing of an Atomic Nucleus". In: *Physical Review Letters* 120.21 (May 2018). ISSN: 1079-7114. URL: http://dx.doi.org/10.1103/PhysRevLett.120.210501.

[245] Shuaining Zhang et al. "Error-mitigated quantum gates exceeding physical fidelities in a trapped-ion system". In: *Nature communications* 11.1 (2020), pp. 1–8.

[246] Xavi Bonet-Monroig et al. "Low-cost error mitigation by symmetry verification". In: *Physical Review A* 98.6 (2018), p. 062339.

[247] Sam McArdle, Xiao Yuan, and Simon Benjamin. "Error-mitigated digital quantum simulation". In: *Physical review letters* 122.18 (2019), p. 180501.

[248] David C McKay et al. "Correlated randomized benchmarking". In: *arXiv preprint arXiv:2003.02354* (2020).

[249] Sergey Bravyi et al. "Mitigating measurement errors in multiqubit experiments". In: *Physical Review A* 103.4 (2021), p. 042605.

[250] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. "A quantum approximate optimization algorithm". In: *arXiv preprint arXiv:1411.4028* (2014).

[251] Edward Farhi and Aram W Harrow. "Quantum supremacy through the quantum approximate optimization algorithm". In: *arXiv preprint arXiv:1602.07674* (2016).

[252] Teague Tomesh et al. "Coreset clustering on small quantum computers". In: *Electronics* 10.14 (2021), p. 1690.

[253] Vinay Tripathi et al. "Suppression of crosstalk in superconducting qubits using dynamical decoupling". In: *arXiv preprint arXiv:2108.04530* (2021).

[254] Guang Hao Low, Theodore J. Yoder, and Isaac L. Chuang. "Optimal arbitrarily accurate composite pulse sequences". In: *Phys. Rev. A* 89 (2 Feb. 2014), p. 022341. DOI: 10.1103/PhysRevA.89.022341. URL: https://link.aps.org/doi/10.1103/PhysRevA.89.022341.

[255] Christophe Piveteau et al. "Error mitigation for universal gates on encoded qubits". In: *Physical Review Letters* 127.20 (2021), p. 200505.

[256] Yasunari Suzuki et al. "Quantum error mitigation as a universal error-minimization technique: Applications from NISQ to FTQC eras". In: *arXiv preprint arXiv:2010.03887* (2020).

[257] Jesse Fern et al. "Generalized performance of concatenated quantum codes—a dynamical systems approach". In: *IEEE transactions on automatic control* 51.3 (2006), pp. 448–459.

[258] Mauricio Gutiérrez et al. "Errors and pseudothresholds for incoherent and coherent noise". In: *Physical Review A* 94.4 (2016), p. 042338.

[259] Andrew S Darmawan and David Poulin. "Tensor-network simulations of the surface code under realistic noise". In: *Physical review letters* 119.4 (2017), p. 040502.

[260] Daniel Greenbaum and Zachary Dutton. "Modeling coherent errors in quantum error correction". In: *Quantum Science and Technology* 3.1 (2017), p. 015007.

[261] Juan P Dehollain et al. "Optimization of a solid-state electron spin qubit using gate set tomography". In: *New Journal of Physics* 18.10 (2016), p. 103018.

[262] AC Hughes et al. "Benchmarking a high-fidelity mixed-species entangling gate". In: *Physical Review Letters* 125.8 (2020), p. 080504.

[263] Gregory AL White, Charles D Hill, and Lloyd CL Hollenberg. "Performance optimization for drift-robust fidelity improvement of two-qubit gates". In: *Physical Review Applied* 15.1 (2021), p. 014023.