THE UNIVERSITY OF CHICAGO

TOPICS IN PAULI CHANNEL LEARNING: QUANTUM ADVANTAGES, QUANTUM NOISE CHARACTERIZATION, AND QUANTUM ERROR MITIGATION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE PRITZKER SCHOOL OF MOLECULAR ENGINEERING
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
SENRUI CHEN

CHICAGO, ILLINOIS
JUNE 2025

ABSTRACT

This dissertation investigates fundamental and practical aspects of quantum information science through the lens of Pauli channels — a simple yet powerful class of quantum processes. Three interconnecting topics will be discussed: quantum advantages, quantum noise characterization, and quantum error mitigation. (1) We introduce a novel type of quantum advantages in Pauli channel learning, showing that entanglement with ancillary quantum memory provides exponential advantages in sample complexity for learning a Pauli channel, shedding new light on the role of entanglement as a quantum resource. (2) We develop a comprehensive theory and novel protocols on Pauli noise characterization in quantum computers, which are self-consistent, complete, and scalable. We also show how entanglement can be used to accelerate noise characterization, even in the presence of imperfect quantum memory. (3) Leveraging our self-consistent Pauli noise characterization framework, we propose and experimentally demonstrate a gauge-consistent quantum error mitigation protocol, resolving an outstanding issue of noise non-identifiability in standard error mitigation methods. Altogether, this dissertation establishes Pauli channels as a theoretically interesting and practically impactful object in the research of quantum information science.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor Prof. Liang Jiang for being a tremendous source of support and inspiration throughout my PhD journey. I am grateful to Liang for giving me the freedom to explore topics of my interest, while always being there with insightful suggestions whenever I faced challenges. I have learned a lot from him – not only about science and physics, but also about how to be a good researcher and how to survive in academia.

There are many other professors and senior researchers who have provided me with invaluable guidance and support. I would like to thank Prof. Steve Flammia for advising me during my internship and for many fun collaborations on hacking quantum noise; Prof. John Preskill for hosting my visit in Caltech and encouraging me to pursue new research directions; Prof. Bill Fefferman for teaching me a lot about random circuit sampling and many other things in theoretic computer science; Dr. Robin Blume-Kohout and Dr. Kevin Young for inviting me to participate in the Assessing Performance of Quantum Computers (APQC) workshop, which is a great experience. Prof. Hannes Bernien, Prof. Norbert Linke, Prof. Andrew Cleland, Prof. Ulrik Lund Andersen, Prof. Jonas Schou Neergaard-Nielsen for offering experimental insights from various platforms. To all of you – and many others – I sincerely appreciate your support for me as an early-career researcher.

I would also like to thank all my colleagues and friends along my academic journey, including Connor Hann, Kyungjoo Noh, Sisi Zhou, Qian Xu, Yat Wong, Ming Yuan, Guo (Jerry) Zheng, Gideon Lee, Su-un Lee, Han Zheng, Changhun Oh, Pei Zeng, Alireza Seif and others from the Jiang group; Zlatko Minev, Edward Chen, Laurin Fischer, Haoran Liao, Swarnadeep Majumder, Kristan Temme, and others from IBM Quantum; and Hsin-Yuan (Robert) Huang, Yunchao Liu, Zhihan Zhang, Yingyue Zhu, Akel Hashim, Noah Goss, Ze-Pei Cian and many others from different institutes. Without all of you, this journey could never be as fun and fruitful.

I am thankful to the Pritzker School of Molecular Engineering at the University of Chicago for fostering a interdisciplinary environment that allowed me to engage with people with diverse scientific backgrounds and interests. These interactions have significantly shaped my research directions. I am also thankful to all the staff members who supported me through various administrative processes.

Finally, I would like to thank my wife, Sisi Zhou, who has been a significant part of my life. Together, we have shared the joys and challenges of this journey, both in the academic world and in life. I am truly grateful for her companionship. I am also deeply thankful to my parents for supporting me in pursuing my dream of science.

CHAPTER 1

INTRODUCTION

Quantum Information Science (QIS) — The usage of quantum mechanics in information processing — holds the promise to revolutionize our ways to interact with the nature. Arguably, the most fundamental question in QIS is to identify quantum advantages, i.e., for which tasks quantum resources can bring significant advancement over classical information processing methods. Such quantum advantages have been theoretically explored in many aspects, including computing with unprecedented efficiency [NC11, Sho99, AAB$^+$19], communication with unprecedented security [GT07, Kim08], and sensing with unprecedented accuracy [GLM06, GLM11]. It is still an on-going quest to understand the power and limitation of QIS and how it can be applied to solve problems that are of real-world relevance.

Along with its extraordinary power, QIS comes with a unique challenge. That is, being highly susceptible to noise and error. Fortunately, the paradigm of quantum error correction (QEC) and fault-tolerant quantum computation (FTQC) [Sho96, ABO08] gives a road map towards noise-robust quantum information processing. Over the past few decades, QEC has achieved significant advancement both in theory (e.g., design of good codes and novel fault-tolerant architecture) [PK22, XBAP$^+$24, BCG$^+$24] and in experiment (e.g., demonstration of early-stage fault-tolerant quantum computation on various experimental platforms including superconducting circuits, neutral atom arrays, trapped ions) [BEG$^+$24, AAAB$^+$25, PBP$^+$24]. Still, quantum noise remains as the major obstacle that blocks our way towards scalable and robust quantum applications. It requires experimental and theoretical co-design to overcome this challenge.

The goal of this thesis is to explore the challenges, opportunities, and methodologies of modern quantum information science, through the lens of *Pauli channel*. Pauli channels are

defined as a probabilistic mixture of Pauli operators. Mathematically,

$$\Lambda(\rho) = \sum_{P \in \{I,X,Y,Z\}^{\otimes n}} p_P P \rho P^\dagger. \tag{1.1}$$

Here $\{I, X, Y, Z\}^{\otimes n}$ denote all n-qubit Pauli operators, i.e., n-fold tensor products of standard single-qubit Pauli operators. p_P is a probability distribution over n-qubit Pauli operators. Detailed definitions are given in Chapter 2. Pauli channels are one of the simplest families of quantum processes. Yet, they are able to capture some of the most fundamental aspects of quantum information science, including the notion of quantum incompatibility. They have also found wide applications in the theory and practice of quantum noise robustness.

This thesis will study the roles of Pauli channels in the following three subareas of QIS. A roadmap is given in Fig. 1.1

1. **Quantum Learning Advantages**. One particularly interesting class of information processing tasks are learning from the nature. This include, for example, searching for dark matter, detecting gravitational waves, probing structures of certain materials, etc. It is interesting to ask whether quantum protocols can significantly speed up certain learning tasks over their classical counterpart. If so, what is the enabling quantum resources behind such advantages. Chapter 3 of this thesis will explore a novel type of quantum advantage in *Pauli channel learning*, where quantum entanglement — a characteristic quantum resource — is shown to be the source of advantages. Using information-theoretic methods, we rigorously obtain tight sample complexity bound for learning Pauli channel using entanglement with an ancillary quantum memory ("quantum protocol") or not using entanglement ("classical protocol"), and establish an exponential separation between them. Notably, the bounds for classical protocol obtained here are valid against the most-general entanglement-free schemes, allowing

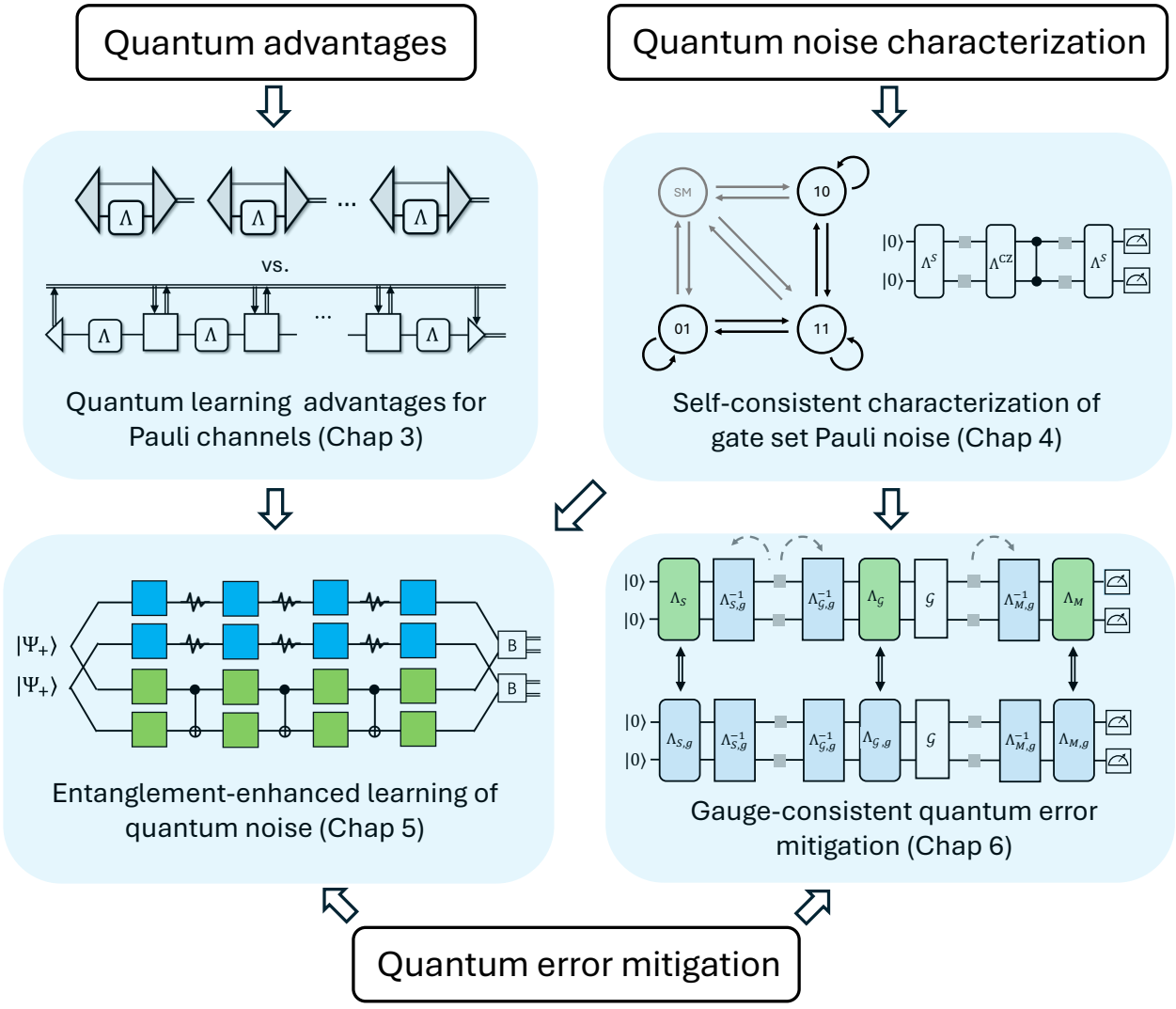


Figure 1.1: Thesis roadmap.

adaptive control, concatenation of channels, and even mid-circuit measurement and feed-forward control. Finally, we show this learning advantage can be readily seen in experiments and have good noise resilience, making them not only fundamentally interesting but also practically relevant.

2. **Quantum Noise Characterization.** Since noise is the biggest practical challenge for QIS, it is important to gives a precise characterization about the noise. In fact, the most well-known application of Pauli channel is in quantum noise modeling. On the one

hand, Pauli channel naturally describe a wide range of common noise processes, such as bit-flip, phase-flip, and depolarizing noise. On the other hand, techniques known as randomized compiling [WE16, HNM+21] exists that can tailor general noise processes into Pauli channels (given high-quality single-qubit quantum gates). This motivates the study of *Pauli noise model*, which roughly speaking models both gate noise and state-preparation-and-measurement (SPAM) noise by Pauli channels. Pauli noise model has been applied in quantum benchmarking, quantum error mitigation, and for quantum error correction. Despite intensive study, some core aspects of Pauli noise is by far not well understood. In particular, due to the existence of unknown SPAM noise, it is unclear how to fully characterize the Pauli channels associated with the noisy gates. State-of-the-art Pauli noise learning protocols can only characterize the gate-dependent Pauli channels up to "degeneracies". Chapter 4 of this thesis fills this gap by introducing a theory that fully characterizes the "learnability" of a Pauli noise model, i.e., which noise parameters are identifiable and which are not. The characterization is achieved using ideas from gate-set tomography and techniques from algebraic graph theory. By further generalizing the theory to allow locality constraints on the noise channels, we developed an efficient self-consistent protocol to learn Pauli noises over a gate set. The relevance of the above theory and protocols are experimentally validated on superconducting quantum circuits.

It turns out that the quantum learning advantages discussed above also find their applications in quantum noise characterization, as reported in Chapter 5. Most existing Pauli noise characterization schemes use no ancillary quantum memory, which means they generally suffers from an exponential sample overhead if no assumption on the noise structure is made (which is a corollary of our theorems in Chapter 3). In contrast, we design an entanglement-enhanced learning protocol that utilizes entanglement with ancillary quantum memory to greatly speedup the noise characterization. Given ideal

quantum memory, this protocol provably breaks the exponential sample complexity barrier. For noisy quantum memory in lab, we develop novel techniques to suppress and mitigate the errors, with an significantly smaller overhead compared to the best possible entanglement-free learning protocols. The advantages has been experimentally demonstrated on a superconducting quantum computers. This series of works opens up new possibilities for practical quantum learning advantages as well as for quantum noise characterization.

3. **Quantum Error Mitigation.** With the knowledge of noise, one can design methods to suppress, mitigate, or correct the errors in a quantum system. One well-known example of such methods (besides quantum error correction) is known as quantum error mitigation. Recently, quantum error mitigation protocols based on Pauli noise model has been developed and experimentally implemented. The basic idea is to first characterize the Pauli noise channel, then either cancel the noise channel via quasi-probability sampling, or amplify the noise and extrapolate back to zero-noise point. However, due to the non-identifiability of Pauli noise model discussed above, existing methods have to rely on unjustified assumptions to determine the noise channels for the protocol to be applied. Built on our theory about the learnability of Pauli noise, Chapter 6 of the thesis resolves this issue by developing an *gauge-consistent error mitigation* procedure for Pauli noise. By learning all the learnable degrees of freedom of a Pauli noise model, and arbitrarily fixing the unlearnable degrees of freedom (called *gauge parameters*), we can correctly mitigate the errors. One can even optimize over the gauge parameters to minimize the sampling overhead. Furthermore, the whole procedure is scalable given appropriate locality constraints of the noise process. In an experimental demonstration, we show that for situations where existing error mitigation scheme suffers from severe bias, our gauge-consistent protocol remains accurate and efficient.

The above covers non-exclusively the theory and practice of Pauli channels in QIS, in-

cluding the exploration of useful quantum advantages and for how to tackle with quantum noise. These aspects are deeply connected with each other. They also represent an theory-experiment co-design on different types of experimental platforms. I hope this thesis provides a unique and timely perspective into opportunities and challenges of QIS, which is at its turning point from theoretic fantasy into real-world usefulness.

CHAPTER 2

PRELIMINARIES

This chapter provide some basic definitions and preliminaries which will be used throughout this thesis. More definitions and backgrounds will be given in the relevant sections where they will be used.

Quantum information basics. Here, we review some basics of quantum information. More details can be found in standard textbooks such as [NC11]. Throughout this thesis, we consider quantum systems consists of qubits. Let $\mathcal{H}_{2^n} = \mathcal{H}_2^{\otimes n}$ denote the n-qubit Hilbert space (which is an 2^n-dimensional complex linear space equipped with standard inner product). An n-qubit quantum state can be describe by a density operator ρ acting on $\mathcal{H}_{2^n}$ that is positive semidefinite ($\rho \geq 0$) and trace-one ($\operatorname{Tr}\rho = 1$). A quantum channel $\mathcal{E}^{A\to B}$ from quantum system A to B is a *linear* map from $L(\mathcal{H}_A)$ to $L(\mathcal{H}_B)$, where $L(\mathcal{H})$ denotes the space of all linear operators on $\mathcal{H}$, such that $\mathcal{E}$ is

1. Completely positive (CP): $\mathcal{E}^{A\to B} \otimes \mathrm{id}^R(\rho^{AR}) \geq 0$ for any system R and any $\rho^{AR} \geq 0$.

2. Trace-preserving (TP): $\operatorname{Tr}(\mathcal{E}(\rho)) = \operatorname{Tr}(\rho)$ for any $\rho \in L(H_A)$.

Intuitively, the CPTP conditions ensures a quantum channel transform any quantum state to a valid quantum state. We generally omit the superscripts when the underlying quantum systems are clear from the context. It is known that any quantum channel admits a Kraus representation, $\mathcal{E}(\cdot) = \sum_j K_j(\cdot)K_j^\dagger$, for a set of operator $\{K_j\}_j$ satisfying $\sum_j K_j^\dagger K_j = I$ where I is the identity operator. Conversely, any map with a Kraus representation is a quantum channel.

Pauli group and Pauli channel. Define the four single-qubit Pauli operators with their matrix representation in the computational basis as

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{2.1}$$

For an n-qubit Hilbert space, define P^n to be the Pauli group modulo the non-physical phase. P^n is an Abelian group isomorphic to $\mathbb{Z}_2^{2n}$, so we can use elements of $\mathbb{Z}_2^{2n}$ to uniquely label elements of P^n. Specifically, we view every $a \in \mathbb{Z}_2^{2n}$ as a $2n$-bit string $a = a_{x,1}a_{z,1}a_{x,2}a_{z,2}\cdots a_{x,n}a_{z,n}$ corresponding to the Pauli operator

$$P_a = \otimes_{k=1}^{n} i^{a_{x,k}a_{z,k}} X^{a_{x,k}} Z^{a_{z,k}} \tag{2.2}$$

where the phase is chosen to ensure Hermiticity. We also define a sympletic inner product $\langle \cdot, \cdot \rangle$ within $\mathbb{Z}_2^{2n}$ as

$$\langle a, b \rangle = \sum_{k=1}^{n} (a_{x,k}b_{z,k} + a_{z,k}b_{x,k}) \mod 2.$$

One can verify that $P_aP_b = (-1)^{\langle a,b \rangle} P_bP_a$ [NC11].

An n-qubit Pauli channel Λ is a quantum channel of the following form

$$\Lambda(\cdot) = \sum_{a \in \mathbb{Z}_2^{2n}} p_a P_a(\cdot)P_a, \tag{2.3}$$

where $\boldsymbol{p} \coloneqq \{p_a\}_a$ is called the *Pauli error rates*. An important property of Pauli channels is that their eigen-operators are exactly the 4^n Pauli operators. Thus, an alternative expression for Λ is

$$\Lambda(\cdot) = \frac{1}{2^n} \sum_{b \in \mathbb{Z}_2^{2n}} \lambda_b \operatorname{Tr}(P_b(\cdot))P_b, \tag{2.4}$$

where $\boldsymbol{\lambda} \coloneqq \{\lambda_b\}_b$ is called the *Pauli eigenvalues* [FW20, FO21]. These two sets of parameters,

$\boldsymbol{p}$ and $\boldsymbol{\lambda}$, are related by the Walsh-Hadamard transform

$$\lambda_b = \sum_{a\in\mathbb{Z}_2^{2n}} p_a(-1)^{\langle a,b\rangle}, \quad p_a = \frac{1}{4^n}\sum_{b\in\mathbb{Z}_2^{2n}} \lambda_b(-1)^{\langle a,b\rangle}. \tag{2.5}$$

Both $\boldsymbol{p}$ and $\boldsymbol{\lambda}$ are physically interesting parameters: The Pauli error rates are directly related to the error thresholds in fault-tolerant quantum computation [Sho96, ABO08] and have been the quantities of interest for many quantum benchmarking protocols [FW20, HFW20, HYF21a, FO21, LOB+21]; The Pauli eigenvalues quantify how well a Pauli observable is preserved through the noise channel (hence also known as *Pauli fidelities*) and have applications in quantum error mitigation (see *e.g.* [CYZF21]).

We will also use the *Pauli-transfer-matrix* (PTM) representation to simplify notations. A linear operator O acting on a 2^n-dimensional Hilbert space can be viewed as a vector in a 4^n-dimensional Hilbert space. We denote this vectorization of O as $|O\rangle\!\rangle$ and the corresponding Hermitian conjugate as $\langle\!\langle O|$. The inner product within this space is the *Hilbert-Schmidt* product defined as $\langle\!\langle A|B\rangle\!\rangle \coloneqq \mathrm{Tr}(A^\dagger B)$. The *normalized Pauli operators* $\{\sigma_a \coloneqq P_a/\sqrt{2^n},\ a \in \mathbb{Z}_2^{2n}\}$ forms an orthonormal basis for this space. In the PTM representation, a superoperator (*i.e.*, quantum channel) becomes an operator acting on the 4^n-dimensional Hilbert space, sometimes called the Pauli transfer operator. Explicitly, we have $|\Lambda(\rho)\rangle\!\rangle = \Lambda^{\mathrm{PTM}}|\rho\rangle\!\rangle \equiv \Lambda|\rho\rangle\!\rangle$, where we use the same notation to denote a channel and its Pauli transfer operator, which should be clear from the context. Specifically, a general Pauli channel Λ has the following Pauli transfer operator

$$\Lambda = \sum_{a\in\mathbb{Z}_2^{2n}} \lambda_a |\sigma_a\rangle\!\rangle\langle\!\langle\sigma_a|.$$

Two additional definitions about Pauli channels will often be used. We define the *pattern* of an n-qubit Pauli operator P_a, denoted by $\mathrm{pt}(a)$, to be an n-bit string that takes 0 at its ith entry if $P_{a_i} = I$ and takes 1 if $P_{a_i} \in \{X, Y, Z\}$. For example, $\mathrm{pt}(XYIZI) = 11010$. The pattern of a Pauli operator is equivalent to its (non-trivial) support. If a Pauli channel Λ

satisfies that λ_a depends only on $\mathrm{pt}(a)$ for all a, we call it a *symmetric Pauli channel* or *generalized depolarizing channel* as it does not distinguish between X, Y, or Z.

CHAPTER 3

QUANTUM LEARNING ADVANTAGES FOR PAULI CHANNELS

Chapter Notes. This chapter contains a series of works of mine that explore quantum advantages in learning Pauli channels, mainly based on [CZSJ22, COZ+24a].

3.1 Introduction

One important challenge for the Noisy Intermediate-Scale Quantum (NISQ) era [Pre18] is to demonstrate quantum advantages on near-term devices. Recent works have made groundbreaking progress towards demonstrating *quantum computational advantages* [HM17, AAB+19, ZWD+20, WBC+21], which means quantum computers can efficiently solve certain computational problems outside the reach of the most advanced classical computers. However, computation is not the only aspect quantum computers can achieve meaningful advantages. Learning is yet another possibility, where the basic question is whether quantum computers (with resources such as *quantum memory* and *entangled measurements*) can help us learn certain properties of a physical system more efficiently. This kind of quantum advantages have been explored by several recent works [BCL20, ACQ22, HKP21, CLO21, RYCS21] with positive examples including mixedness testing [BCL20], unitarity testing [ACQ22], Pauli expectation values estimation [HKP21], *etc.* However, most of these quantum advantages proposals so far either focus on artificial problems or have no noise-resilient implementation. It is thus highly desirable to identify a practically-interesting learning task that can be used to demonstrate a robust quantum advantage. Another interesting theoretic open question is to identify the source of quantum learning advantages from familiar quantum resources [CG19], such as quantum entanglement.

A particularly interesting type of learning tasks, known as *quantum noise characteriza-*

tion or *quantum benchmarking* [EHW$^+$20], aims at characterizing the noise on a quantum device. This is yet another challenge for the NISQ era which is crucial to building better quantum hardware. Pauli noise is one of the most important quantum noise models: On the one hand, it can describe a wide range of incoherent noise, including dephasing, deplorizing, bit-flip, etc. On the other hand, the recently developed "randomized compiling" technique [WE16, HNM$^+$21] can tailor any noise model in a universal gate set into Pauli noise. Many benchmarking protocols also rely on twirling the noise into a Pauli channel before extracting any information [EWP$^+$19, MGE11, HFW20, LOB$^+$21]. Because of the important role played by Pauli channels, it is of great interest to study the estimation of these objects in an efficient and practical way. Despite a long line of research [FI03, Hay10, CRV$^+$11, RVH12, Col13, FW20, HYF21a, FO21, MRL08, FO21], the ultimate sample complexity for Pauli channel estimation has not yet been fully characterized.

In this section, we present an *exponential* quantum advantage for Pauli channel learning. We show that, for the task of estimating the eigenvalues of an n-qubit Pauli channel (*i.e.*, *Pauli eigenvalues*) to additive error ε in l_∞ distance, there exists a measurement protocol assisted with an n-qubit ancilla that succeeds with high probability using $\mathcal{O}(n/\varepsilon^2)$ samples, while any entanglement-free protocol (possibly with adaptive control, concatenation, mid-circuit measurements, or feed-forward control) would require at least $\Omega(2^n/\varepsilon^2)$ rounds of measurements. As a byproduct, this provides a lower bound for randomized benchmarking (RB) [MGE11, KLR$^+$08] based Pauli noise estimation protocol, resolving an open problem raised in [FW20].

Furthermore, we study the sample efficiency advantages provided by a restricted amount of ancilla. While an ancilla larger than n qubits will not help further improve the efficiency, given a k-qubit ($0 \leq k \leq n$) ancilla, we obtain a lower bound of $\Omega(2^{(n-k)/3})$ for any non-concatenating protocol, and a stronger lower bound of $\Omega(n2^{n-k})$ for non-adaptive and non-concatenating protocols. The latter is shown to be tight by an explicitly constructed

Schemes	Upper bound	Lower bound
Ancilla-free	$O(n2^n/\varepsilon^2)$	$\Omega(2^n/\varepsilon^2)$, $\forall\varepsilon \leq 1/6$
k-qubit ancilla, concatenation-free	$O(n2^{n-k}/\varepsilon^2)$	$\Omega(2^{n-k})$, for $\varepsilon = 1/2$
Unbounded ancilla	$O(n/\varepsilon^2)$	$\Omega(n)$, for $\varepsilon = 1/2$

Table 3.1: Summary of contributions: sample complexity for learning Pauli channels. The task we consider is to estimate all Pauli eigenvalues of any n-qubit Pauli channel to ε additive precision with at least 2/3 success probability. All of these can be found in Theorem 3.1, 3.2, 3.3.

protocol (see Algorithm 1).

3.2 Models of quantum learning

To study the advantages provided by different kinds of quantum resources, we categorize measurement strategies into ancilla-assisted *vs.* ancilla-free, concatenating *vs.* non-concatenating, and adaptive *vs.* non-adaptive measurements. See Fig. 3.1 and explanations in the captions. We remark that, recent works on quantum advantages in property learning [HKP21, ACQ22] have been focusing on the difference between the full-fledged entangled measurement Fig. 3.1(a) and the un-entangled measurement Fig. 3.1(d). Here, we introduce two intermediate measurement models Fig. 3.1(b) and Fig. 3.1(c) to separately study the role of ancilla and concatenation. There are also practical reasons to care about those intermediate models. As an example, the ancilla-free concatenating measurement model Fig. 3.1(c) exactly describes most existing RB protocols (see *e.g.* [HRO$^+$20]), including the RB-type Pauli channel estimation protocol in [FW20]. Indeed, we show that it is the ancilla that provides an exponential advantage in sample complexity for Pauli channel estimation, while concatenation does not help improve the sample efficiency. These results further advances our understanding of quantum advantages in property learning [HKP21, ACQ22, BCL20, CLO21].

Regarding these measurement models, one further question to ask is whether a restricted amount of ancilla can provide any sample efficiency advantages. In this chapter, we will characterize the advantages provided by a k-qubit ancilla ($0 \leq k \leq n$) for the non-concatenating

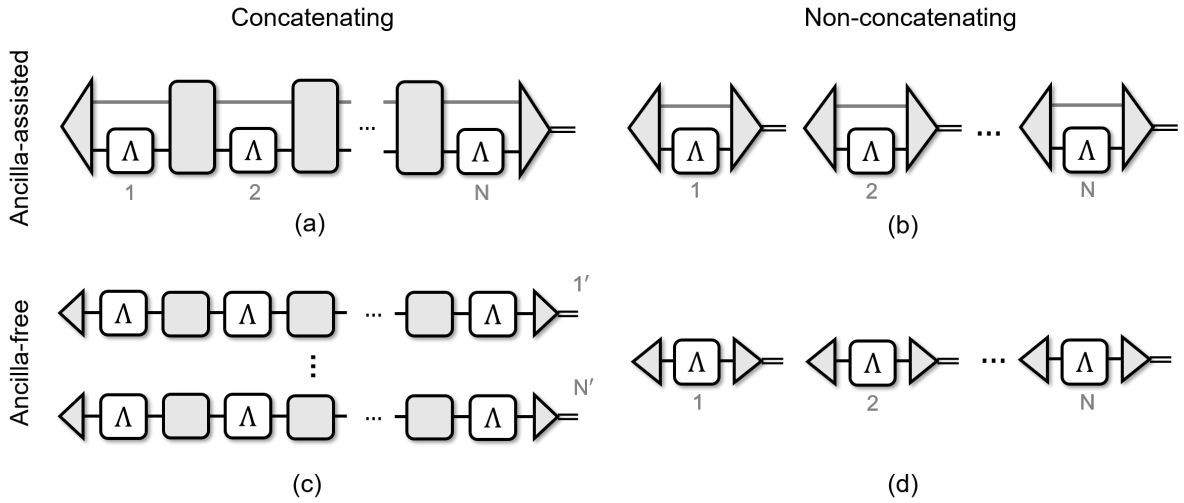


Figure 3.1: Different learning models on N copies of an unknown channel Λ. The gray triangles denote state preparation and measurement, and the gray boxes denote some known processing channels. (a) *Fully entangled measurement*: the most general way to measure a quantum channel where arbitrarily large entanglement and quantum memory is allowed; (b) *Ancilla-assisted non-concatenating measurement*: For each sample of Λ, one input some entangled state and conduct an entangled measurement. The measurement must be completely destructive, which means no quantum memory is allowed; (c) *Ancilla-free concatenating measurement*: One is allowed to sequally apply multiple copies of Λ and some processing channels before applying a single round of measurement, with no ancilla allowed. Here N' denotes the number of measurement rounds which is no larger than the number of samples; (d) *Un-entangled measurement*: Neither ancilla nor concatenation is allowed. Additionally, measurement models (b), (c), (d) can be either *adaptive* or *non-adaptive*. If the measurement setting at a certain round depends on previous measurement outcomes, it is an adaptive protocol. Otherwise, it is non-adaptive.

measurement models. This kind of quantitative trade-off relations between quantum resources and sample efficiency has not been explored in previous works [HKP21, ACQ22, BCL20, CLO21] and may potentially lead to a new resource-theoretical interpretation of quantum entanglement [CG19].

Alternative definitions of entanglement-free schemes. To study the role of quantum entanglement in quantum learning advantages, we introduce two alternative classes of learning schemes as follows [COZ$^+$24a], see Fig. 3.2.

Consider the task of learning properties of an n-qubit quantum channel Λ from certain

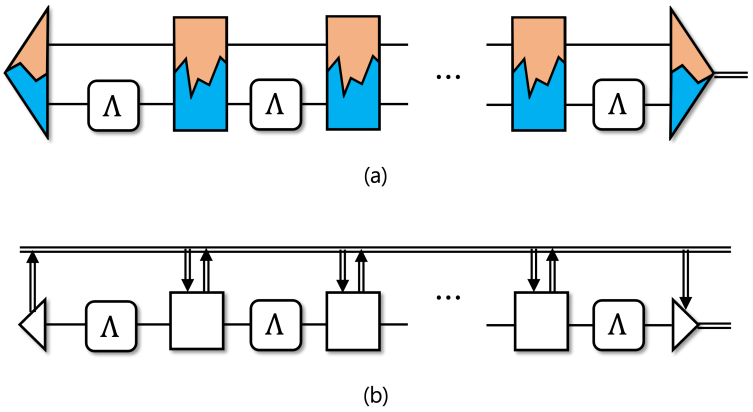


Figure 3.2: Alternative definitions of entanglement-free schemes. (a) Separable schemes. The two different colors indicate the operations are separable. (b) Classical-memory-assisted schemes. The double line represents classical registers. The square box represents adaptively-chosen (arrows coming from the classical registers) quantum instruments (outcomes sent to the classical registers). We will show the two schemes are equivalent in terms of sample complexity, so we call both entanglement-free schemes.

family by querying multiple copies of the channel. We start by defining entanglement-free learning schemes. We introduce the class of *separable schemes*, where a main system $\mathcal{H}_S$ and an ancillary system $\mathcal{H}_A$ are given. The main system is where Λ acts on and has a fixed dimension of 2^n, while the ancillary system can be arbitrarily large. Now, a separable scheme allows interleaving copies of Λ on $\mathcal{H}_S$ with any processing operations (including state preparation, measurement, and quantum channels) on $\mathcal{H}_S \otimes \mathcal{H}_A$, with the only restriction that all the processing operations are separable [CDKL01] across $\mathcal{H}_A$ and $\mathcal{H}_S$. A schematic for separable schemes is shown in Fig. 3.2 (a). The separable operations are known to be the largest set of operations that do not generate entanglement from un-entangled states (even when acting only on a subspace) [CDKL01], and is thus a suitable model for entanglement-free strategies from a quantum-resource-theoretic [CG19] point of view. Furthermore, separable operations contain as a subset other physically-motivated classes of operations like LOCC (local operation and classical communications) [BBC$^+$93, CLM$^+$14], which is also a standard choice of free operation in the resource theory of entanglement [BBPS96, BDSW96, VPRK97]. A lower bound on sample complexity for the former implies a lower bound for the latter.

Besides separable schemes, we introduce another operationally motivated class of schemes called *classical-memory-assisted schemes*. Here, one can only access the main system $\mathcal{H}_S$ (with a fixed dimension of 2^n) and arbitrarily many classical registers. The allowed operations are to interleave copies of Λ with adaptively-chosen *quantum instruments*, which are defined as quantum channels associated with outputs to classical registers. By "adaptively" we mean the quantum instruments can be chosen according to the classical registers. A schematic is given in Fig. 3.2 (b). One can think of such schemes as quantum circuits assisted by mid-circuit measurement and classical feedforward control. We remark that the classical-memory-assisted schemes include the ancilla-free concatenating scheme introduced in [CZSJ22] as a special case, which can describe most randomized benchmarking (RB) [EAŻ05, KLR$^+$08] type protocols. The lower bounds obtained in this thesis thus also holds for those protocols.

While the above two schemes are introduced with different motivations, perhaps surprisingly, they are equivalent in terms of sample complexity. We have the following result.

Proposition 3.1. *For any separable scheme A, there exists a classical-memory-assisted scheme B that generates the same outcome distribution as A for any underlying* Λ *using the same number of copies. Vice versa.*

The formal definitions of both schemes, rigorous statement of Proposition 3.1, and the proof are given in Appendix A. Thanks to Proposition 3.1, we see that both schemes capture the power of learning without entanglement, and be treated interchangeably when studying the sample complexity. In the remaining part of this chapter, we will focus on the classical-memory-assisted scheme, which has a clearer operational meaning. Specifically, there can be two different notions of complexity: One is the sample complexity N_{samp}, which is the number of copies of Λ; The other is the number of measurements N_{meas}, which is the number of quantum instruments with non-trivial measurement. Clearly, $N_{\text{samp}} \geq N_{\text{meas}}$, as one can concatenate multiple copies of Λ and only make one measurement, just like in RB. The lower

bound we derive later will hold for N_{meas}.

3.3 Upper bounds

Our goal is to estimate the Pauli eigenvalues $\boldsymbol{\lambda}$ of an n-qubit Pauli channel Λ to ε precision in l_∞ distance, *i.e.*, estimating each λ_a to ε additive error.

When an n-qubit ancillary system is available, a simple protocol is as follows: prepare n Bell pairs, input one qubit from each pair to the Pauli channel, and apply a Bell measurement on the output. Since the Pauli channel can be viewed as randomly applying one of the 4^n Pauli operators P_a with probability p_a, and each P_a is mapped to a unique measurement outcome[1], we are effectively sampling from the probability distribution $\boldsymbol{p}$. The sample of $\boldsymbol{p}$ can then be used to construct an estimator for $\boldsymbol{\lambda}$ according to the Walsh-Hadamard transform. As shown in Theorem 3.1, this protocol has sample complexity $\mathcal{O}(n)$; on the other hand, when no ancilla is allowed, there is no way to sample from $\boldsymbol{p}$ and simultaneously estimate all elements of $\boldsymbol{\lambda}$ from a single measurement setting. Intuitively, this is what makes the task difficult.

In the following, we give a unified estimation protocol using k ancilla qubits for $0 \leq k \leq n$. Roughly speaking, we divide the n qubits of the Pauli channel into two disjoint subsystems containing k and $n-k$ qubits, respectively, and deal with them separately. For the first subsystem, we introduce a k-qubit ancilla, input k Bell pairs to both systems, and apply a Bell measurement; For the second subsystem, we will use *stabilizer states* input and *syndrome measurements* to be defined later. The measurement scheme is depicted in Fig. 3.3.

A rigorous description of the protocol is given in Algorithm 1. Here, $|\Psi_v\rangle$ represents the Bell states on the $2k$-qubit subsystem defined as

$$|\Psi_v\rangle = P_v \otimes I \, |\Psi^+\rangle \,, \quad |\Psi^+\rangle = \frac{1}{2^k} \sum_{i=0}^{2^k-1} |i\rangle \, |i\rangle \,. \tag{3.1}$$

1. This can be viewed as a consequence of the well-known *superdense coding* protocol [BW92].

A *stabilizer group* S on the $(n-k)$-qubit subsystem is a group of 2^{n-k} commuting Pauli operators. Mathematically, S can be viewed as an $(n-k)$-dimensional subspace of $\mathbb{Z}_2^{2(n-k)}$. The *stabilizer states* $|\phi_e^{\mathsf{S}}\rangle$ are the simultaneous eigenstates of all Pauli operators in S, which can be expressed as

$$|\phi_e^{\mathsf{S}}\rangle\langle\phi_e^{\mathsf{S}}| = \frac{1}{2^{n-k}} \sum_{s\in\mathsf{S}} (-1)^{\langle s,e\rangle} P_s, \tag{3.2}$$

for $e \in \mathsf{S}^{\perp} \coloneqq \mathbb{Z}_2^{2(n-k)}/\mathsf{S}$ known as the *error syndrome*. Note that $\{|\phi_e^{\mathsf{S}}\rangle\}_e$ forms an orthonormal basis for the $(n-k)$-qubit subsystem. The *stabilizer covering* O is a set of stabilizer groups $\{\mathsf{S}_i\}_i$ such that every Pauli operator belongs to at least one S_i [FW20].

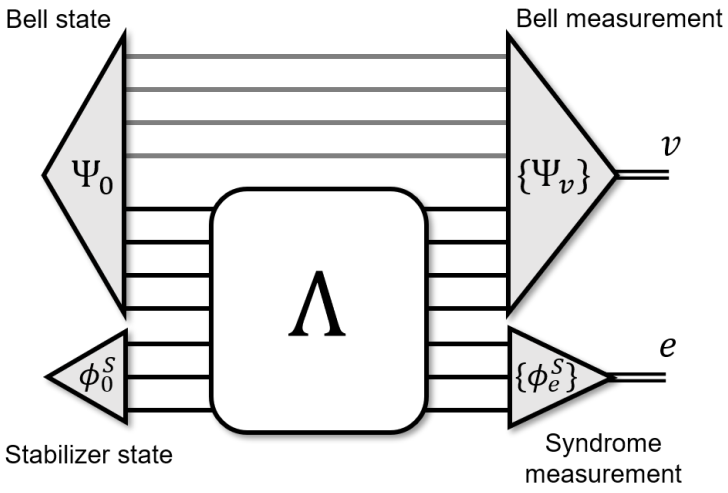


Figure 3.3: A single round of measurement for the k-qubit-assisted Pauli channel estimation protocol in Algorithm 1. Here, a 4-qubit ancilla is used to estimate a 7-qubit Pauli channel.

Theorem 3.1. *Algorithm 1 gives an estimate $\widehat{\boldsymbol{\lambda}}$ for the Pauli eigenvalues $\boldsymbol{\lambda}$ of any n-qubit Pauli channels that satisfies*

$$|\widehat{\lambda}_a - \lambda_a| \leq \varepsilon, \quad \forall a \in \mathbb{Z}_2^{2n} \tag{3.3}$$

with success probability at least $1-\delta$, given the following number of samples

$$N = \mathcal{O}(|\mathsf{O}| \times n\varepsilon^{-2} \log \delta^{-1}). \tag{3.4}$$

Algorithm 1 k-qubit-ancilla-assisted Pauli channel estimation

Input: (1) N copies of an n-qubit Pauli channel Λ. (2) A stabilizer covering of $\mathsf{P}^{\mathsf{n-k}}$ denoted as O.

Output: Estimates $\widehat{\boldsymbol{\lambda}}$ for the Pauli eigenvalues of Λ.

1: $\widehat{\lambda}_a \coloneqq 0$, $N_a \coloneqq 0$ **for all** $a \in \mathbb{Z}_2^{2n}$.
2: **for** $\mathsf{S} \in \mathsf{O}$ **do**
3: **for** $i = 1$ **to** $\lfloor N/|\mathsf{O}| \rfloor$ **do**
4: Input $|\Psi_0\rangle \otimes |\phi_0^{\mathsf{S}}\rangle$ to $\mathbb{1} \otimes \Lambda$.
5: Measure in basis $\{|\Psi_v\rangle \otimes |\phi_e^{\mathsf{S}}\rangle\}$ with outcomes v, e.
6: **for** $u \in \mathbb{Z}_2^{2k}, s \in \mathsf{S}$ **do**
7: $\widehat{\lambda}_{u \oplus s}$ += $(-1)^{\langle u,v\rangle + \langle s,e\rangle}$, $N_{u\oplus s}$ += 1.
8: $\widehat{\lambda}_a \coloneqq \widehat{\lambda}_a / N_a$ **for all** $a \in \mathbb{Z}_2^{2n}$.
9: **return** $\widehat{\boldsymbol{\lambda}} \coloneqq \{\widehat{\lambda}_a\}_a$.
(Note: $\oplus$ stands for string concatenation.)

Proof of Theorem 3.1. The probability distribution of measurement outcomes at Line 5 in Algorithm 1 can be calculated as

$$p(v,e) = \frac{1}{2^{n+k}} \sum_{u \in \mathbb{Z}_2^{2k}} \sum_{s \in \mathsf{S}} \lambda_{u\oplus s} (-1)^{\langle u,v\rangle} (-1)^{\langle s,e\rangle}. \tag{3.5}$$

See Appendix A.1 for more details. Therefore, $p(v,e)$ and $\lambda_{u\oplus s}$ are related by the Walsh-Hadamard transform. Taking the inverse transform, we get

$$\lambda_{u\oplus s} = \sum_{v \in \mathbb{Z}_2^{2k}} \sum_{e \in \mathsf{S}^\perp} p(v,e) (-1)^{\langle u,v\rangle + \langle s,e\rangle}. \tag{3.6}$$

Thus, $(-1)^{\langle u,v\rangle + \langle s,e\rangle}$ is an un-biased estimator of $\lambda_{u\oplus s}$. According to Hoeffding's bound, $N_0 = \mathcal{O}(\varepsilon^{-2} \log \delta_0^{-1})$ samples are enough to estimate a single $\lambda_{u\oplus s}$ to additive precision ε with success probability at least $1 - \delta_0$. Since every λ_a is covered by some stabilizer group $\mathsf{S} \in \mathsf{O}$, a total sample complexity of $N = \mathcal{O}(|\mathsf{O}| \times n\varepsilon^{-2} \log \delta^{-1})$ is enough to estimate all λ_a to additive error ε simultaneously with success probability at least $1 - \delta$, by setting $\delta_0 \coloneqq 4^{-n}\delta$ and applying the union bound. □

Corollary 3.2. *There exists a k-qubit-ancilla-assisted non-adaptive non-concatenating Pauli channel estimation protocol achieving* $|\widehat{\lambda}_a - \lambda_a| \leq \varepsilon$ *for all* $a \in \mathbb{Z}_2^{2n}$ *with probability* $1-\delta$ *using* $N = \mathcal{O}(n2^{n-k}\varepsilon^{-2}\log\delta^{-1})$ *samples.*

Proof. This follows from the existence of a stabilizer covering for P^{n-k} of size $2^{n-k}+1$ [LBZ02], which in turn follows from the existence of $2^{n-k}+1$ mutually-unbiased bases for $(n-k)$-qubit systems [WF89]. □

We remark that, the choice of stabilizer covering in Corollary 3.2 gives the optimal sample complexity for all *non-adaptive* and *non-concatenating* Pauli eigenvalues estimation protocols, as proved in the next section. From a practical point of view, it involves $(n-k)$-qubit stabilizer states that might be difficult to prepare. A more experimental friendly version is to choose the stabilizer covering generated by all 3^{n-k} possible Pauli measurements, in which case only Pauli eigenstates preparation and Pauli measurements are required (on the $(n-k)$-qubit subsystem), at the expense of a sub-optimal sample complexity $N = \mathcal{O}(n3^{n-k}\varepsilon^{-2}\log\delta^{-1})$.

3.4 Lower bounds

We have established a sample complexity upper bound of $\mathcal{O}(n2^{n-k})$ for non-adaptive non-concatenating k-qubit-ancilla-assisted Pauli eigenvalues estimation protocols, which implies a $\mathcal{O}(n)$ upper bound for the n-qubit ancilla case and a $\mathcal{O}(n2^n)$ upper bound for the ancilla-free case. In the following Theorem 3.2, we provide corresponding lower bounds to justify the exponential advantage provided by ancilla in this task.

Theorem 3.2. *For any estimation protocol that give an estimate* $\widehat{\boldsymbol{\lambda}}$ *for the Pauli eigenvalues* $\boldsymbol{\lambda}$ *of an arbitrary unknwon n-qubit Pauli channel* Λ *such that*

$$|\widehat{\lambda}_a - \lambda_a| \leq 1/2, \quad \forall a \in \mathbb{Z}_2^{2n} \tag{3.7}$$

holds with high probability, the number of samples of Λ *must satisfies (recall Fig. 3.1)*

(A) $N = \Omega(n2^{n-k})$*, for non-adaptive non-concatenating k-qubit-ancilla measurements.*

(B) $N = \Omega(2^{(n-k)/3})$*, for adaptive non-concatenating k-qubit-ancilla measurements.*

(C) $N \geq N' = \Omega(2^{n/3})$*, for adaptive concatenating ancilla-free measurements, where* N' *stands for the number of measurement rounds.*

(D) $N = \Omega(n)$*, for fully entangled measurements.*

Indeed, Theorem 3.2 and Corollary 3.2 establishes an exponential advantage of ancilla-assisted measurements over ancilla-free measurements even with channel concatenation (as in the RB-type Pauli channel estimation protocols in [FW20]). Furthermore, for the non-concatenating cases, we see a roughly matching bounds for all ancilla size $0 \leq k \leq n$, which can be interpreted as that a small number of ancilla ($k = o(n)$) would not help much in improving the sample efficiency. We also see from (D) that the sample complexity of Algorithm 1 with n ancilla qubits is optimal among all entangled strategies, thus we need not study protocols with more than n ancilla qubits.

Sketch of the proof. Our proof generalizes the techniques of Huang *et al.* [HKP21] in proving lower bounds for learning Pauli expectation values of quantum states. The key is to construct the following set of Pauli channels

$$\left\{\Lambda_{(a,s)}(\cdot) = \frac{1}{2^n}(I\,\mathrm{Tr}(\cdot) + sP_a\,\mathrm{Tr}(P_a(\cdot)))\right\}_{a,s}, \tag{3.8}$$

for $a \in \{1, \cdots, 4^n - 1\}$ and $s = \pm 1$. An estimation protocol satisfying the assumption of Theorem 3.2 is able to identify an arbitrary element of this set using N copies of the channel. We can then use information-theoretical arguments to lower bound N for (A); The bounds in (B) and (C) are proved by reducing the learning problem to a channel discrimination problem between the completely-deplorizing channel and the channels in Eq. (3.8); To prove

the bound in (D), we first use *teleportation stretching* [PLOB17, PLLP19] to reduce any estimation protocols on N copies of the Pauli channel into a POVM measurement on N copies of their Choi states [Cho75, Jam72], and then apply the Holevo's theorem [Hol73]. See Appendix A.2 for a full proof of Theorem 3.2. □

A tight lower bound for entanglement-free learning schemes. After Theorem 3.2 is obtained in [CZSJ22], an improved lower bound of $\Omega(2^n/\varepsilon^2)$ is obtained by me and my collaborators for entanglement-free learning schemes (as in Fig. 3.2). This bound closes a gap between the lower bound from Theorem 3.2(C) and the upper bound from Theorem 3.1. It also holds against a more general class of schemes and achieves am optimal scaling for ε.

Theorem 3.3. *If there exists an entanglement-free scheme that, for any n-qubit Pauli channel Λ, outputs an estimator $\widehat{\lambda}_a$ such that $|\widehat{\lambda}_a - \lambda_a| \leq \varepsilon \leq 1/6$ with probability at least $2/3$ for any $a \in \mathsf{P}^n$, after making N rounds of measurement, then $N = \Omega(2^n/\varepsilon^2)$.*

This matches the upper bound of $O(n2^n/\varepsilon^2)$ up to a logarithmic factor. In fact, if we consider a modified task of estimating each λ_a with $2/3$ success probability *individually* rather than *simultaneously*, the upper bound is easily seen to be $O(2^n/\varepsilon^2)$, in which case our bound is tight without the logarithmic factor. Combined with the bound of $\Theta(1/\varepsilon^2)$ using n-qubit entanglement-assisted scheme [CZSJ22], this gives a tight exponential separation for learning with and without entanglement in the task of Pauli channel learning. Another noteworthy feature is that our lower bound has a moderate constant factor. For example, with $\varepsilon \leq 0.1$ and $n \geq 5$ we have $N \geq 0.01 \times 2^n/\varepsilon^2$.

3.5 Discussion

In this chapter, we show a provable quantum advantage provided by entangled measurements for a learning task [ACQ22, HKP21] of Pauli channel estimation, which is a practically useful tool urgently needed to characterize large quantum systems. For quantum benchmarking, our

results provide fundamental efficiency limits for Pauli noise estimation, which solves an open problem raised in [FW20]. In the following chapter, we also describe how the ancilla-assisted Pauli channel estimation protocol can be applied to a practical quantum benchmarking tasks in a noise-resilient and sample-efficient manner. Our results provide a promising tool for both characterizing near-term quantum devices and demonstrating quantum advantages in those systems

One interesting open problem is to explore a deeper connection between the resource theory of entanglement [BBPS96, BDSW96, VPRK97] and quantum learning theory. Specifically, since the tight bounds for Pauli channel learning with no entanglement and arbitrary entanglement have been settled, it is natural to ask what about learning with a certain amount of entanglement. There could be different ways of defining the entanglement cost of a learning scheme, one of which is to allow a bounded amount of ancillary qubits [CZSJ22, CCHL22]. As discussed in this chapter, an upper bound of $\widetilde{O}(2^{n-k}/\varepsilon^2)$ for k-ancillary-qubit-assisted scheme is known and proved optimal for some restricted class of schemes [CZSJ22], but a general answer to this question is yet to be found.

3.5.1 Related Works

The problem of Pauli channel learning has been studied with different figures of merit. For Pauli error rates, [FW20] provides an ancilla-free protocol that learns the Pauli error rates to precision ε in l_2-distance with $\widetilde{O}(2^n/\varepsilon^2)$ samples, which implies an upper bound of $\widetilde{O}(2^{3n}/\varepsilon^2)$ for learning in l_1-distance. Ref. [FO21] shows that $\widetilde{O}(\log n/\varepsilon^2)$ samples is sufficient to learn the Pauli error rates to precision ε in l_∞-distance without using ancilla; In the case that ancilla is allowed, one can use Bell states and Bell measurements to directly sample from the Pauli error rates, which implies an $O(2^{2n}/\varepsilon^2)$ upper bound for learning in l_1-distance [FO21, CZSJ22]; For Pauli eigenvalues, [CZSJ22] gives a family of k-qubit ancilla-assisted protocols using $O(n2^{n-k}/\varepsilon^2)$ samples for learning in l_∞-distance, for $0 \leq k \leq n$.

In terms of the lower bounds, [CZSJ22] focuses on learning Pauli eigenvalues to constant error in l_∞-distance, and in particular obtains a lower bound $\Omega(2^{n/3})$ for the number of measurements for any ancilla-free schemes with concatenation. The current work improves this lower bound to be tight (for more general schemes). Another recent work studies learning Pauli error rate to error ε in l_1-distance [FOF23]. For ancilla-free schemes with adaptivity, they obtains $\Omega(2^{2n}/\varepsilon^2)$ for general case and $\Omega(2^{2.5n}/\varepsilon^2)$ when ε is exponentially small in n. They allows concatenation with *unital* processing channels and the lower bound holds for the number of measurements. The results of the current work and [FOF23] do not imply each other[2]. Whether a tighter lower bound can be established with other figures of merit remains an open problem.

An independent and concurrent work [CG23b] obtains a tight lower bound of $\Omega(2^n/\varepsilon^2)$ on the number of measurements for learning every Pauli eigenvalues to ε precision with ancilla-free concatenating schemes. Their proof is based on a different technique, which leads to different features in their results compared to ours: On the one hand, the ε in their bound can be any value within $(0, 1]$, while ours is restricted to $\varepsilon \in (0, 1/6]$; On the other hand, our bound has a better constant factor and applies for the more general setting of classical-memory-assisted schemes.

2. We remark that, a lower bound of $\Omega(2^{3n}/\varepsilon^2)$ for learning Pauli error rates in l_1-distance would imply a bound of $\Omega(2^n/\varepsilon^2)$ for learning Pauli eigenvalues in l_∞-distance, via the Parseval–Plancherel identity and relations of l_p-norm.

CHAPTER 4

SELF-CONSISTENT CHARACTERIZATION OF PAULI NOISE

Chapter Notes. This chapter contains a series of my works on quantum noise characterization based on learning Pauli noise models, mainly based on [CLO+23, CZJF24].

4.1 Introduction

Despite recent experimental progresses, noise remains as a major challenge in building a practical quantum information processing platform. In order to suppress, mitigate, and correct noise affecting a quantum device, it is essential to first gain knowledge about the noise. However, learning quantum noise is not easy. Noise is complicated, containing a large number of parameters that grows exponentially with the system size. Meanwhile, the available quantum control for noise learning is limited and can itself suffer from imperfection. Developing efficient, reliable, and informative quantum noise characterization protocols is thus an important problem and an active field of research [EHW+20, PYBBK24, HNG+24].

Among different noise characterization approaches, Pauli noise learning is a simple but powerful formalism that has gained increasing attention recently [FW20, EWP+19, FO21, HFW20, Fla22, WKBK23, VDBMKT23, CLO+23, CZSJ22, CDDH+23, vdBW24, HDH25]. Here, the noise processes affecting different components (i.e., state preparation, measurements, gates) of a quantum device are modeled by Pauli channels. Pauli channels can naturally describe a wide spectrum of incoherent noises, including depolarizing and dephasing noises. It is also a widely used model in the study of quantum error correction and fault-tolerant quantum computing [CYK+22]. Besides, there exist techniques known as randomized compiling [Kni05, WE16, HNM+21] that, given mild assumptions about the noise and gate set, can tailor an arbitrary noise channel into a Pauli channel, justifying the validity of this noise model.

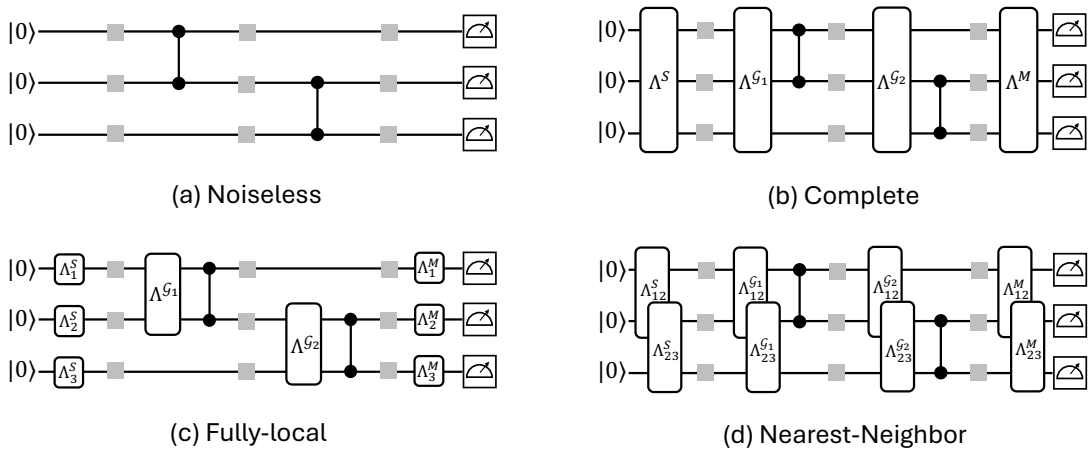


Figure 4.1: Examples of different noise ansatz on a 3-qubit system. The available set of multi-qubit gates are two CZ gates. The small square box represents single-qubit gates whose noise is ignored (or absorbed to multi-qubit gates and SPAM). (a) A noiseless circuit. (b) Circuit with complete Pauli noise, where the noise for SPAM and both gates are all generic 3-qubit Pauli channels. (c) Circuit with fully local Pauli noise, where the SPAM noise channels are qubit-wise independent, while the gate noise channels has the same support as the gate. (d) Circuit with nearest-neighbor Pauli noise (a special case of quasi-local model), where each Pauli noise channel is nearest-neighbor 2-local. See Sec. B.5 for a rigorous definition.

Despite intensive recent study, some core aspects of Pauli noise learning have not been well-understood. In particular, a learning algorithm that is simultaneously *self-consistent*, *comprehensive*, and *efficient* is yet to be established. Here, self-consistency means using only operations that are noisy to learn about themselves, comprehensiveness means to learn all the information that is self-consistently learnable, and efficiency means the number of measurements and classical processing time scales favorable with the number of qubits. In this chapter, we build up a framework called *gate set Pauli noise learning*, give analytic characterization of the self-consistently learnable degrees of freedom, and develop an algorithm with the above three desirable features.

Problem Setup. Consider an n-qubit quantum system. Suppose, by design, the following operations can be implemented on the platform: (1) Initialize the system to the ground state; (2) Apply a layer of n single-qubit unitary gates; (3) Apply some multi-qubit Clifford

gates gate from a finite set denoted by $\mathfrak{G}$; (4) Measure all qubits on the computational basis. We refer to the collection of all the above components as a *gate set*, following the language of gate set tomography (GST) [NGR+21], although we are also including noisy SPAM (state preparation and measurement). In practice, all the operations of a gate set can be noisy. Our goal is to characterize the noisy gate set *self-consistently*. That is, using only operations from the gate set to learn about themselves.

Instead of dealing with the most general form of noise, we focus on what is known as the *Pauli noise model*: We ignore the noise of any single-qubit gates, and assume the noise channels acting on the initial state, measurements, and any multi-qubit Clifford gate to be a (component-dependent) Pauli channel, which takes the following form,

$$\Lambda(\rho) = \sum_{a \in \mathsf{P}^n} p_a P_a \rho P_a = \sum_{b \in \mathsf{P}^n} \lambda_b P_b \operatorname{Tr}(P_b \rho)/2^n. \tag{4.1}$$

Here, $\mathsf{P}^n \coloneqq \{I, X, Y, Z\}^{\otimes n}$ is the n-qubit Pauli group (modulo phase). $\{p_a\}$ are the *Pauli error rates* while $\{\lambda_b\}$ are the *Pauli fidelities* or *Pauli eigenvalues*. We assume $\lambda_b > 0$ throughout this chapter. Such a model can be ensured via randomized compiling under mild assumptions [WE16, HNM+21]. The action of noise channel on the initial state ρ_0, POVM element E_j, and gate $\mathcal{G} \in \mathfrak{G}$ is given by

$$\tilde{\rho}_0 = \Lambda^S(\rho_0), \quad \tilde{E}_j = \Lambda^M(E_j), \quad \tilde{\mathcal{G}} = \mathcal{G} \circ \Lambda^{\mathcal{G}}, \tag{4.2}$$

respectively. A realization of the gate set Pauli noise model is specified by the involved Pauli channels $\{\Lambda^S, \Lambda^M, \Lambda^{\mathcal{G}} : \mathcal{G} \in \mathfrak{G}\}$. Since a generic Pauli channel contains exponentially many parameters, we often need to assume an efficient parametrization of the Pauli channels, which we call a *noise ansatz*. Examples of noise ansatzes include spatially local or quasi-local Pauli channels, as shown in Fig. 4.1(c) and (d), respectively. We refer to a Pauli noise model with a noise ansatz as a *reduced model*, and the one without a noise ansatz as a *complete model*.

Throughout this chapter, we assume the true noise model is *realizable*. That is, the true noise model can be represented exactly as a point in the model space.

To characterize the noisy gate set means to learn all the Pauli noise channels with ansatz. However, it is generally impossible to fully determine those Pauli noise channels self-consistently [HFP22, CLO$^+$23]. Indeed, consider the following *gauge transformation* [NGR$^+$21] on the noisy operations,

$$\tilde{\rho}_0 \mapsto \mathcal{D}(\tilde{\rho}_0), \quad \tilde{E}_j \mapsto \mathcal{D}^{-1}(\tilde{E}_j), \quad \tilde{\mathcal{G}} \mapsto \mathcal{D}\tilde{\mathcal{G}}\mathcal{D}^{-1}, \quad \bigotimes_{i=1}^{n} \mathcal{U}_i \mapsto \mathcal{D}\left(\bigotimes_{i=1}^{n} \mathcal{U}_i\right)\mathcal{D}^{-1}, \tag{4.3}$$

where $\mathcal{D}$ is some invertible linear map. It is not hard to see that any experimental outcomes statistics remain unchanged under such transformation. Thus, two realizations of a noisy gate set related by a gauge transformation are *indistinguishable*. Meanwhile, one can carefully choose $\mathcal{D}$ (e.g., as a depolarizing channel on a subset of qubits) such that the transformed noisy gate set will remain a valid Pauli noise model within ansatz. That is, the single-qubit gates $\bigotimes_i \mathcal{U}_i$ remain noiseless, while $\Lambda^S, \Lambda^M, \Lambda^{\mathcal{G}}$ remain Pauli channels within the ansatz (possibly with different parameters).

Now, consider a function of the gate set noise parameters. If the function value changes under certain valid gauge transformations, then the function is not self-consistently learnable. This raises the first question of gate set Pauli noise learning as follows,

Problem 4.1. *How to classify the learnable and unlearnable functions of a gate set Pauli noise model?*

After understanding what are the learnable degrees of freedom, the next natural problem is to find an efficient learning protocol, which can be summarized as,

Problem 4.2. *How to efficiently learn all the learnable information of a get set Pauli noise model?*

In this chapter, we will establish a generic theoretic framework to address the above two problems. To demonstrate the power and flexibility of our framework, we will apply it to Pauli noise models with concrete gate set and noise ansatzes that are considered in the literature (e.g. [Fla22, VDBMKT23]) and resolve open problems therein. I will then report an experimental demonstration of our framework in a 2-qubit example.

In Sec. 4.2, we summarize our theory contribution. In Sec. 4.3 we provides a minimal example on learning noises of a single CNOT gate. More details about our framework are put in Appendix B.

4.2 Main Results

Learnability of gate set Pauli noise. Our approach to characterize the learnability (i.e. classification of learnable and unlearnable information) is to first encode all the noise parameters into a linear space. Let us start with the complete Pauli noise model (i.e. the one without a noise ansatz). Similar results have been obtained in [CLO$^+$23] for this case, but here we will use a more formal way that can be extended to the reduced model later.

Define the logarithmic Pauli eigenvalues as $x_b \coloneqq -\log \lambda_b$. We collect all logarithmic Pauli eigenvalues (excluding the trivial ones corresponding to identity operator) from all Pauli noise channels into a ordered list denoted by $\boldsymbol{x}$, which can be viewed as a vector of a real linear space according to a standard basis. We denote the space $\boldsymbol{x}$ lives on as X, called the *parameter space*. An *experiment* is defined as a mapping $X \mapsto \mathbb{R}^{2^n}$ specified by a sequence of gates $\mathcal{C}$. It maps $\boldsymbol{x}$ to a measurement outcome probability distribution given by the Born rule, $\Pr(j) = \mathrm{Tr}[\tilde{E}_j \tilde{\mathcal{C}}(\tilde{\rho}_0)]$, where the noisy realization of state, gates, and measurement is determined by $\boldsymbol{x}$. Now, we define a linear function $\boldsymbol{f} \in \mathcal{L}(X \mapsto \mathbb{R}) =: X'$ to be *learnable* if its value can be completely specified by a set of experiments[1]; We also define

1. One may wonder why we only consider *linear* functions of the logarithmic Pauli eigenvalues. This is because any experiments can be determined by some learnable linear functions, so they suffices to characterize the learnability. See Theorem 2.1 for more details.

a vector $\mathfrak{d} \in X$ to be a *gauge* vector if the two lists of noise parameters $\boldsymbol{x}$ and $\boldsymbol{x} + \mathfrak{d}$ cannot be distinguished by any experiment, for all $\boldsymbol{x} \in X$. The linear subspace of all learnable linear function and all gauge vectors are denoted by L and T, respectively. The problem then becomes characterizing X, L, and T.

For this purpose, we introduce a directed graph called the Pauli pattern transfer graph (PTG). This graph is first defined in [CLO+23] with some twist needed to fit into the current framework [CZJF24]. The graph consists of 2^n nodes. One of which is a root node (also called a SPAM node), the others correspond to the $2^n - 1$ non-zero n-bit strings, each denotes a Pauli pattern (the pattern of a n-qubit Pauli operator can be understood as its support); For each logarithmic Pauli eigenvalues, we assign a unique edge to the PTG: For gate noise parameters, each edge describe how a Clifford gate transform certain Pauli operators (where we only track the pattern); For each state preparation (or, measurement) noise parameter, we assign an edge from the root node to certain pattern node (or, from a pattern node to root). An example of PTG is given in Fig. 4.2.

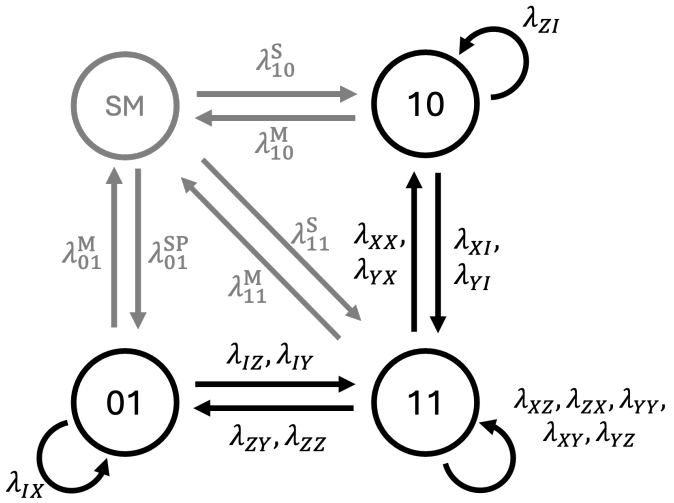


Figure 4.2: The PTG for $n = 2$, $\mathfrak{G} = \{\text{CNOT}\}$. The black edges corresponds to gate noise parameters, while the gray edges corresponds to SPAM noise parameters. Multiple edges are shown as one edge with multiple labels.

PTG contains all information about a complete Pauli noise model. Intuitively, a path

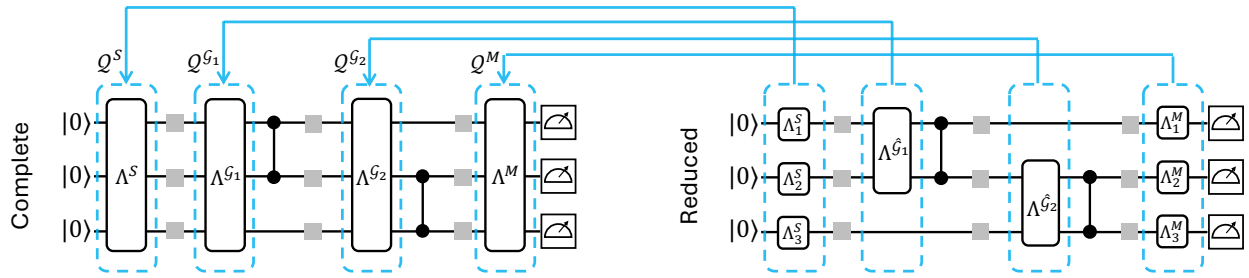


Figure 4.3: A 3-qubit example of complete Pauli noise model (Left) and reduced Pauli noise model (Right). The gate set contains $\{\mathrm{CZ}_{12}, \mathrm{CZ}_{23}\}$. The ansatz for the reduced model assumes the state-preparation-and-measurement noise channels to be a tensor product of single-qubit Pauli channel, and the gate noise channel to only act on the gate support. The blue arrow represents the embedding map $\mathcal{Q}$ from the reduced to the complete model.

starting from and ending at the root node gives a combination of logarithmic Pauli eigenvalues that can be learned from an experiment. This intuition can be made rigorous using tools from algebraic graph theory [Bol98, GLS03]. Informally, let the edge space E to be the real linear space spanned by all edges of PTG. For each cycle (or, cut) on PTG, we can assign it with a vector of E in a proper way (see Sec. B.2.2). The subspace spanned by all cycle (or, cut) vectors is defined as the *cycle space* Z (or, the *cut space* U). It is known that Z and U are orthogonal complements to each other within E. Our main theorem is that, up to the natural identification between the parameter space X and the edge space E, the learnable space L is identical to the cycle space Z, while the gauge space T is identical to the cut space U. This implies that L and T are also orthogonal complements to each other within X.[2]

Next, we turn to study the learnability of a reduced model. A reduced model is specified by a tuple $(X_R, \mathcal{Q})$. Here X_R is the *reduced parameter space*, a linear space whose components are viewed as lists of the reduced noise parameters (denoted by $\boldsymbol{r}$). $\mathcal{Q} : X_R \mapsto X$ is the *embedding map* describing how the reduced parameters specifies the Pauli channels. Specifically, the logarithmic Pauli eigenvalues are given by $\boldsymbol{x} = \mathcal{Q}(\boldsymbol{r})$. We will assume $\mathcal{Q}$

2. Rigorously speaking, L is a subspace of the dual space X'. Here we identify X with X' using the standard basis for defining X.

is linear and injective. See Fig. 4.3 as an illustrative example. The notion of experiments on reduced model can be naturally induced from the definition for complete model. We can also similarly define the reduced learnable space L_R (learnable functions of X_R) and the reduced gauge space T_R (indistinguishable transformations within X_R). The problem is then to characterize X_R, L_R, T_R. The challenge here is that, it is not easy to construct a graph describing the reduced noise parameters as in the complete model. Our approach is to connect the reduced spaces back to the complete spaces, the use the characterization that we have already obtained. We proved that $L_R = \mathcal{Q}^{\mathsf{T}}(L)$, $T_R = \mathcal{Q}^{-1}(T)$. In words, the reduced learnable space is the complete learnable space projected by $\mathcal{Q}^{\mathsf{T}}$, while the reduced gauge space is the preimage of the complete gauge space via $\mathcal{Q}$.

The relation between all the linear spaces introduced above can be summarized as follows.

Theorem 4.1 (Learnability, informal)**.** *For any reduced model $(X_R, \mathcal{Q})$ such that $\mathcal{Q}$ is linear and injective,*

Reduced model:	X_R	$=$	L_R	$\oplus^{\perp}$	T_R
	(reduced parameter)		*(reduced learnable)*		*(reduced gauge)*
	$\downarrow_{\mathcal{Q}}$		$\uparrow_{\mathcal{Q}^{\mathsf{T}}}$		$\uparrow_{\mathcal{Q}^{-1}}$
Complete model:	X	$=$	L	$\oplus^{\perp}$	T
	(parameter)		*(learnable)*		*(gauge)*
	$\|\|$		$\|\|$		$\|\|$
Pattern transfer graph:	E	$=$	Z	$\oplus^{\perp}$	U
	(edge)		*(cycle)*		*(cut)*

The equality is up to the natural identification between X, X', and E. $\bigoplus^{\perp}$ means orthogonal complements.

Theorem 4.1 resolved Problem 4.1, giving a complete characterization about the learnable degrees of freedom for a Pauli noise model with general gate sets and noise ansatzes. They also serve as the basis for developing efficient self-consistent learning algorithms. Rigorous

statements and detailed proofs for Theorem 4.1 are presented in Sec. B.2.

Efficient learning algorithms for gate set Pauli noise. We now consider Problem 4.2, how to efficiently and self-consistently learn a Pauli noise model. Thanks to Theorem 4.1, to learn a reduced Pauli noise model means to learn every functions of the reduced learnable space L_R. It suffices to find a basis for L_R and learn the value of each basis function, as all the other functions can be decided by linearity. This can be done by first finding certain cycle basis for $E = L$ with each basis function learnable from some experiment, and then use the fact $L_R = \mathcal{Q}^{\mathsf{T}}(L)$ to pick up a subset of functions to form a basis for L_R. We give such a construction with the additional feature that each experiments uses at most one noisy gate, summarized as follows.

Theorem 4.2 (Simple experiment construction, informal)**.** *There exists a set of* $M = \dim(L_R)$ *experiments that can determine the value of any functions of* L_R*. Moreover, any experiment from this set either apply one gate from* $\mathfrak{G}$ *exactly once, or apply no gates from* $\mathfrak{G}$ *at all.*

To show this, we first show that any cycle containing the root node exactly once (called a *rooted cycle*) can by learned by one experiment. Then notice that, thanks to the connectivity of PTG, there exists a rooted cycle basis where every cycle is of length 2 or 3, corresponding to experiment with no gate or exactly one gate from $\mathfrak{G}$. For an efficient noise ansatz, the number of reduced parameters should be at most polynomial in n, which means the number of experiments M needed in Theorem 4.2 is also polynomial. The protocol can thus be efficiently executed. Details about this protocol are presented in Sec. B.3.1.

In the literature of quantum noise learning, an important consideration is whether the noise parameter can be learned to relative precision. That is, given a sufficiently small infidelity parameter r of a noisy gate $\tilde{\mathcal{G}}$, can we obtain an estimator $\hat{r}$ such that $|\hat{r}-r| \leq O(\varepsilon r)$ with high probability, by making a number of measurement that scales as $\varepsilon^{-2}\mathrm{polylog}(r^{-1})$?

It is known that, if one can concatenate m copies of $\tilde{\mathcal{G}}$ to amplify the infidelity to order r^m for any m, this scaling can be achieved [HHF$^+$19, Theorem 1]. We should that such noise amplification is indeed possible for gate set Pauli noise model. We have that,

Theorem 4.3 (Amplifying gate noise parameters, informal)**.** *For any function that is a cycle vector and only supported on gate noise parameters, let* $\lambda \coloneqq \exp(-\boldsymbol{f}(\boldsymbol{x}))$*. Then, there exists a family of experiments* $\{F_m\}_{m\in\mathbb{N}}$ *and a Pauli observable, such that the expectation value of the* m*th experiment is given by* $A\lambda^m$*, where* A *is an* m*-independent parameter.*

Theorem 4.3 ensures one can achieve relative precision for a basis of learnable functions supported on the gate noise parameters, confirming that gate set Pauli noise learning has the same desirable feature as other quantum noise learning schemes such as randomized benchmarking [GCM$^+$12a]. Details about Theorem 4.3 are presented in Sec. B.3.2.

Theorem 4.2 and 4.3 address Problem 4.2 from two different aspects, providing efficient experiment construction for gate set Pauli noise learning. The key is to find an appropriate basis for the reduced learnable space. One caveat is that, since the numbers of edges and nodes in PTG grow exponentially with the number of qubits n, an algorithm that first finds a cycle basis in PTG and then select a subset to construct a basis for L_R will inevitably suffer from exponential run time. However, when the noise ansatzes have some structure, we can find such a basis analytically, thus circumventing the run time issue. In what follows, we demonstrate this point by applying our theory to concrete physically relevant noise ansatzes.

Applications to spatially-local noise ansatzes We now apply the general theory of learnability and learning algorithms to concrete noise ansatzes. We will focus on spatially local or quasi-local noise model as an example. Such noise models have been widely studied in the literature of Pauli noise learning [FW20, Fla22, HDH25, WKBK23, VDBMKT23, CDDH$^+$23, PCOG$^+$24], but our results for the first time give a clear classification of the learnable and gauge degrees of freedom within such models.

We first consider the fully-local model, where the SPAM noise channels are assumed to be qubit-wise product channels, while the gate noise channels act non-trivially only within the gate's support. This model is also sometimes known a crosstalk-free model and have been considered in, e.g., [Fla22, HDH25, PCOG$^+$24]. To understand the learnable and gauge spaces for this model, we introduce a family of gauge vectors called the *subsystem depolarizing gauges* (SDG), denoted by $\{\mathfrak{d}_\nu\}_{\nu\in 2^{[n]}}$, where $\mathfrak{d}_\nu$ represents a gauge transformation as in Eq. (4.3) with $\mathcal{D}$ chosen as a partially depolarizing channel acting on a subsystem of qubits ν. One can show the SDGs forms a basis for the complete cut space T. Interestingly, we show that the subset of SDG's acting on one qubits fully characterize the reduced gauge space T_R for a fully-local model.

Theorem 4.4. *For any fully-local noise model $(X_R, \mathcal{Q})$, its reduced gauge space satisfies,*

$$\mathcal{Q}(T_R) = \mathrm{span}\{\mathfrak{d}_s : |s| = 1\}. \tag{4.4}$$

Since $\mathcal{Q}$ is injective, this allows one to uniquely determine T_R. Intuitively, any multi-qubit SDG transformation will violate the locality assumption of qubit-wise independent SPAM noise (and sometimes also local gate noise), and is thus an invalid gauge transformation. We also studied the more involved cases where either SPAM or gates has a fully-local noise assumption while the other does not. Even in that case, we can show that certain subset of SDG's still spanned $\mathcal{Q}(T_R)$. Detailed are presented in Sec. B.4.2.

Regarding learning protocols, the reduced learnable space L_R and a set of $\dim(L_R)$ experiments can be analytically written down. see Sec. B.4.3. We further study the task of relative precision learning, which boils down to finding a cycle basis for the reduced gate noise parameters. Though we believe this is a hard problem in general, we give solutions to concrete examples in Sec. B.6. Discussion about general procedure is presented in the paper [CZJF24].

Going beyond the fully local model, we apply our theory to a quasi-local noise model that describes certain degrees of spatial correlation while remains parameter-efficient. The quasi-local Pauli noise channels we consider are those whose Pauli eigenvalues factorize according to a factor graph [MK05]. We first introduce some notations that are needed for defining the quasi-local noise model, defined as follows

Definition 4.3. *An n-qubit Pauli channel Λ is Ω-local if its Pauli eigenvalues satisfies*

$$\lambda_a = \prod_{b \triangleleft a, b \sim \Omega} \exp(-r_b), \quad \forall a \neq I_n, \tag{4.5}$$

for some $\{r_b \in \mathbb{R} : b \sim \Omega, b \neq I_n\}$ and a factor set $\Omega \subseteq 2^{[n]}$.

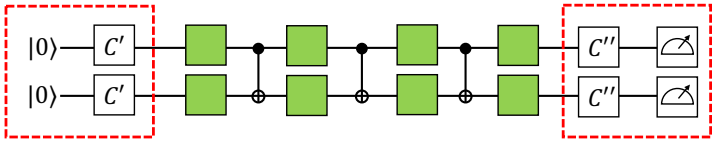
Here, $\Omega \subseteq 2^{[n]}$ is a *factor set* if it satisfies $\forall \nu \in \Omega$, $\forall \varnothing \subsetneq \mu \subseteq \nu$, $\mu \in \Omega$. We write $b \triangleleft a$ if $b_i \neq I \Rightarrow a_i = b_i$ for all $i \in [n]$. We write $b \sim \Omega$ if the support of b belongs to Ω. We show that, under certain "covariance" assumptions of the quasi-local model, the reduced gauge space is again characterized by a subset of SDGs that are consistent with the channel factorization,

Theorem 4.5. *(Informal) For any quasi-local noise model $(X_R, \mathcal{Q})$, given that all Pauli channels are Ω-local and is covariant for all gate in $\mathcal{G}$, then its reduced gauge space satisfies,*

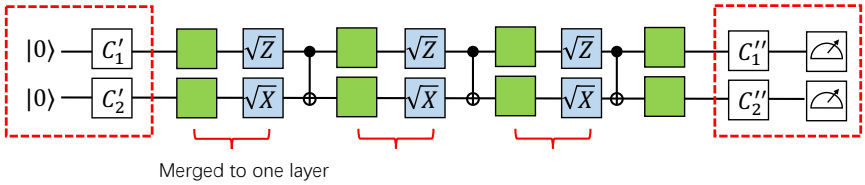
$$\mathcal{Q}(T_R) = \mathrm{span}\{\mathfrak{d}_\nu : \nu \in \Omega\}. \tag{4.6}$$

Rigorous statements and detailed proofs are presented in Sec. B.5, where we also review useful properties of such noise model and discuss its connection and distinction with existing models studied in the literature [FW20, VDBMKT23, WKBK23]. A cycle basis for the reduced learnable space L_R is also analytically constructed for this case.

Finally, in Sec. B.6, we analyze a concrete gate set where the Control-Z (CZ) gates are the only multi-qubit entangling gates. We characterize the learnability of this noise model



(a) standard CB



(b) CB with interleaved gates

Figure 4.4: Cycle benchmarking for learning the Pauli noise channel of a CNOT gate. (a) Standard CB circuits, where CNOT gates are interleaved by random Pauli gates (green boxes), with initial stabilizer states and Pauli basis measurements (red boxes). (b) CB circuits with additional interleaved single qubit Clifford gates (blue boxes).

with more than one noise ansatzes (which are directly related to those that have been studied in the literature [VDBMKT23]) and give efficient learning algorithms in each case. These examples serve as an explicit recipe of gate set Pauli noise learning that is ready to be applied in current quantum devices.

4.3 Example: learning noise of a CNOT

In this section, we present a minimal example of our theory in learning noises of a single CNOT gate, along with experimental implementation on IBM's cloud quantum computers [ibm22]. These data are reported in [CLO+23].

Consider the CNOT gate which maps the Pauli operator IX to itself. Fig. 4.4 (a) shows the standard *cycle benchmarking* (CB) circuit. Imagine that we put the Pauli operator IX after the left red box and evolve it with the circuit, then the evolved operator (before the right red box) equals $\lambda_{IX}^3 \cdot IX$, up to a $\pm$ sign (which comes from the random Pauli gates and can always be accounted for during post-processing). Recall that we use the convention that the noise channel happens before each CNOT gate, $\widetilde{\mathcal{G}} = \mathcal{G} \circ \Lambda^{\mathcal{G}}$ (we omit the superscript

of $\mathcal{G}$ from here on). In experiments, we prepare a $+1$ eigenstate of IX (such as $|+\rangle|+\rangle$), measure the expectation value of IX at the end, and average over random Pauli twirling sequences. These SPAM operations are noisy and are represented as the red boxes. It is shown [EWP+19, Theorem 1 in Supplementary Information] that the measured expectation value equals

$$\mathbb{E}\langle IX\rangle = A_{IX} \cdot \lambda_{IX}^{d} \tag{4.7}$$

where the expectation is over random Pauli twirling gates and randomness of quantum measurement, and $A_{IX} = \lambda_{IX}^{S}\lambda_{IX}^{M}$ depends on SPAM noise but is independent of circuit depth d. From this λ_{IX} can be learned by estimating the observable IX at several different depths and perform a curve fitting.

The Pauli operator IX is special as it is invariant under CNOT. Consider another example: CNOT maps XZ to YY and vice versa. To learn $\lambda_{XZ}, \lambda_{YY}$ we propose an *interleaved cycle benchmarking* protocol as shown in Fig. 4.4 (b), where we insert additional layers of single-qubit Clifford gates $\sqrt{Z} \otimes \sqrt{X}$ that also maps XZ to YY and vice versa (up to a minus sign that can always be accounted for during post-processing). After XZ picks up a coefficient λ_{XZ} in front of the CNOT gate, it gets mapped to $\lambda_{XZ} \cdot YY$ by CNOT but then rotated back to $\lambda_{XZ} \cdot XZ$ by $\sqrt{Z} \otimes \sqrt{X}$. Following the same argument we conclude that both λ_{XZ} and λ_{YY} are learnable. Recall that we have assume single-qubit gates to be noiseless, as they can be merged into the single-qubit twirling layers, and our noise characterization result can be interpreted as the noise channel induced by a dressed cycle which consists of a CNOT gate and a layer of single-qubit gates [WE16].

The main challenge comes with the next example: CNOT maps IZ to ZZ and vice versa. By directly applying cycle benchmarking as in Fig. 4.4 (a) (with even depth d) we obtain

$$\mathbb{E}\langle IZ\rangle = A_{IZ} \cdot \lambda_{IZ}\lambda_{ZZ}\lambda_{IZ}\lambda_{ZZ}\cdots = A_{IZ}\left(\lambda_{IZ}\lambda_{ZZ}\right)^{d/2}, \tag{4.8}$$

and curve fitting gives $\sqrt{\lambda_{IZ}\lambda_{ZZ}}$ (similar results have been obtained in [EWP+19, HNM+21, VDBMKT23, FHV+24]). To learn λ_{IZ}, we may consider applying the same technique in Fig. 4.4 (b). However, the problem is that once IZ gets mapped to ZZ, it cannot be rotated back to IZ because I is invariant under single qubit unitary gates. The main difference between this example and previous examples is that here the *Pauli weight pattern* (an n-bit binary string with 0 indicating identity and 1 indicating non-identity) changes from 01 to 11, thus making the single qubit interleaving tricks inapplicable.

In fact, our Theorem 4.1 can be applied to show which eigenvalues of Λ is learnable. Recall the Pattern transfer graph (PTG) of a single CNOT as given in Fig. 4.2. Theorem 4.1 says all cycles on the graph are learnable parameters (e.g., $\lambda_{IX}, \lambda_{YY}, \lambda_{IZ} \cdot \lambda_{ZZ}$), while all parameters not in the cycle space are unlearnable (e.g., λ_{IZ}). For the purpose of this example, we only aims at learning gate-noise parameter, so we focus on the subgraph of PTG excluding the root node (In fact, this is our original definition of PTG in [CLO+23]). In Table 4.1, we lists a cycle basis, a basis of learnable Pauli fidelities and Pauli error rates (see the next paragraph), and a choices of unlearnable parameters. Note that there are 13 learnable degrees of freedom and 2 gauge degrees of freedom in this example.

Gate	CNOT
(a) Cycle basis	$e_{ZI}, e_{IX}, e_{ZX}, e_{XZ}, e_{YY}, e_{XY}, e_{YZ}$, $e_{IZ}+e_{ZZ}, e_{IY}+e_{ZY}, e_{IZ}+e_{ZY}$, $e_{XI}+e_{XX}, e_{YI}+e_{YX}, e_{XI}+e_{YX}$
(b) Learnable Pauli fidelities	$\lambda_{ZI}, \lambda_{IX}, \lambda_{ZX}, \lambda_{XZ}, \lambda_{YY}, \lambda_{XY}, \lambda_{YZ}$, $\lambda_{IZ}\cdot\lambda_{ZZ}, \lambda_{IY}\cdot\lambda_{ZY}, \lambda_{IZ}\cdot\lambda_{ZY}$, $\lambda_{XI}\cdot\lambda_{XX}, \lambda_{YI}\cdot\lambda_{YX}, \lambda_{XI}\cdot\lambda_{YX}$
(c) Learnable Pauli errors (Approx.)	$p_{ZI}, p_{IX}, p_{ZX}, p_{XZ}, p_{YY}, p_{XY}, p_{YZ}$, $p_{IZ}+p_{ZZ}, p_{IY}+p_{ZY}, p_{IZ}+p_{ZY}$, $p_{XI}+p_{XX}, p_{YI}+p_{YX}, p_{XI}+p_{YX}$
(d) Unlearnable degrees of freedom	$\lambda_{XI}, \lambda_{IZ}$

Table 4.1: A complete basis for the learnable linear functions of log Pauli fidelities and Pauli error rates for a single CNOT gate. This is a consequence of our master Theorem 4.1 and the PTG of CNOT given in Fig. 4.2.

Finally, the learnability of Pauli errors can be determined by the learnability of Pauli

fidelities according to the Walsh-Hadamard transform $p_a = \frac{1}{4^n}\sum_{b\in\mathsf{P}^n}\lambda_b(-1)^{\langle a,b\rangle}$. An issue here is that Pauli errors are linear functions of $\{\lambda_b\}$ instead of $\{\log\lambda_b\}$. Here we make a standard assumption in the literature [EWP+19, FW20] that the total Pauli error is sufficiently small. In this case all individual Pauli errors are close to 0 while all individual Pauli fidelities are close to 1. Therefore the Pauli errors can be estimated via

$$p_a = \frac{1}{4^n}\sum_{b\in\mathsf{P}^n}\lambda_b(-1)^{\langle a,b\rangle} \approx \frac{1}{4^n}\sum_{b\in\mathsf{P}^n}(-1)^{\langle a,b\rangle}\left(1+\log\lambda_b\right), \tag{4.9}$$

which means that their learnability can be determined by Theorem 4.1 plus a first-order approximation. Interestingly, the (first-order) learnable Pauli error rates have the same label as the learnable Pauli eigenvalues. We remark that this is related to the fact that the cycle space is invariant under the Walsh-Hadamard transform as proved in the supplemental materials of [ZCLJ25].

4.3.1 Experiments

We demonstrate our theory on IBM quantum hardware [ibm22] using a minimal example – characterizing the noise channel of a CNOT gate. In our experiments both the gate noise and SPAM noise are twirled into Pauli noise using randomized compiling. In the following we show how to extract all learnable information of Pauli noise SPAM-robustly, and also attempt to estimate the unlearnable degrees of freedom by making additional assumptions.

First, we conduct two types of cycle benchmarking (CB) experiments, the standard CB and CB with interleaving single-qubit gates (called *interleaved CB*), as shown in Fig. 4.4. The results are shown in Fig. 4.5. Here a set of two Pauli labels in the x-axis (*e.g.*, $\{IZ, ZZ\}$) corresponds to the geometric mean of the Pauli fidelity (*e.g.*, $\sqrt{\lambda_{IZ}\lambda_{ZZ}}$). Comparing to Table 4.1, we see that all learnable information of Pauli fidelities (including learnable individual and 2-product) are successfully extracted. Also note from Fig. 4.5 that the two types of CB

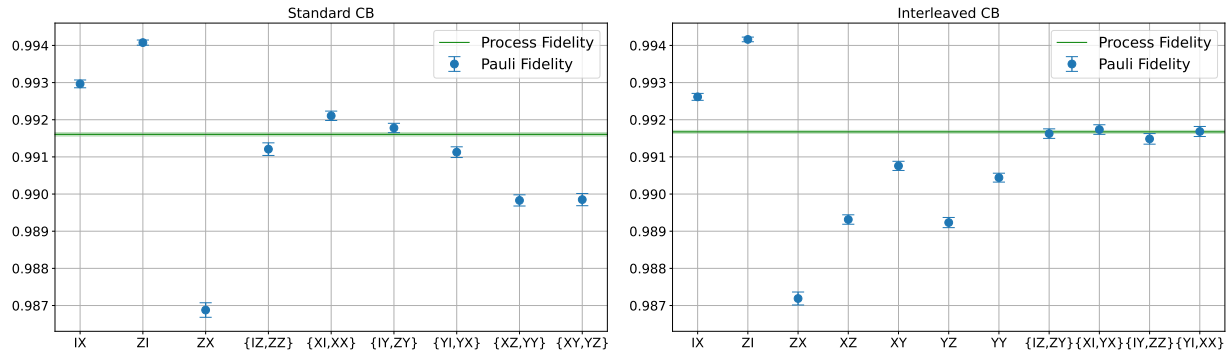
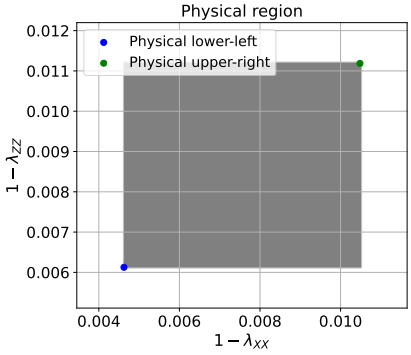


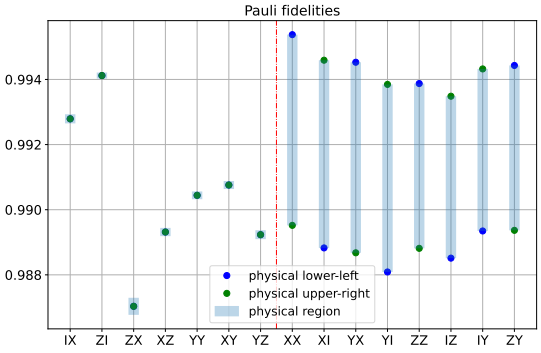
Figure 4.5: Estimates of Pauli fidelities of IBM's CNOT gate via standard CB (left) and CB with interleaved gates (right), using circuits shown in Fig. 4.4. Data are collected from `ibmq_montreal` on 2022-03-23. Each Pauli fidelity is fitted using seven different circuit depths $L = [2, 2^2, ..., 2^7]$. For each depth $C = 60$ random circuits and 1000 shots of measurements are used. Throughout this paper, the error bar represents the standard error.

experiments give consistent estimates, in terms of both the process fidelity and individual Pauli fidelities (*e.g.*, $\sqrt{\lambda_{XZ}\lambda_{YY}}$ estimated from standard CB is consistent with λ_{XZ} and λ_{YY} from interleaved CB).

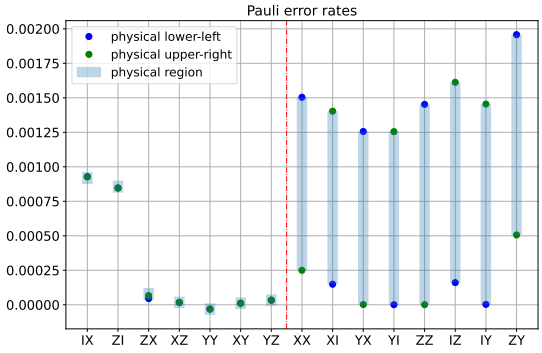
We can further bound the feasible region of the 2 unlearnable degrees of freedom using physical constraints, *i.e.*, the reconstructed Pauli noise channel must be completely positive. This is equivalent to requiring $p_a \geq 0$ for all Pauli error rates p_a. We choose λ_{XX} and λ_{ZZ} as a representation of the unlearnable degrees of freedom, and plot the calculated feasible region in Fig. 4.6 (a), which happens to be a rectangular area. We also calculate the feasible region for each unlearnable Pauli fidelity and Pauli error rate, which are presented in Fig. 4.6 (b), (c). In particular, we choose two extreme points (blue and green dots in Fig. 4.6 (a)) in the feasible region and plot the corresponding noise model in Fig. 4.6 (b), (c). Note that the (approximately) learnable Pauli error rates (on the left of the red vertical dashed line) are nearly invariant under change of gauge degrees of freedom, but they can be estimated to be negative due to statistical fluctuation. Thus, when we calculate the physical constraints, we only require those unlearnable Pauli error rates (on the right of the red vertical dashed line) to be non-negative.



(a) feasible region



(b) Pauli fidelities



(c) Pauli errors

Figure 4.6: Feasible region of the learned Pauli noise model, using data from Fig. 4.5. (a) Feasible region of the unlearnable degrees of freedom in terms of λ_{XX} and λ_{ZZ}. (b) Feasible region of individual Pauli fidelities. (c) Feasible region of individual Pauli errors.

Next, we explore an approach to estimate the unlearnable information with additional assumptions. Suppose that one can prepare $|0\rangle^{\otimes n}$ perfectly. Since we assume noiseless single-qubit gates, this means we can prepare a set of perfect tomographically complete states $\{|0/1\rangle, |\pm\rangle, |\pm i\rangle\}$. In this case, all the unlearnable degrees of freedom become learnable, as one can first perform a measurement device tomography, and then directly estimate the process matrix of a noisy gate with measurement error mitigated [MZO20]. Following this general idea, we propose a variant of cycle benchmarking for Pauli noise characterization, which we call *intercept CB* as it uses the information of intercept in a standard cycle benchmarking protocol. Given an n-qubit Clifford gate $\mathcal{G}$, let m_0 be the smallest positive

integer such that $\mathcal{G}^{m_0} = \mathcal{I}$. For any Pauli fidelity λ_a (regardless of whether learnable or not), consider the following two CB experiments using the standard circuit as in Fig. 4.4 (a). First, prepare an eigenstate of P_a, run CB with depth $lm_0 + 1$ for some non-negative integer l, and estimate the expectation value of $P_b \coloneqq \mathcal{G}(P_a)$. The result equals

$$\mathbb{E}\langle P_b\rangle_{lm_0+1} = \lambda_{P_a}^S \lambda_{P_b}^M \lambda_a \left(\prod_{k=1}^{m_0} \lambda_{\mathcal{G}^k(P_a)}\right)^l, \tag{4.10}$$

where $\lambda_{P_{a/b}}^{S/M}$ is the Pauli fidelity of the state preparation and measurement noise channel, respectively (earlier we have absorbed these two coefficients into a single coefficient A for simplicity). Second, prepare an eigenstate of P_b, run CB with depth lm_0, and estimate the expectation value of P_b. The result equals

$$\mathbb{E}\langle P_b\rangle_{lm_0} = \lambda_{P_b}^S \lambda_{P_b}^M \left(\prod_{k=1}^{m_0} \lambda_{\mathcal{G}^k(P_a)}\right)^l. \tag{4.11}$$

By fitting both $\mathbb{E}\langle P_b\rangle_{lm_0+1}$ and $\mathbb{E}\langle P_b\rangle_{lm_0}$ as exponential decays in l, extracting the intercepts (function values at $l = 0$), and taking the ratio, we obtain an estimator $\widehat{\lambda}_a^{\mathrm{ICB}}$ that is asymptotically unbiased to $\lambda_a \cdot \lambda_{P_a}^S / \lambda_{P_b}^S$. This estimator is robust against measurement noise. Note that $\lambda_{P_a}^S = \lambda_{P_b}^S = 1$ if we assume perfect initial state preparation, and in this case the above shows that λ_a is learnable, and thus the entire Pauli noise channel is learnable. We note that, instead of fitting an exponential decay in l, one could in principle just take $l = 0$ and estimate the ratio of $\mathbb{E}\langle P_b\rangle_0$ and $\mathbb{E}\langle P_b\rangle_1$, which also yields a consistent estimate for $\lambda_a \cdot \lambda_{P_a}^S / \lambda_{P_b}^S$. If one has already obtained all the learnable information from previous experiments, this could be a more efficient approach. However, if one has not done those experiments, the intercept CB with multiple depths can estimate the intercept (unlearnable information) and slope (learnable information) simultaneously, which is more sample efficient.

We numerically simulate intercept CB for characterizing the CNOT gate under different state preparation (SP) and measurement (M) noise. As shown in Fig. 4.7, this method yields relatively precise estimate when there is only measurement noise even if the noise is orders of magnitude stronger than the gate noise, but will have large deviation from the true noise model even under small state preparation noise. We refer the reader to Supplementary Section III for more details about the numerical simulation.

Finally, we experimentally implement intercept CB to estimate λ_{XX} and λ_{ZZ}, which are the two unlearnable degrees of freedom of CNOT, allowing us to determine all the Pauli fidelities and Pauli error rates. One challenge in interpreting the results is that we do not know in general whether the low SP noise assumption holds, therefore it is unclear if the learned results should be trusted. However, for the estimate to be correct, it should at least lie in the physically feasible region we obtained earlier in Fig. 4.6. In Fig. 4.8, we present our experimental results of intercept CB. It turns out that certain Pauli fidelities are far away from the physical region by several standard deviations. This gives strong evidence that the low SP noise assumption was *not* true on the platform we used.

The data collected here can further be used to give a lower bound for the SP noise. Suppose we obtain the physical region of λ_a to be $[\widehat{\lambda}_{a,\min}, \widehat{\lambda}_{a,\max}]$. Combining with the expression of intercept CB, we have

$$\widehat{\lambda}_a^{\mathrm{ICB}}/\widehat{\lambda}_{a,\max} \leq \lambda_{P_a}^S/\lambda_{P_b}^S \leq \widehat{\lambda}_a^{\mathrm{ICB}}/\widehat{\lambda}_{a,\min}. \tag{4.12}$$

Applying this to the data of IZ and ZZ in Fig. 4.8 (a), we have $\lambda_{IZ}^S/\lambda_{ZZ}^S \leq 0.9879(23)$. If we make a physical assumption that the state preparation noise is a random bit-flip during the qubit initialization, one can conclude the bit-flip rate on the first qubit is lower bounded by 0.61(12)%. One can in principle bound the bit-flip rate on the second qubit by looking at $\lambda_{XX}^S/\lambda_{XI}^S$. Unfortunately, our estimate of λ_{XX}^S from intercept CB falls in the physical region within one standard deviation, so there is no nontrivial lower bound. One could

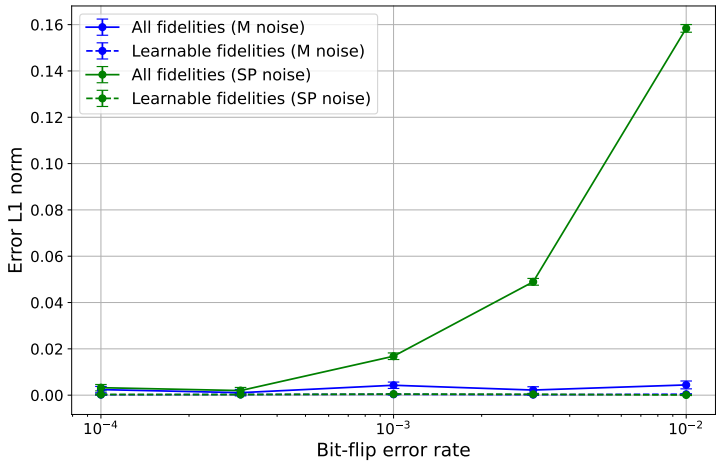


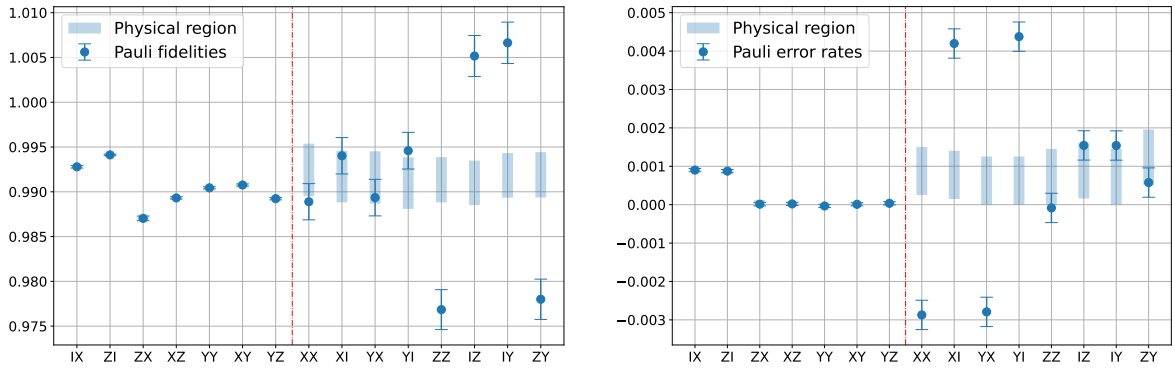
Figure 4.7: Simulation of intercept CB on CNOT under different SPAM noise rate. The simulated noise channel is a 2-qubit amplitude damping channel with effective noise rate 5%, and SPAM noise are modeled as bit-flip errors. For the blue (green) lines, we introduce random bit-flip errors to the measurement (state preparation). The solid lines show the l_1-distance of the estimated Pauli fidelities from the true Pauli fidelities. The solid lines show the l_1-distance of the (individually) learnable Pauli fidelities from the ground truth.

expect obtaining a useful lower bound by looking at a CNOT gate with reversed control and target. The lower bound of SP noise obtained here is completely independent of the measurement noise and does not suffer from the issue of gauge freedom [NGR+21], as long as all of our noise assumptions are valid, *i.e.*, there is no significant contribution from time non-stationary, non-Markovian, or single-qubit gate-dependent noise.

4.4 Discussion

4.4.1 Related works

Learning the Pauli noise channel Λ affecting a Clifford gate $\mathcal{G}$ (i.e., $\tilde{\mathcal{G}} = \Lambda \circ \mathcal{G}$) in the presence of unknown state-preparation-and-measurement (SPAM) noise is a standard task studied in multiple references [FW20, FO21, HFW20, Fla22, VDBMKT23, CLO+23, CDDH+23, vdBW24, HDH25]. When the ideal Clifford gate $\mathcal{G}$ equals to the identity, it has been shown that Λ can be fully determined independent of the SPAM noise [FW20]; When $\mathcal{G}$ is a non-



(a) (b)

Figure 4.8: The learned Pauli noise model using intercept CB. The feasible region (blue bars) are taken from Fig. 4.6. Estimates of Pauli fidelities (a) and Pauli error rates (b). Each data point is fitted using seven different circuit depths $L = [2, 2^2, ..., 2^7]$. For each depth $C = 150$ random circuits and 2000 shots of measurements are used. Data are collected from `ibmq_montreal` on 2022-03-23.

trivial Clifford gate, it is known that certain properties of Λ (e.g., the average fidelity) can be learned independent of SPAM [EWP$^+$19], but no existing protocols can completely and SPAM-robustly learn Λ in this case. In fact, it was pointed out [CLO$^+$23, HFP22] that this is a fundamental limitation related to the model non-identifiability, also known as gauge ambiguity in the literature on gate set tomography [NGR$^+$21]. Specifically, [CLO$^+$23] have developed a theory that completely characterize the degrees of freedom of a Pauli noise channel associated to any set of Clifford gates and provides a "gauge-consistent" algorithm to learn all the learnable degrees of freedom. The main limitation in [CLO$^+$23] is that it considers the case that each n-qubit gate has a generic n-qubit Pauli noise channel, which has exponentially many parameters, and consequently the learning algorithm would take exponential time. In contrast, efficient parameterization of Pauli noise channels have been studied in the literature and applied in experiments [Fla22, VDBMKT23, CDDH$^+$23, WKBK23, HDH25, PCOG$^+$24], but a good understanding about the gauge and learnable degrees of freedom in such models is yet to be obtained, which hinders the application of these protocols (specifically, how to learning such models self-consistently). Our work combines these two lines of

research to achieve a self-consistent and efficient Pauli noise learning algorithm.

Our framework resembles ACES (i.e., averaged circuit eigenvalue sampling), a Pauli noise learning protocol proposed in [Fla22] and further analyzed and implemented in [COG+24] and [HDH25]. Similar to our setting, ACES models every different Clifford gate as having a gate-dependent Pauli noise channel (with an efficient noise ansatz such as fully local noise), and aims at learning every Pauli eigenvalue of every Pauli noise channel. ACES constructs linear equations of the form $\boldsymbol{b} = A\boldsymbol{x}$, where $\boldsymbol{x}$ are the logarithmic Pauli eigenvalues, $\boldsymbol{b}$ are the logarithmic circuit eigenvalues which can be experimentally measured, and A is the design matrix determined by a set of Clifford circuits and Pauli measurements. The original ACES effectively fixes a gauge by ignoring state preparation noise, and is thus able to construct a full column-rank A using random Clifford gates. Within ACES's framework, our work basically says that if SPAM noise parameters are included in $\boldsymbol{x}$, there is a maximum possible rank of A equal to the number of learnable degrees of freedom, and we provide an explicit recipe for constructing such an A. Moreover, the explicit construction can learn gate parameters to relative precision; see Theorem 2.4. It is interesting to extend the statistical analysis for ACES [HDH25] to our setting. It is also interesting to study optimal learning and experimental design within our framework [vdBW24, ORS+23]. Conversely, it is interesting to study the ultimate sample complexity limit for learning gate set Pauli noise models. Such bounds for learning Pauli channels in a black-box setting have been previously investigated [CZSJ22, COZ+24a, FOF23, CG23a, TY24].

4.4.2 Outlook

Quantum error mitigation. One promising practical application of our theory is in quantum error mitigation (QEM) [CBB+23]. There have been QEM protocols based on the Pauli noise models [VDBMKT23, FHV+24, KEA+23, CYZF21] which work by first learning the Pauli noise channels and then either canceling the noise via quasi-probability

sampling, or amplifying the noise and then extrapolating to the noiseless limit. Due to the existence of gauge, these protocols generally make assumptions that are not physically justified (e.g., "symmetry assumption" [VDBMKT23, vdBW24] or perfect state preparation [CYZF21]) in order to determine the noisy Pauli channels. However, it is known that the issue of gauge ambiguity should not hinder self-consistent QEM if assisted by gate set tomography [EBL18, SEFT22], although the latter method is not easily scalable. Using our framework, by efficiently learning all the learnable degrees of freedom and anchoring at an arbitrary point of the gauge, one will be able to mitigate the Pauli noise without any additional assumptions.

Beyond Pauli noise model. We believe that our framework and graph-theoretic analysis can be extended beyond the Pauli noise model. We expect the statement that the cycle space equals the learnable space to be held more generally. Note that a Pauli noise model has a linear parameterization (i.e., the log of the Pauli eigenvalues), and one can study the learnable and gauge degrees of freedom as linear subspaces of the parameter spaces. For more general noise models, one might need to study the tangent space of the parameters, which has been considered in the theory of the first-order gauge-invariant (FOGI) model [NYBK22, BKdSN$^+$22]. It is interesting to see if our method can be extended there to have a more quantitative understanding of the learnable and gauge degrees of freedom.

Another important potential extension is to include mid-circuit measurement (MCM, also known as quantum non-demolition measurement), which is a crucial building block for fault-tolerant quantum computation and is being experimentally implemented [RABL$^+$21, SBA$^+$23, AAA$^+$23, BEG$^+$24]. It has been shown that MCM can be incorporated into Pauli noise models via a generalized randomized compiling technique [BW23]. More recently, results on the learnability of MCM within the Pauli noise model have been reported [ZCLJ25, HP25]. It is practically interesting to incorporate those results into the current framework, e.g., to have a self-consistent and efficient protocol for characterizing

MCM with a scalable noise ansatz. More generally, it is interesting to study how our framework can be used for noise characterization of early fault-tolerant quantum computing devices [CGFF17, WKBK23].

Beyond realizable setting. The learning algorithm we proposed relies on the fact that the noise model of interest is realizable, i.e., there is no out-of-model error. It would be interesting to analyze the effect of such errors. Our current analysis and algorithms require all the noise to be Pauli channels with certain constraints, specifically, the locality constraints as discussed in Sec. B.4, B.5, and B.6. In practice, one can fit the same data into models with different locality assumption, and there will be a tradeoff between model expressivity and trainability. One can also utilize techniques from structure learning to find the optimal (quasi-)local Pauli noise model (see e.g. [RF23]). It would be interesting to explore learning algorithms with those features within our framework.

CHAPTER 5

ENTANGLEMENT-ENHANCED QUANTUM NOISE CHARACTERIZATION

Chapter Notes. This chapter presents a protocol that applies the entanglement-enabled quantum learning advantages in practical Pauli noise characterization tasks. A preliminary version of the protocol and related theoretic results is presented in [CZSJ22]. The protocol has then been substantially improved and experimentally implemented in a recent collaboration with IBM Quantum [SCM+24]. The presentation here is mainly based on the latter manuscript. The experiments are conducted by my IBM Quantum colleagues, while I lead the theory development and participating in analyze and processing the data.

5.1 Introduction

In the previous section, we have explored the task of Pauli noise characterization. According to Theorem 3.2, a generic n-qubit Pauli channels is exponentially hard to learn using ancilla-free schemes (which are schemes used for most noise characterization protocols). To circumvent this barrier, one common approach is to make locality assumptions of the noise and reduce the number of noise parameters, just as the *reduced models* discussed in the last section. However, such locality assumptions might not always be justified. There could be unexpected long-range correlation overlooked by a local model. One might still want to learn the whole noise channel without any assumptions to diagnose such uncommon effect of noise.

Recall that in Chapter 3 we show entangled quantum memory can exponentially speedup learning of Pauli channels. A natural question is whether we can use such an entanglement-enhanced learning (EEL) protocol to overcome the exponential barrier for Pauli noise learning in practice. This might not seem practically feasible at the first glimpse: the entangled states, measurements, and ancillary quantum memory are all subject to noise in practice. How can

we precisely characterize one interested source of noise (gate-dependent Pauli noise) while introducing many other sources of noise (SPAM noise, ancilla noise, etc.)? It turns out that, by developing several quantum error suppression and mitigation techniques, our EEL protocol can not only estimate the gate noise consistent with state-of-the-art entanglement-free learning (EFL) protocols, but also achieve this which much fewer number of samples, as validated by experiments. Our protocol is an manifestation of robust and significant quantum learning advantages in a practical information processing task.

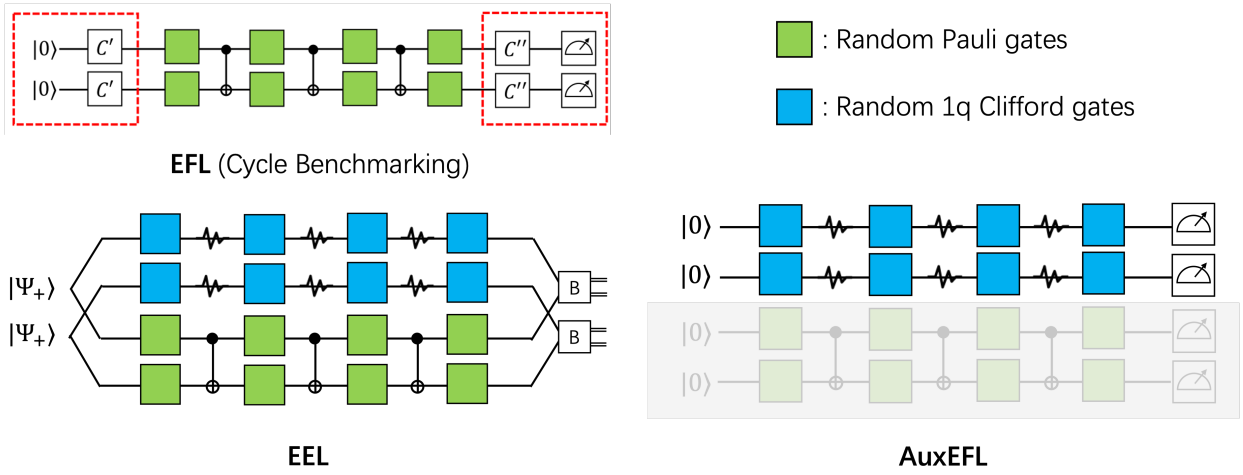


Figure 5.1: Schematics of EFL and EMEEL (which consists of EEL and AuxEFL). (Top left) EFL, or the standard cycle benchmarking circuits, which consists of stabilizer state preparation and Pauli measurements. Green boxes represent random Pauli twirling gates. (Bottom left) EEL, consisting of Bell state and bell measurements. The blue boxes represent random single-qubit Clifford twirling gates. The giggle on the ancilla denotes dynamical decoupling for idling protection. (Bottom right) AuxEFL, whose circuit is the same as EEL, but with computational basis state preparation and measurements. The combination of EEL and AuxEFL gives our EMEEL protocols.

5.2 Setup

Let A, S each be an n-qubit quantum system, referred to as ancillary system and the main system, respectively. In a high level, our goal is to characterize the noise associated with certain multi-qubit Clifford gate $\mathcal{G}$ that can be implemented on system S. We will make the

following assumptions about the noise model:

1. Layers of any single-qubit Clifford gates have negligible noise.

2. The noises acting on the system and ancilla is time-stationary, Markovian, and have no crosstalk or contextual dependence between each other. Specifically, applying the multi-qubit Clifford gate $\mathcal{G}$ on the system and idling the ancilla is realized as

$$\mathcal{I}^A \otimes \mathcal{G}^S \mapsto \widetilde{\mathcal{I}}^A \otimes \widetilde{\mathcal{G}}^S \coloneqq \Lambda_{\text{anc}}^A \otimes (\Lambda_{\mathcal{G}} \circ \mathcal{G})^S$$

3. The initial state and measurement suffer from noise channel acting jointly on the system and ancilla,

$$\rho_0^{AS} \mapsto \mathcal{E}^{\mathsf{S}}(\rho_0^{AS}), \quad \mathcal{M}^{AS} \mapsto (\mathcal{M} \circ \mathcal{E}^{\mathsf{M}})^{AS}.$$

The first assumption are standard [WE16, HNM$^+$21, CLO$^+$23, FW20]. The second assumption requires the system and ancilla to be well isolated, which can be ensured via dynamical decoupling, turning off certain couplers on superconducting platforms, or moving qubits far apart on ion or atom platforms, etc. The goal is to characterize the gate noise $\Lambda_{\mathcal{G}}$ in a SPAM-robust manner. Since we are going to apply Pauli twirling, we will only be interested in learning the Pauli eigenvalue of the Pauli twirl of $\Lambda_{\mathcal{G}}$, defined as $\{\lambda_a^{\mathcal{G}}\}$. We also denote the Pauli eigenvalue of λ_{anc} as $\{\lambda_a^{\text{anc}}\}$.

It has been proved that not every $\lambda_a^{\mathcal{G}}$ can be learned SPAM-independently for a generic Clifford gate $\mathcal{G}$ [CLO$^+$23, HFP22], which is a fundamental limitation posted by model identifiability that is generic to any learning scheme. Take $\mathcal{G} = \text{CNOT}$ as an example, one can show that $\lambda_{XX}, \lambda_{XI}$ cannot be individually learned SPAM-independently, while their product $\lambda_{XX}\lambda_{XI}$ can be. In our work, we only focus on learning a subset of learnable degrees of freedom of $\Lambda_{\mathcal{G}}$. Formally, we aim at learning the following quantities,

Definition 5.1 (Symmetrized Pauli eigenvalues)**.** *Given an n-qubit Clifford gate $\mathcal{G}$, let d_0 be*

the smallest positive integer such that $\mathcal{G}^{d_0} = \mathcal{I}$. *We define the* symmetrized Pauli eigenvalues *as follows:*

$$\bar{\lambda}_a \coloneqq \left(\prod_{k=0}^{d_0-1} \lambda_{\mathcal{G}^k(a)} \right)^{1/d_0}, \quad \forall a \in \mathsf{P}^n. \tag{5.1}$$

where we drop the potential $\pm$ *sign of* $\mathcal{G}^k(a)$.

As an example, the symmetrized Pauli eigenvalues of CNOT are listed in Table 5.1.

a	$\bar{\lambda}_a$
ZI	λ_{ZI}
IX	λ_{IX}
ZX	λ_{ZX}
IZ, ZZ	$(\lambda_{IZ}\lambda_{ZZ})^{1/2}$
XI, XX	$(\lambda_{XI}\lambda_{XX})^{1/2}$
XZ, YY	$(\lambda_{XZ}\lambda_{YY})^{1/2}$
YI, YX	$(\lambda_{YI}\lambda_{YX})^{1/2}$
IY, ZY	$(\lambda_{IZ}\lambda_{ZY})^{1/2}$
XY, YZ	$(\lambda_{XY}\lambda_{YZ})^{1/2}$

Table 5.1: Symmetrized Pauli eigenvalues for a single CNOT gate.

The task we are about to deal with is the following:

Task 5.2 (Gate-dependent Pauli noise estimation)**.** *For a given multi-qubit Clifford gate* $\mathcal{G}$*, under the aforementioned assumptions, estimate the symmetrized Pauli eigenvalues* $\bar{\lambda}_a^{\mathcal{G}}$ *in a SPAM-robust manner.*

We remark that learning the symmetrized Pauli eigenvalues is a standard task in the literature [EWP+19, HNM+21]. By interleaving with carefully chosen single-qubit Clifford gates, one can learn more than those symmetrized eigenvalues [VDBMKT23, CLO+23, vdBW24]. Such protocols can also be incorporated with EMEEL, but we leave such demonstration for future work. Also note that, even learning those symmetrized Pauli eigenvalues is proved to cost exponentially many measurements for EFL schemes [COZ+24a, Theorem 2].

5.3 Protocols

In this section, we describe Error-Mitigated Entanglement-Enhanced Learning (EMEEL), which is our main protocol. We give an intuitive explanation in this section, leaving a rigorous treatment to the technical details section.

To understand how EMEEL works, it is helpful to first look at the state-of-the-art entanglement-free learning protocols, Cycle Benchmarking [EWP+19] or Cycle Error Reconstruction [CDDH+23], see Fig. 5.1. We use EFL to refer to these protocols. The key step is to prepare certain (noisy) stabilizer state, repeat the noisy gate $\widetilde{\mathcal{G}}$ multiple times (interleaved with Pauli twirling gates), and measure (noisy) Pauli observables. With appropriate choices of the input and measurements, we obtain an expectation value of the form $\xi\lambda^m$, where ξ is a SPAM noise-dependent factor, λ is certain Pauli eigenvalue of $\Lambda^{\mathcal{G}}$, and m is the number of repetitions. By choosing different values of m and fit the circuit expectation values to this exponential model, one can obtain a *SPAM-robust* estimation for λ. The main sample complexity stems from the need to switch among exponentially many input state and measurement configurations.

Now we can explain EMEEL. It consists of two subroutines: EEL and AuxEFL (the latter stands for auxiliary entanglement-free learning), see Fig. 5.1. EEL uses a similar repeating-and-fitting strategy as EFL, but uses Bell states and Bell measurements with an ancillary system. Thanks to the entangled memory, only 1 state preparation and measurement configuration is needed to estimate all Pauli eigenvalues SPAM-robustly.

In practice, there will be ancilla idling noise besides the SPAM noise. What is learned by EEL is composed by both the gate noise and ancilla noise. Our strategy for that is to first apply dynamical decoupling [SLT+24] and single-qubit Clifford twirling on the ancilla, suppressing the noise as much as possible, and then apply *ancilla error mitigation* as follows: First conduct AuxEFL on the ancilla, which is basically EFL for identity gate on the ancilla systems with the random Pauli twirling gates replaced by random single-qubit Clifford gates.

Thanks to the Clifford twirling, the noise channel on the ancilla is tailored as a symmetric Pauli channel (i.e., λ_a depends only on $\mathrm{pt}(a)$), which can be learned sample-efficiently. We then use the learned information from AuxEFL to mitigate ancilla errors in EEL, thus obtaining the final estimators of EMEEL.

5.4 Experiments

Here we present some experimental results from [SCM+24] to justify the robustness and efficiency of EMEEL, contrasting it with standard EFL.

Learning noise of a two-qubit gate. The first experiments validate EMEEL in noise characterization for a two-qubit ECR gate. We learned all the symmetrized Pauli eigenvalues using both EMEEL and EFL. The goal is to see if the estimates from both scheme match each other, which would indicate we successfully mitigate the ancilla and memory errors. Note that EFL need 9 state preparation and measurement settings while EMEEL need 2. The results are reported in Fig. 5.2. For EEL without additional error mitigation, a mean absolute error of $(3.84 \pm 0.14) \times 10^{-3}$ is reported. For EMEEL (error mitigated with AuxEFL data), the error is reduced to $(4.4 \pm 1.4) \times 10^{-4}$. This suggests that EMEEL can indeed provide convincing estimates.

Learning noise of parallel two-qubit gates. While EMEEL drastically reduces the number of measurement settings from an exponential in n for protocols without memory to 2 (one for EEL and another one for AuxEFL as shown in Fig. 5.2b and c), for each experiment setting, EMEEL has a larger variance due to ancillary and memory error as well as the error mitigation overhead. The central question is then how this overhead compared to the saving of measurement settings. For EMEEL to beat EFL, the saving in measurement settings must be more drastic compare to the increase of variance at each setting. To validate this, we conduct learning experiments of up to 4 parallel two-qubit gates. see Fig. 5.3.

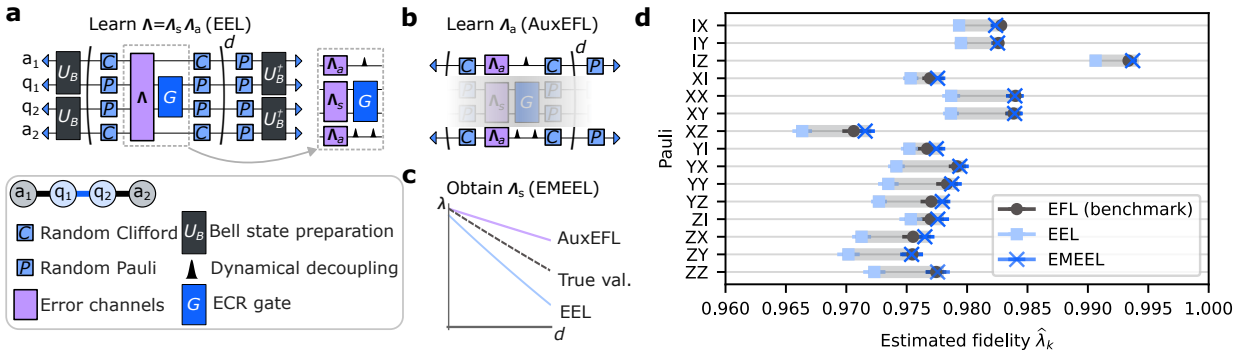


Figure 5.2: **Learning the noise of a two-qubit gate.** (Figures courtesy of Alireza Seif.) (a)(b) shows the experiment layout. (c) gives a illustration of the fidelity decay curve (in log scale) with number of circuit depth d. In this experiments we use $d = 0, 16, 32$. (d) Report the estimation for all symmetrized Pauli eigenvalues using EFL, EEL, and EMEEL. One can see that while EEL significantly overestimate the fidelities compared to EFL, EMEEL almost retrieves consistent estimates as EFL. This experiments were conducted on *ibm_peekskill* using 100 twirls and 1000 measurement shots at each depth $d \in \{0, 16, 32\}$.

For up to 8 qubits, it becomes challenging to measure all Pauli eigenvalues. For validation purpose, we only measure and compare eigenvalues consists solely of X and I or Z and I. Again, we see EEL has an error of $(2.081 \pm 0.004) \times 10^{-2}$ while EMEEL has an error of $(1.59 \pm 0.04) \times 10^{-3}$, compared to be EFL benchmark. For the overhead, we empirically fit that for a measurement setting of a weight w Pauli eigenvalue, the variance from EMEEL is $(1.33 \pm 0.05)^w$ times larger than EFL. On the other hand, the saving in measurement setting for EMEEL is at least 2^w by our rigorous lower bound. In Sec. C.3, we present more detailed analysis of the resource cost. These justify the efficiency and robustness of EMEEL.

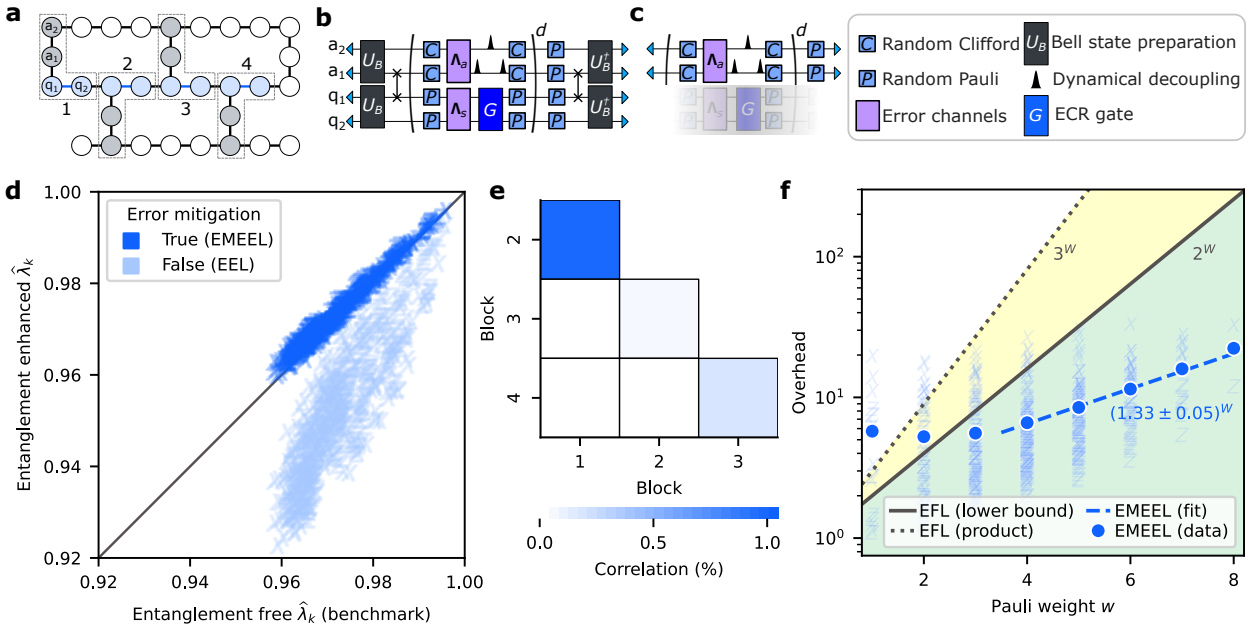


Figure 5.3: **Learning the noise of parallel two-qubit gates.** (Figures courtesy of Alireza Seif.) (a) Shows the our qubit topology. The blue qubits are our main systems where parallel gates act on, and the gray qubits are used as ancilla. (b)(c) layouts for EEL and AuxEFL. Note that an additional pair of SWAP gates are need to create the Bell state and measurement due to our qubit topology. (d) Comparing the estimates from EFL (horizontal) with EEL and EMEEL (vertical). We only measure all X-type and all Z-type eigenvalues. While EEL significantly underestate fidelities, EMEEL only slightly overestimates. (e) Using learned information to validate correlated noise on the systems. We omit discussion on this block and direct the readers of interest to [SCM+24]. (f) Comparing the variance overhead (fit as 1.33^w) with the saving of measurement setting by EMEEL. The variance is estimated for each all X-type and all Z-type eigenvalues. Here 3^w referes to naive entanglement-free strategies of going through all Pauli basis, while 2^w is a rigorous lower bound for entanglement-free schemes.

CHAPTER 6

GAUGE-CONSISTENT QUANTUM ERROR MITIGATION

Chapter Notes. This chapter is based on an on-going work about applying the self-consistent Pauli noise learning framework into quantum error mitigation [CCF$^+$25]. The experimental data presented here are taken by Edward Chen, Laurin Fischer, Alireza Seif, and other collaborators from IBM Quantum.

6.1 Introduction

We have seen in the last few chapters how to characterize quantum noise. Eventually, one wants to use the learned noise information to improve the performance of a quantum information processing system. A paradigm of methods that is particularly suitable for near-term quantum computers is known as *Quantum Error Mitigation* (QEM) [EBL18, CXB20]. Generally speaking, QEM uses additional classical computing resources to reduced bias in the outcome of noisy quantum computation. While QEM alone is not able to replace quantum error correction or fault-tolerance computation due to its intrinsic exponential costs [TSY23, QSFK$^+$24], there are demonstration that QEM can enhance the utility of existing and near-future quantum computers with limited number of qubits and limited capacity of fault tolerance [KEA$^+$23, AAA$^+$25]. What's more, many existing error mitigation protocols are learning-based, meaning that the protocol first learn the noise model and then use the learned information to mitigate errores. For such protocols, QEM can be a validation tool for noise modeling and learning, as imprecise QEM results can indicate errors not captured by the noise model.

As discussed in details in Chapter. 4, a fundamental issue of quantum noise learning is the existence of non-identifiable or gauge parameters, preventing one from completely learning the noise models. A natural question is how such "unlearnability" would af-

fect the performance of learning-based QEM. As an example, let us look at one popular learning-based quantum error mitigation method, named *probabilistic error cancellation*(PEC) [TBG17, EBL18, VDBMKT23]. PEC cancels any noise channels Λ affecting quantum operations by implementing the inverse channels Λ^{-1}. Though Λ^{-1} is typically not a physical channel due to the violation of positivity, it can be implemented in expectation via quasi-probability sampling and classical post-processing. However, since implementing the inverse channel requires full knowledge of the noise channel, the existence of unlearnable parameters seem to be a problem for PEC to work. Indeed, in a series of recent work about Pauli noise based PEC [VDBMKT23, FHV$^+$24, KEA$^+$23], they have to circumvent the learnability issue by imposing additional assumption of the noise channel that are not physically justified.

In contrast, a more appropriate way to deal with the learnability issue has been hinted in the earlier literature of PEC [EBL18]. Intuitively, if certain noise parameters are unlearnable within the model, their values should not affect any observable outcomes. This means if one correctly learns all the learnable parameters and arbitrarily decide the unlearnable (gauge) parameters, it suffices to predict any circuits within the model, and one should be able to correctly conduct QEM without introducing additional assumptions. Concretely, [EBL18] proposes combining *gate set tomography* (GST) [NGR$^+$21], a generic method to learn the learnable noise parameters, with PEC. However, due to the extremely large resource cost of GST, it is hard to scale up the method beyond a few qubits.

Building the framework of Gate set Pauli noise learning we developed (Chapter. 4), we propose a Gauge-consistent PEC protocols based on Pauli noise models. On one hand, our protocol works for Pauli noise models with arbitrary locality, thus scalable to a large system size, comparable to [KEA$^+$23, VDBMKT23]. On the other hand, our protocol is gauge-consistent in the sense that only learnable parameters is needed and no unjustified assumptions are needed. We experimentally implement and compare our protocol with

existing protocol from [VDBMKT23] for specific families of quantum circuits, and we see our method significantly reduce the bias in PEC results, highlighting the importance of gauge consistency.

6.2 Theory

Setup. We consider a similar model as in Chapter 4, restated in below. For an n-qubit system, we consider the following set of operations and their noisy implementation.

1. Initialization: $|0\rangle\langle 0| \mapsto \tilde{\rho}_0 = \Lambda_S(|0\rangle\langle 0|)$.
2. Computational-basis measurement: $\{|b\rangle\langle b|\}_{b\in\{0,1\}^n} \mapsto \{\tilde{E}_b = \Lambda_M(|b\rangle\langle b|)\}_{b\in\{0,1\}^n}$.
3. Layer of single-qubit unitary: $\mathcal{U} = \otimes_{k=1}^n \mathcal{U}_k$, implemented without noise.
4. Layer of multi-qubit Clifford: $\mathcal{G} \mapsto \tilde{\mathcal{G}} = \Lambda_{\mathcal{G}} \circ \mathcal{G}$, for all $\mathcal{G}$ from a finite set $\mathfrak{G}$.

Here, we further assume $\Lambda_{\mathcal{G}}$ are $\mathcal{G}$-dependent Pauli channels, and Λ_S, Λ_M are symmetric Pauli channels (i.e., λ_a only depends on $\mathrm{pt}(a)$). We use $\mathbf{\Lambda}$ to denote the collection of all noise channels. Again, we assume all the Pauli eigenvalues are strictly positive. All these assumptions are experimentally reasonable based on the use of randomized compiling [WE16].

Standard PEC. Suppose we want to estimate the expectation value of certain observable O at the output state of a quantum circuit (which is a natural task in, say, Hamiltonian simulation). Denote the ideal value as follows,

$$o = \langle\langle O|\mathcal{G}_T\mathcal{U}_T \cdots \mathcal{G}_1\mathcal{U}_1|\rho_0\rangle\rangle. \tag{6.1}$$

Here $\mathcal{U}_j$'s are layers of (possibly non-Clifford) single-qubit gates, and $\mathcal{G}_j$'s are layers of multi-qubit Clifford gates from $\mathfrak{G}$. Because of noise, a direct execution of the above gate sequence

will instead give

$$\begin{aligned} o^{(\text{noisy})} &= \langle\langle \widetilde{O}|\widetilde{\mathcal{G}}_T\mathcal{U}_T\cdots\widetilde{\mathcal{G}}_1\mathcal{U}_1|\widetilde{\rho}_0\rangle\rangle \\ &= \langle\langle O|\Lambda_M\Lambda_{\mathcal{G}_T}\mathcal{G}_M\mathcal{U}_M\cdots\Lambda_{\mathcal{G}_1}\mathcal{G}_1\mathcal{U}_1\Lambda_S|\rho_0\rangle\rangle. \end{aligned} \tag{6.2}$$

To retrieve the ideal value, a naive idea is to cancel out all noise channels Λ by implementing Λ^{-1}. For any Pauli channel $\Lambda = \sum_b \lambda_b |P_b\rangle\rangle\langle\langle P_b|/2^n$, its inverse is $\Lambda^{-1} = \sum_b \lambda_b^{-1}|P_b\rangle\rangle\langle\langle P_b|/2^n$, which can be expressed as

$$\Lambda^{-1}(\rho) = \sum_{a\in\mathsf{P}^n} p_a^\star P_a\rho P_a, \quad \text{where } \; p_a^\star = \frac{1}{4^n}\sum_{b\in\mathsf{P}^n}(-1)^{\langle a,b\rangle}\lambda_b^{-1}. \tag{6.3}$$

This is a Pauli diagonal map that is not necessarily completely-positive (i.e., $p_a^\star$ can be negative). Consequently, it cannot be directly implemented as a quantum channel. Instead, one can rewrite it in the following form

$$\begin{aligned} \Lambda^{-1}(\rho) &= \sum_a \left(\sum_b |p_b^\star|\right)\frac{\mathrm{sgn}_a|p_a^\star|}{\sum_b|p_b^\star|}P_a\rho P_a \\ &= \sum_a \gamma \mathrm{sgn}_a q_a P_a\rho P_a, \end{aligned} \tag{6.4}$$

where sgn_a is the sign of $p_a^\star$, $\gamma = \sum_b |p_b^\star|$, and $q_a = |p_a^\star|/\gamma$. Note that $\{q_a\}$ forms a probability distribution over P^n. Thus, by sampling Pauli operator P_a according to $\{q_a\}$ and multiplying $\gamma\mathrm{sgn}_a$ in classical post-processing, one can implement Λ^{-1} in expectation. This is the core idea of PEC.

Concretely, consider the following steps of *standard PEC* (which has assumed all noise channels are known a priori):

1. Randomly sample $a_0 \sim \{q_{a_0}^S\}$, $a_j \sim \{q_{a_j}^{\mathcal{G}_j}\}_{j=1}^T$, $a_{T+1} \sim \{q_{a_{T+1}}^M\}$.

2. Implement and measure the following expectation value

$$E_{\boldsymbol{a}} = \langle\langle \widetilde{O}|\mathcal{P}_{a_{T+1}}\mathcal{P}_{a_T}\widetilde{\mathcal{G}}_T\mathcal{U}_T\cdots\mathcal{P}_{a_1}\widetilde{\mathcal{G}}_1\mathcal{U}_1\mathcal{P}_{a_0}|\widetilde{\rho}_0\rangle\rangle \tag{6.5}$$

where $\mathcal{P}_a(\rho) = P_a \rho P_a$.

3. Define the PEC estimator as

$$\hat{o}^{(\mathrm{PEC})} = E_{\boldsymbol{a}} \cdot \prod_{j=0}^{T+1} \gamma_j \mathrm{sgn}_{a_j}. \tag{6.6}$$

where γ_j and sgn_{a_j} are with respect to the jth Pauli noise channel.

The following proposition shows the correctness of PEC.

Proposition 6.1. *Given that one knows* $\boldsymbol{\Lambda}$ *exactly, the standard PEC estimator* $\hat{o}^{(\mathrm{PEC})}$ *is an unbiased estimator for* o*.*

Proof.

$$\begin{aligned}
\mathbb{E}\hat{o}^{(\mathrm{PEC})} &= \sum_{\boldsymbol{a}} E_{\boldsymbol{a}} \cdot \prod_{j=1}^{T+1} \gamma_j \mathrm{sgn}_{a_j} q_{a_j} \\
&= \sum_{\boldsymbol{a}} E_{\boldsymbol{a}} \cdot \prod_{j=1}^{T+1} p^{\star}_{a_j} \\
&= \sum_{\boldsymbol{a}} \langle\!\langle \widetilde{O} | p^{\star,M}_{a_{T+1}} \mathcal{P}_{a_{T+1}} p^{\star,\mathcal{G}_T}_{a_T} \mathcal{P}_{a_T} \widetilde{\mathcal{G}}_T \mathcal{U}_T \cdots p^{\star,\mathcal{G}_1}_{a_1} \mathcal{P}_{a_1} \widetilde{\mathcal{G}}_1 \mathcal{U}_1 p^{\star,S}_{a_0} \mathcal{P}_{a_0} | \widetilde{\rho}_0 \rangle\!\rangle \qquad (6.7) \\
&= \langle\!\langle \widetilde{O} | \Lambda_M^{-1} \Lambda_{\mathcal{G}_T}^{-1} \widetilde{\mathcal{G}}_T \mathcal{U}_T \cdots \Lambda_{\mathcal{G}_1}^{-1} \widetilde{\mathcal{G}}_1 \mathcal{U}_1 \Lambda_0^{-1} | \widetilde{\rho}_0 \rangle\!\rangle \\
&= \langle\!\langle O | \mathcal{G}_T \mathcal{U}_T \cdots \mathcal{G}_1 \mathcal{U}_1 | \rho_0 \rangle\!\rangle \\
&= o.
\end{aligned}$$

The second line is by the definition of q_a. □

We remark that, PEC also works for Pauli noise channels with efficient ansatzes, such as 2-local noise models [VDBMKT23, KEA+23], so that it can scales up to much larger system size. A key ingredient there is a method to efficiently sample from $\{q_a\}$ via the so-called sparse Pauli-Lindblad model. We will omit the discussion here.

Regarding the sample efficiency of PEC, for observable O with bounded spectral norms, PEC needs to pay a sampling overhead of order $\prod_j \gamma_j^2$ to get constant additive precision, which is exponential in system size for fixed error rates. Such exponential scaling is shown to be inevitable for generic QEM protocols [QSFK$^+$24, TSY23].

Gauge-consistent PEC. Clearly, the aforementioned standard PEC procedure requires full knowledge of the noise channels $\mathbf{\Lambda}$, which usually needs to be learned from the model. In Chapter 4, we have proven that there exists unlearnable degrees of freedom in the Pauli noise model, making a direct application of PEC infeasible.

In [VDBMKT23, KEA$^+$23], this issue is circumvented by introducing the *symmetry assumption* that forces sets of unlearnable Pauli fidelities to be equal. As a concrete example, when $\mathcal{G} = \text{CNOT}$ we know for $\Lambda_{\mathcal{G}}$, both λ_{XI} and λ_{XX} are individually unlearnable, while their product $\lambda_{XI}\lambda_{XX}$ is learnable. The *symmetry assumption* introduces the restriction $\lambda_{XI} = \lambda_{XX}$ to determine both eigenvalues, similarly for other eigenvalues. To my knowledge, there has not been a good justification for why such assumption is reasonable. In some cases, such assumptions can even result in equations contradicting each other.

In fact, such assumptions are not necessary to correctly perform PEC. Using the gate set Pauli noise learning framework, one can learn a collection of noise parameters, denoted by $\mathbf{\Lambda}_g$, that is gauge-equivalent to the true noise parameters $\mathbf{\Lambda}$, in the sense that no experiments can distinguish $\mathbf{\Lambda}$ from $\mathbf{\Lambda}_g$. Then, if we just pretend $\mathbf{\Lambda}_g$ is the true noise parameters and conduct PEC, we will get the correct prediction.

Formally, consider the following Gauge-consistent PEC (GC-PEC) protocol. One first learn a set of noise parameters $\mathbf{\Lambda}_g$ that are gauge-equivalent to the true values $\mathbf{\Lambda}$ (assuming the learning is exact for now). Use the superscript g to denote the learned noise channels. Construct our estimator using the following steps:

1. Randomly sample $a_0 \sim \{q_{a_0}^{g,S}\}$, $a_j \sim \{q_{a_j}^{g,\mathcal{G}_j}\}_{j=1}^{T}$, $a_{T+1} \sim \{q_{a_{T+1}}^{g,M}\}$.

2. Implement and measure the following expectation value

$$E_{\boldsymbol{a}} = \langle\langle \widetilde{O} | \mathcal{P}_{a_{T+1}} \mathcal{P}_{a_T} \widetilde{\mathcal{G}}_T \mathcal{U}_T \cdots \mathcal{P}_{a_1} \widetilde{\mathcal{G}}_1 \mathcal{U}_1 \mathcal{P}_{a_0} | \widetilde{\rho}_0 \rangle\rangle \tag{6.8}$$

where $\mathcal{P}_a(\rho) = P_a \rho P_a$.

3. Define the GC-PEC estimator as

$$\hat{o}^{(\mathrm{GC-PEC})} = E_{\boldsymbol{a}} \cdot \prod_{j=0}^{T+1} \gamma_j^g \mathrm{sgn}_{a_j}^g. \tag{6.9}$$

where γ_j^g and $\mathrm{sgn}_{a_j}^g$ are with respect to the jth learned Pauli noise channel (instead of the true noise channel).

The following proposition shows the correctness of GC-PEC. The proof steps are illustrated in Fig. 6.1

Proposition 6.2. *Given that one exactly knows a $\boldsymbol{\Lambda}_g$ that is gauge-equivalent to $\boldsymbol{\Lambda}$, the GC-PEC estimator with respect to $\boldsymbol{\Lambda}_g$ is an unbiased estimator for o.*

Proof. One just need to notice that $E_{\boldsymbol{a}}$ can be re-expressed as

$$\begin{aligned} E_{\boldsymbol{a}} &= \langle\langle O | \Lambda_M \mathcal{P}_{a_{T+1}} \mathcal{P}_{a_T} \Lambda_{\mathcal{G}_T} \mathcal{G}_T \mathcal{U}_T \cdots \mathcal{P}_{a_1} \Lambda_{\mathcal{G}_1} \mathcal{G}_1 \mathcal{U}_1 \mathcal{P}_{a_0} \Lambda_S | \rho_0 \rangle\rangle \\ &= \langle\langle O | \Lambda_M^g \mathcal{P}_{a_{T+1}} \mathcal{P}_{a_T} \Lambda_{\mathcal{G}_T}^g \mathcal{G}_T \mathcal{U}_T \cdots \mathcal{P}_{a_1} \Lambda_{\mathcal{G}_1}^g \mathcal{G}_1 \mathcal{U}_1 \mathcal{P}_{a_0} \Lambda_S^g | \rho_0 \rangle\rangle \end{aligned} \tag{6.10}$$

Because $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Lambda}$ cannot be distinguished by any experiments by assumptions. Then, following the same step as the proof for Proposition 6.1, one can show that

$$\mathbb{E}\hat{o}^{(\mathrm{GC-PEC})} = \langle\langle O | \mathcal{G}_T \mathcal{U}_T \cdots \mathcal{G}_1 \mathcal{U}_1 | \rho_0 \rangle\rangle = o. \tag{6.11}$$

□

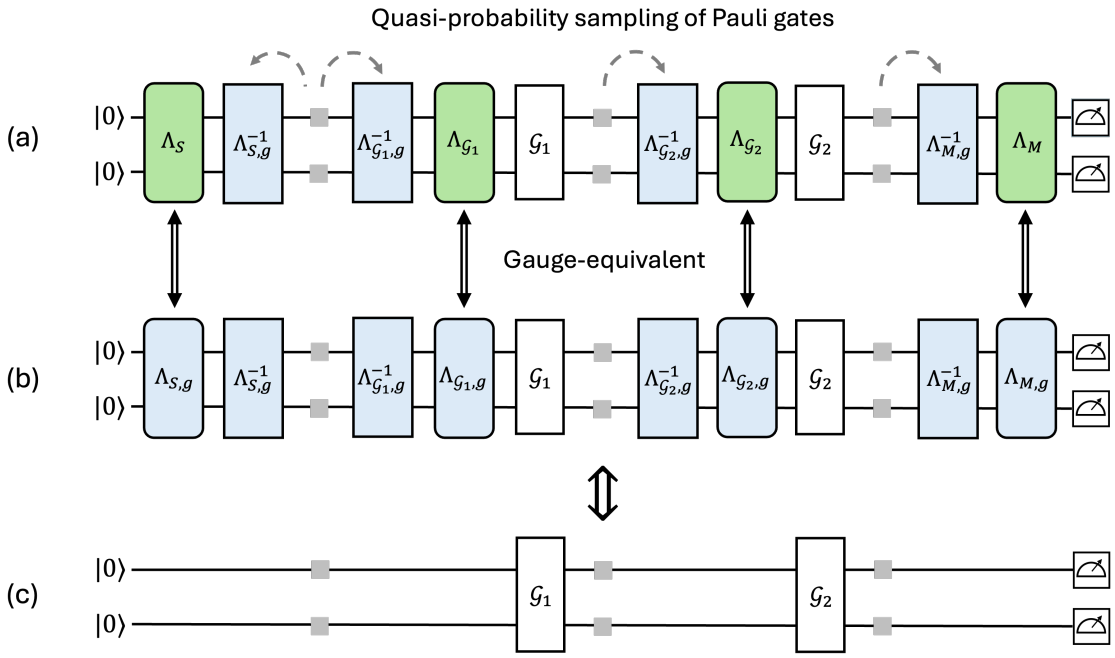


Figure 6.1: Illustration of Gauge-consistent PEC. (a) One use quasi-probability sampling to implement the inverse channels according to the learned gauge-equivalent noise model $\mathbf{\Lambda}_g$. (b) Replacing the true noise models with a gauge-equivalent model have no observable effects on the measurement statistics. This step is only a thought experiment. (c) The noise channels and inverse channels cancel out, returning noise-free results.

The sampling overhead of GC-QEM is determined by $\prod_j \gamma_j^{g,2}$, which, interestingly, depends on the choice of gauge parameters. In practice, one can optimize over all gauge-equivalent noise models that minimize this sampling overhead, we will not delve into the details in this thesis. We also remark that GC-QEM works for Pauli noise model with ansatzes, in which case our framework from Chapter 4 is crucial to determine the learnable and gauge degrees of freedoms.

6.3 Experiments

We experimentally demonstrate a proof-of-principle version of Gauge-consistent PEC. Concretely, we focus solely on Clifford circuits with Pauli observables. The noisy expectation value will take a monomial form $\lambda_{a_0}^{S}\lambda_{a_1}^{\mathcal{G}_1}\cdots\lambda_{a_T}^{\mathcal{G}_T}\lambda_{T+1}^{M}\cdot o$ where o is the ideal expectation value. Instead of doing the quasi-probability sampling, we can just use the learned Pauli eigenvalue to invert the noise factors. Note that doing quasi-probability sampling will give us the same results in expectation. We contrast our protocol with [VDBMKT23, KEA$^+$23], where they ignore state preparation noise to characterize and mitigate measurement noise and applying symmetry assumptions to identify gate noise. We refer to the latter method as Sym-PEC.

We will highlight one experiment in the thesis, leaving more experiments and details in [CCF$^+$25]. The experiment aims at preparing an n-qubit GHZ state for n from 2 to 21, and then measure the expectation value of $X^{\otimes n}$, whose ideal value should be one. We use both GC-PEC and Sym-PEC to mitigate the observable. As reported in Fig. 6.2, GC-PEC successfully mitigate the error and retrieve ideal value within error bars, while Sym-PEC suffers from a huge bias.

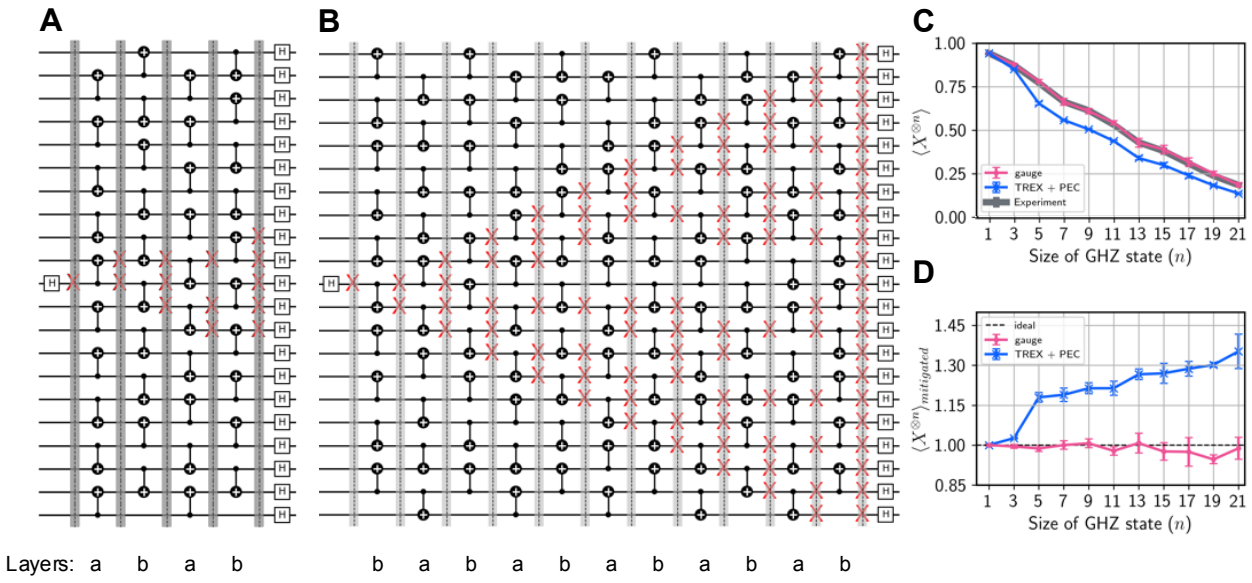


Figure 6.2: **Comparing GC-PEC with Sym-PEC in GHZ state preparation circuits.** (Figures courtesy of Alireza Seif and Edward Chen). (A)(B) shows our GHZ preparation circuits. An X measurements on all qubits is conducted at the end of the circuit. Note that both uses interleaving layers of 2 types of parallel 2-qubit gates. We have conduct separate learning experiments to learn the noise of parallel 2-qubit gates. (C) shows the experimental output vs. the predicted expectation value from gauge-consistent (pink circles) and symmetric (blue crosses) noise models. (D) Error mitigated results using GC-PEC (pink) and Sym-PEC(blue). Error bars are estimated from 6 separate experiments. Clearly GC-PEC obtain true values within error bars while Sym-PEC has a significant bias.

CHAPTER 7

OUTLOOK

In this the thesis, we have explored several theoretic and practical results of Pauli channels. Concretely, we have been discussing quantum learning advantages, quantum noise characterization, and quantum error mitigation. I hope this thesis establish Pauli channels as an invaluable tools for exploring modern quantum information science.

Besides the discussion provided in some of the chapters, I will provide some high-level outlook on the future research directions related to Pauli channel learning and how it can contribute to the development of quantum information science.

Logical noise characterization for enhancing quantum error correction. As we step into the era of quantum fault-tolerance, it is important to develop methods for characterizing noise on quantum devices with limited error correction capability. Understand the nature of noise on such device could provide invaluable insight on how to design better error correction architecture and decoding algorithms, thus expedite the quest towards full-fledged fault-tolerance. Due to the additional complexity added by error correction, it is unclear how much of the existing literature on Pauli noise learning or other physical-level quantum noise characterization toolkit can be applied there. Fortunately, there have been pioneering works going along this direction [WKBK23, ZCLJ25, HP25]. Other questions worth investigating are how to use the information of logical noise to enhance error correction performance [CYK$^+$22], how do you design error correction architecture tailored to the practical noise [AP08], etc.

Novel quantum applications with fault-tolerance. As quantum fault-tolerance gives us more qubits and less error, it is timely to explore novel quantum applications along the quest. Apart from the heavily studied quantum applications in cryptography, chemical

simulation, and other computational problem, I think an exciting but under-explored area is quantum metrology and quantum learning. Previous research (e.g., [ZZPJ18]) has laid down some theoretic background on error-corrected quantum metrology, but large are still unknown on the potential of how quantum error correction can enhance quantum learning more generally. As an example, the protocols described in Chapter 5 in this thesis [SCM+24] shows error mitigation can help de-bias the estimation. Can we use error correction instead to gain better performance for the noise learning task?

Interplay between quantum learning and quantum metrology. Recent years have seen exciting advancement in quantum learning theory. Quantum metrology, a long-standing research area with important real-world applications (such as LIGO and atomic clock), shares similar motivations and problems with quantum learning. It is interesting to better understand the connection between these two. In terms of Pauli channel learning, an interesting direction is to generalize the results to continuous-variable systems which are potentially more natural for certain metrological tasks. Some results have been established along this direction [OCW+24, LBØ+25]. Other example can be to use learning-theoretic methods in multi-parameter and time-dependent quantum metrology problem.

APPENDIX A

APPENDICES TO CHAPTER 3

A.1 Proof of Theorem 3.1

In this section, we provide details in the derivation of Eq. (3.5), *i.e.*, the distribution of measurement outcomes at Line 5 in Algorithm 1. For clarity, denote the k-qubit Hilbert space of the ancilla as A, and divide the n-qubit Hilbert space of the main system into a k-qubit subspace B and an $(n-k)$-qubit subspace C. The input state to the channel is $|\Psi_0\rangle_{AB} \otimes |\phi_0^{\mathsf{S}}\rangle_C$, and the measurement basis is $\{|\Psi_v\rangle_{AB} \otimes |\phi_e^{\mathsf{S}}\rangle_C\}_{v,e}$. Here, $|\Psi_v\rangle_{AB}$ are Bell states and can be expressed as

$$\begin{aligned} |\Psi_v\rangle\langle\Psi_v| &\coloneqq (P_v \otimes I)|\Psi^+\rangle\langle\Psi^+|(P_v \otimes I) \\ &= \frac{1}{4^k} \sum_{u \in \mathbb{Z}_2^{2k}} P_v P_u P_v \otimes P_u^T \\ &= \frac{1}{4^k} \sum_{u \in \mathbb{Z}_2^{2k}} (-1)^{\langle u,v \rangle} P_u \otimes P_u^T, \end{aligned} \tag{A.1}$$

And the stabilizer state $|\phi_e^{\mathsf{S}}\rangle_C$ is defined as

$$|\phi_e^{\mathsf{S}}\rangle\langle\phi_e^{\mathsf{S}}| \coloneqq \frac{1}{2^{n-k}} \sum_{s \in \mathsf{S}} (-1)^{\langle s,e \rangle} P_s, \tag{A.2}$$

for $e \in \mathsf{S}^\perp \coloneqq \mathbb{Z}_2^{2(n-k)}/\mathsf{S}$. We remark that, the stabilizer state is well-defined for all $e \in \mathbb{Z}_2^{2(n-k)}$, but $|\phi_a^{\mathsf{S}}\rangle$ and $|\phi_b^{\mathsf{S}}\rangle$ represent the same state if $a + b \in \mathsf{S}$ (bitwise modulo 2 sum), so we only need to consider the quotient space of $\mathbb{Z}_2^{2(n-k)}$ over S. When calculating the sympletic inner product $\langle s, e \rangle$, one should understand e as an arbitrary representative of the coset it stands for.

Therefore, the measurement outcome distribution can be calculated as

$$
\begin{aligned}
&p(v,e)\\
=&\operatorname{Tr}\left((|\Psi_v\rangle\langle\Psi_v|_{AB}\otimes|\phi_e^{\mathsf{S}}\rangle\langle\phi_e^{\mathsf{S}}|_C)\mathbb{1}^A\otimes\Lambda^{BC}(|\Psi_0\rangle\langle\Psi_0|_{AB}\otimes|\phi_0^{\mathsf{S}}\rangle\langle\phi_0^{\mathsf{S}}|_C)\right)\\
=&\frac{1}{4^{n+k}}\operatorname{Tr}\left(\sum_{u,u'\in\mathbb{Z}_2^{2k}}\sum_{s,s'\in\mathsf{S}}\lambda_{u'\oplus s'}\left((-1)^{\langle u,v\rangle}P_u\otimes P_u^T\otimes(-1)^{\langle s,e\rangle}P_s\right)_{ABC}\left(P_{u'}\otimes P_{u'}^T\otimes P_{s'}\right)_{ABC}\right)\\
=&\frac{1}{2^{n+k}}\sum_{u\in\mathbb{Z}_2^{2k}}\sum_{s\in\mathsf{S}}\lambda_{u\oplus s}(-1)^{\langle u,v\rangle}(-1)^{\langle s,e\rangle},
\end{aligned}
\tag{A.3}
$$

which is exactly Eq. (3.5) in the main text. It is then obvious that $p(v,e)$ and $\lambda_{u\oplus s}$ are related by the Walsh-Hadamard transform (see [FW20, Lemma 4]).

A.2 Proof of Theorem 3.2

A tight lower bound for non-adaptive and non-concatenating strategies In this section, we prove a matching lower bound of $\Omega(n2^{n-k})$ for all non-adaptive and non-concatenating k-qubit ancilla-assisted Pauli channel estimation protocols, which include the ancilla-free strategies ($k=0$) and n-qubit ancilla assisted strategies ($k=n$) as two special cases. Recall that, by a non-adaptive and non-concatenating protocol we mean that, for each sample of the Pauli channel Λ, we prepare an $n+k$ qubits state, input it to $\Lambda\otimes\mathbb{1}$, and apply a POVM measurement on the joint output state. The input state and measurement setting for the ith sample is not allowed to depend on previous measurement outcomes, nor do we allow concatenating multiple samples of Λ in a single round of measurement (which is used in RB-type Pauli error estimation protocols [FW20]).

Theorem 1.1. *For any non-adaptive, non-concatenating k-qubit ancilla-assisted protocols that give an estimate $\widehat{\boldsymbol{\lambda}}$ of the Pauli eigenvalues $\boldsymbol{\lambda}$ of an arbitrary unknown n-qubit Pauli channel such that*

$$
|\widehat{\lambda}_a-\lambda_a|<\frac{1}{2},\quad\forall a\in\mathbb{Z}_2^{2n}
\tag{A.4}
$$

holds with high probability, the number of samples of Λ *required is at least* $\Omega(n2^{n-k})$.

Our proof generalizes the information-theoretical arguments by Huang *et al.* [HKP21], which in turn stem from previous work on sample complexity lower bounds for quantum tomography [FGLE12, HHJ$^+$17, RKK$^+$18]. Consider a communication protocol between Alice and Bob where their share the following "codebook":

$$(a,s) \in \{1,\cdots,4^n-1\}\times\{\pm 1\} \;\longrightarrow\; \Lambda_{(a,s)}(\cdot) = \frac{1}{2^n}\left(I\,\mathrm{Tr}(\cdot) + sP_a\,\mathrm{Tr}(P_a(\cdot))\right). \tag{A.5}$$

Now, Alice picks at uniform random one out of the $2(4^n-1)$ possible pairs of (a,s) and then send N copies of the channel $\Lambda_{(a,s)}$ to Bob. If there exists a Pauli channel estimation protocol using N samples and satisfying the assumption of Theorem 1.1, Bob can use that protocol to uniquely determine Alice's choice of (a,s) with high probability, since the Pauli eigenvalues of any $\Lambda_{(a,s)}$ only take values from $\{-1,0,+1\}$. Suppose Bob's input state and POVM outcome for the ith sample is $\{\rho_i, E_i\}$. According to Fano's inequality, the mutual information between the random variable pair (a,s) and Bob's measurement results has the following lower bound

$$I\left((a,s):\{\rho_1,E_1\},...,\{\rho_N,E_N\}\right) \geq \Omega(\log(2(4^n-1))) = \Omega(n). \tag{A.6}$$

We also know by assumption that the measurement outcomes $\{\rho_i, E_i\}$ are independent from each other, conditioned on (a,s). The chain rule of mutual information then gives that

$$\sum_{i=1}^{N} I\left((a,s),\{\rho_i,E_i\}\right) = I((a,s):\{\rho_1,E_1\},...,\{\rho_N,E_N\}) \geq \Omega(n). \tag{A.7}$$

We will show that $I\left((a,s):\{\rho_i,E_i\}\right) \leq \mathcal{O}(2^{k-n})$ in the following lemma. This would then give us the desired sample complexity lower bound $N \geq \Omega(n2^{n-k})$, which completes the proof of Theorem 1.1.

Lemma 1.1. $I\left((a,s):\{\rho_i,E_i\}\right)\leq\dfrac{2^k}{2^n-1}$.

Proof. First notice that it suffices to consider pure state input and rank-1 POVM measurements. The latter comes from the fact that every POVM measurement can be viewed as a coarse-graining of some rank-1 POVM measurement. To see the former, consider the underlying distribution $p((a,s),\{\rho_i,E_i\})=p(a,s)p(\{E_i,\rho_i\}|a,s)$. The mutual information $I((a,s):\{\rho_i,E_i\})$ is convex about $p(\{E_i,\rho_i\}|a,s)$ when fixing $p(a,s)$ (see *e.g.*, [CT12, Theorem 2.7.4]), thus using mixed state input can never provide a larger mutual information.

Thanks to this observation, we can without loss of generality let the input state be $|A\rangle$ and let the POVM measurement be $\{w_j 2^{n+k}|B_j\rangle\langle B_j|\}_j$, where $|A\rangle\,,|B_j\rangle\in\mathbb{C}^{2^n\times 2^k}$ are unit vectors, and $\sum_j w_j=1$ by normalization. We also abuse notations a little bit to let A and B_j denote the $2^n\times 2^k$ matrices that satisfy

$$|A\rangle=\sum_{p=0}^{2^n-1}\sum_{q=0}^{2^k-1}\langle p|A|q\rangle\,|p\rangle\,|q\rangle\,,\quad |B_j\rangle=\sum_{p=0}^{2^n-1}\sum_{q=0}^{2^k-1}\langle p|B_j|q\rangle\,|p\rangle\,|q\rangle\,, \tag{A.8}$$

where $\{|p\rangle\}$ and $\{|q\rangle\}$ are computational basis states. The normalization condition of $|A\rangle$ and $|B_j\rangle$ is equivalent to

$$\mathrm{Tr}(A^\dagger A)=\mathrm{Tr}(B_j^\dagger B_j)=1. \tag{A.9}$$

We also define $C_j\coloneqq B_jA^\dagger$ which is a $2^n\times 2^n$ matrix of rank less or equal to 2^k.

The mutual information between (a,s) and a single round of measurement outcome j can be upper bounded as

$$\begin{aligned}
I((a,s):j)&=H(j)-H(j|a,s)\\
&=-\sum_j\left(\mathop{\mathbb{E}}_{(a,s)}p(j|a,s)\right)\log\left(\mathop{\mathbb{E}}_{(a,s)}p(j|a,s)\right)+\mathop{\mathbb{E}}_{(a,s)}\sum_j p(j|a,s)\log p(j|a,s)\\
&\leq\sum_j\frac{\mathbb{E}_{(a,s)}[p(j|a,s)^2]-\mathbb{E}_{(a,s)}[p(j|a,s)]^2}{\mathbb{E}_{(a,s)}[p(j|a,s)]},
\end{aligned} \tag{A.10}$$

where the inequality follows from the fact that $\log(x) \leq \log(y) + \frac{x-y}{y}$ in which we take $x \coloneqq p(j|a,s)$ and $y \coloneqq \mathbb{E}_{(a,s)}[p(j|a,s)]$.

The conditional probability $p(j|a,s)$ can be calculated as

$$\begin{aligned}
p(j|a,s) &= w_j 2^{n+k} \langle B_j | \Lambda_{(a,s)} \otimes \mathbb{1}(|A\rangle\langle A|) | B_j \rangle \\
&= w_j \sum_{b=0}^{4^k} \Big(\langle B_j | I \otimes P_b | B_j \rangle \langle A | I \otimes P_b | A \rangle + s \langle B_j | P_a \otimes P_b | B_j \rangle \langle A | P_a \otimes P_b | A \rangle \Big) \\
&= w_j \sum_{b=0}^{4^k} \Big(\operatorname{Tr}(B_j^\dagger B_j P_b^T) \operatorname{Tr}(A^\dagger A P_b^T) + s \operatorname{Tr}(B_j^\dagger P_a B_j P_b^T) \operatorname{Tr}(A^\dagger P_a A P_b^T) \Big) \\
&= w_j 2^k \Big(\operatorname{Tr}(B_j^\dagger B_j A^\dagger A) + s \operatorname{Tr}(B_j^\dagger P_a B_j A^\dagger P_a A) \Big) \\
&= w_j 2^k \Big(\operatorname{Tr}(C_j^\dagger C_j) + s \operatorname{Tr}(P_a C_j P_a C_j^\dagger) \Big),
\end{aligned} \tag{A.11}$$

where we expand the identity channel as $\mathbb{1}(\cdot) = 2^{-k} \sum_{b=0}^{4^k} P_b \operatorname{Tr}(P_b(\cdot))$ in the second line, and use the fact $2^{-k} \sum_{b=0}^{4^k} P_b \otimes P_b$ equals to the swap operator in the fourth line.

The average value and second moment of $p(j|a,s)$ according to the distribution of (a,s) are

$$\begin{aligned}
\mathop{\mathbb{E}}_{(a,s)} [p(j|a,s)] &= w_j 2^k \operatorname{Tr}(C_j^\dagger C_j), \\
\mathop{\mathbb{E}}_{(a,s)} [p(j|a,s)^2] &= w_j^2 4^k \left(\operatorname{Tr}^2(C_j^\dagger C_j) + \frac{1}{4^n - 1} \sum_{a=1}^{4^n - 1} \operatorname{Tr}^2(P_a C_j P_a C_j^\dagger) \right).
\end{aligned} \tag{A.12}$$

Hence we have the following bound for the mutual information

$$I\left((a,s) : j\right) \leq \sum_j w_j 2^k \operatorname{Tr}(C_j^\dagger C_j) \left(\frac{1}{4^n - 1} \sum_{a=1}^{4^n - 1} \frac{\operatorname{Tr}(P_a C_j^\dagger P_a C_j)^2}{\operatorname{Tr}(C_j^\dagger C_j)^2} \right). \tag{A.13}$$

Now we further calculate the R.H.S. of the above inequality. Let $M = P_a C_j^\dagger P_a C_j$. Notice that

$$\operatorname{rank}(M) \leq \operatorname{rank}(C_j) \leq 2^k, \tag{A.14}$$

which means there exists a rank-2^k projector Π such that $\mathrm{Tr}(M) = \mathrm{Tr}(M\Pi)$. According to Cauchy-Schwarz inequality,

$$\begin{aligned}\mathrm{Tr}(M)^2 &= \mathrm{Tr}(M\Pi)^2 \\ &\leq \mathrm{Tr}(MM^\dagger)\,\mathrm{Tr}(\Pi^\dagger\Pi) \\ &= \mathrm{Tr}(C_jC_j^\dagger P_aC_jC_j^\dagger P_a)\times 2^k\end{aligned} \tag{A.15}$$

Substitute this into Eq. (A.13),

$$\begin{aligned}I((a,s):j) &\leq \sum_j w_j 2^k\,\mathrm{Tr}(C_j^\dagger C_j)\left(\frac{2^k}{4^n-1}\sum_{a=1}^{4^n-1}\frac{\mathrm{Tr}(C_jC_j^\dagger P_aC_jC_j^\dagger P_a)}{\mathrm{Tr}(C_j^\dagger C_j)^2}\right) \\ &= \sum_j w_j 2^k\,\mathrm{Tr}(C_j^\dagger C_j)\left(\frac{2^k}{4^n-1}\times\frac{\sum_{a=0}^{4^n-1}\mathrm{Tr}(C_jC_j^\dagger P_aC_jC_j^\dagger P_a)-\mathrm{Tr}(C_jC_j^\dagger C_jC_j^\dagger)}{\mathrm{Tr}(C_j^\dagger C_j)^2}\right) \\ &= \sum_j w_j 2^k\,\mathrm{Tr}(C_j^\dagger C_j)\left(\frac{2^k}{4^n-1}\times\frac{2^n\,\mathrm{Tr}(C_j^\dagger C_j)^2-\mathrm{Tr}(C_jC_j^\dagger C_jC_j^\dagger)}{\mathrm{Tr}(C_j^\dagger C_j)^2}\right) \\ &\leq \sum_j w_j 2^k\,\mathrm{Tr}(C_j^\dagger C_j)\frac{2^k}{2^n-1} \\ &= \frac{2^k}{2^n-1},\end{aligned} \tag{A.16}$$

where the third line uses the following formula of Pauli twirling,

$$\frac{1}{4^n}\sum_{a=0}^{4^n}P_aXP_a = \frac{1}{2^n}\,\mathrm{Tr}(X)I, \tag{A.17}$$

and the last line follows from the fact that

$$\sum_j w_j 2^k\,\mathrm{Tr}(C_j^\dagger C_j) = \sum_j \mathop{\mathbb{E}}_{(a,s)}[p(j|a,s)] = 1. \tag{A.18}$$

This completes the proof of Lemma 1.1. □

A lower bound for adaptive but non-concatenating strategies In this section, we prove a lower bound of $\Omega(2^{(n-k)/3})$ for all adaptive and non-concatenating k-qubit ancilla-assisted Pauli channel estimation protocols, as stated in the following theorem. This bound is improved to be tight in subsequent works which will be presented in Sec.[[]]. Here I decide to include this loose bound because it uses a different proof strategies that might be of independent interest.

Theorem 1.2. *For any adaptive, non-concatenating k-qubit ancilla-assisted protocols that give an estimate $\widehat{\boldsymbol{\lambda}}$ of the Pauli eigenvalues $\boldsymbol{\lambda}$ of an arbitrary unknown n-qubit Pauli channel such that*

$$|\widehat{\lambda}_a - \lambda_a| < \frac{1}{2}, \quad \forall a \in \mathbb{Z}_2^{2n} \tag{A.19}$$

holds with high probability, the number of samples of Λ required is at least $\Omega(2^{(n-k)/3})$.

Our proof techniques generalize the methods of Huang *et al.* [HKP21] for proving adaptive sample complexity lower bounds for Pauli expectation values estimation of unknown quantum states. Consider the following 4^n possible Pauli channels

$$\begin{cases} \Lambda_{\text{dep}}(\cdot) = \dfrac{1}{2^n} I \operatorname{Tr}(\cdot), \\ \Lambda_a(\cdot) = \dfrac{1}{2^n} \left(I \operatorname{Tr}(\cdot) + P_a \operatorname{Tr}(P_a(\cdot)) \right), \ \forall a \in \{1, \cdots, 4^n - 1\}. \end{cases} \tag{A.20}$$

Here Λ_{dep} is known as the completely deplorizing channel. If there exists an Pauli channel estimation protocol satisfying the requirement of Theorem 1.2, one can unambiguously identify each one of the 4^n possible Pauli channels appearing above with high probability, given sufficient number of samples of Λ. This in turn implies that one should be able to distinguish the following two equal-probable hypotheses with high success probability.

1. Given N copies of $\Lambda = \Lambda_{\text{dep}}$.

2. Given N copies of $\Lambda = \Lambda_a$ for a uniformly-randomly picked $a \in \{1, \cdots, 4^n - 1\}$.

For an adaptive but non-concatenating protocol, one has to choose an $2^n \times 2^k$ dimensional input state and a POVM measurement for the ith sample of Λ, where the choice may depend on previous measurement outcomes. Denote the measurement outcome of the ith round as o_i. We explicitly write the state and measurement at the ith round as $\rho^{o<i}$ and $\{E_j^{o<i}\}_j$ to emphasize their dependence on $o_{<i} \coloneqq [o_1, ..., o_{i-1}]$.Denote the measurement outcomes among all the N samples as $o_{1:N} \coloneqq [o_1, \cdots, o_N]$. The probability distribution of $o_{1:N}$ under the above two hypothesis can be expressed as

$$\begin{cases} \text{Hypothesis 1:} \quad p_1(o_{1:N}) = \prod_{i=1}^{N} \mathrm{Tr}\left(E_{o_i}^{o<i} \Lambda_{\mathrm{dep}} \otimes \mathbb{1}(\rho^{o<i})\right), \\ \text{Hypothesis 2:} \quad p_2(o_{1:N}) = \mathop{\mathbb{E}}_{a \neq 0} \prod_{i=1}^{N} \mathrm{Tr}\left(E_{o_i}^{o<i} \Lambda_a \otimes \mathbb{1}(\rho^{o<i})\right). \end{cases} \tag{A.21}$$

The ability to distinguish these two hypotheses is equivalent to the ability to distinguish p_1 from p_2. The maximal success probability of distinguishing two probability distributions is given by $\frac{1}{2}(1+\mathrm{TV}(p_1, p_2))$ where TV stands for the total variance distance defined as follows

$$\mathrm{TV}(p_1, p_2) \coloneqq \sum_{\substack{o_{1:N} \text{ s.t.} \\ p_1(o_{1:N}) \geq p_2(o_{1:N})}} (p_1(o_{1:N}) - p_2(o_{1:N})). \tag{A.22}$$

We will show in the following Lemma that $\mathrm{TV}(p_1, p_2) = \mathcal{O}(N 2^{(k-n)/3})$, which immediately implies that one must have $N = \Omega(2^{(n-k)/3})$ in order to obtain high success probability in distinguishing p_1 from p_2. This then completes the proof of Theorem 1.2.

Lemma 1.2. $\mathrm{TV}(p_1, p_2) \leq 2N \left(\dfrac{2^k}{2^n - 1}\right)^{1/3}$.

Proof. First notice that it suffices to consider pure state input and rank-1 POVM measurement. For the former, the probability distribution obtained from mixed state input can be viewed as a convex combination of distributions obtained from pure state input. Thanks to the joint convexity, mixed state input can not yield a larger total variance distance; For the

latter, every POVM measurement can be viewed as a coarse-graining of some rank-1 POVM measurement. Because of the data-processing property, this coarse-graining would not yield a larger total variance distance.

In light of this observation, we can without loss of generality let the input state at the ith round be $|A^{o_{<i}}\rangle$ and let the POVM measurement be $\{w_{o_i}^{o_{<i}} 2^{n+k} |B_{o_i}^{o_{<i}}\rangle\langle B_{o_i}^{o_{<i}}|\}_{o_i}$, conditioned on previous measurement outcomes $o_{<i}$, where $|A^{o_{<i}}\rangle, |B_{o_i}^{o_{<i}}\rangle \in \mathbb{C}^{2^n \times 2^k}$ are unit vectors, and $\sum_{o_i} w_{o_i}^{o_{<i}} = 1$ by normalization. We also introduce the two $2^n \times 2^k$ matrices $A^{o_{<i}}$, $B_{o_i}^{o_{<i}}$ defined similarly as in Eq. (A.8), and define $C_{o_i}^{o_{<i}} \coloneqq B_{o_i}^{o_{<i}} A^{o_{<i}\dagger}$ which is a $2^n \times 2^n$ matrix of rank less or equal to 2^k. With the above definitions, one can verify that p_1 and p_2 can be expressed as follows

$$\begin{aligned}
p_1(o_{1:N}) &= \prod_{i=1}^{N} w_{o_i}^{o_{<i}} 2^{n+k} \langle B_{o_i}^{o_{<i}}| \Lambda_{\text{dep}} \otimes \mathbb{1} \left((|A^{o_{<i}}\rangle\langle A^{o_{<i}}|) \right) |B_{o_i}^{o_{<i}}\rangle , \\
&= \prod_{i=1}^{N} w_{o_i}^{o_{<i}} 2^{k} \operatorname{Tr}\left(C_{o_i}^{o_{<i}\dagger} C_{o_i}^{o_{<i}} \right) \\
p_2(o_{1:N}) &= \mathop{\mathbb{E}}_{a \neq 0} \prod_{i=1}^{N} w_{o_i}^{o_{<i}} 2^{n+k} \langle B_{o_i}^{o_{<i}}| \Lambda_a \otimes \mathbb{1} \left((|A^{o_{<i}}\rangle\langle A^{o_{<i}}|) \right) |B_{o_i}^{o_{<i}}\rangle \\
&= \mathop{\mathbb{E}}_{a \neq 0} \prod_{i=1}^{N} w_{o_i}^{o_{<i}} 2^{k} \left(\operatorname{Tr}\left(C_{o_i}^{o_{<i}\dagger} C_{o_i}^{o_{<i}} \right) + \operatorname{Tr}\left(C_{o_i}^{o_{<i}\dagger} P_a C_{o_i}^{o_{<i}} P_a \right) \right) .
\end{aligned} \tag{A.23}$$

The total variance between p_1 and p_2 can then be bounded as

$$\begin{aligned}
&\operatorname{TV}(p_1, p_2) \\
&= \mathop{\mathbb{E}}_{a \neq 0} \sum_{\substack{o_{1:N} \text{ s.t.} \\ p_1(o_{1:N}) \geq p_2(o_{1:N})}} \left(\prod_{i=1}^{N} w_{o_i}^{o_{<i}} 2^{k} \operatorname{Tr}\left(C_{o_i}^{o_{<i}\dagger} C_{o_i}^{o_{<i}} \right) \right) \left(1 - \prod_{i=1}^{N} \left(1 + \frac{\operatorname{Tr}\left(C_{o_i}^{o_{<i}\dagger} P_a C_{o_i}^{o_{<i}} P_a \right)}{\operatorname{Tr}\left(C_{o_i}^{o_{<i}\dagger} C_{o_i}^{o_{<i}} \right)} \right) \right) \\
&= \mathop{\mathbb{E}}_{a \neq 0} \sum_{\substack{o_{1:N} \text{ s.t.} \\ p_1(o_{1:N}) \geq p_2(o_{1:N})}} p_1(o_{1:N}) \left(1 - \prod_{i=1}^{N} \left(1 + \frac{\operatorname{Tr}\left(C_{o_i}^{o_{<i}\dagger} P_a C_{o_i}^{o_{<i}} P_a \right)}{\operatorname{Tr}\left(C_{o_i}^{o_{<i}\dagger} C_{o_i}^{o_{<i}} \right)} \right) \right) ,
\end{aligned} \tag{A.24}$$

In order to bound the R.H.S., we make use of a technique from Huang *et al.* [HKP21]. Let C

denote an arbitrary $2^n \times 2^n$ complex matrix of rank no more than 2^k, consider the following subset of n-qubit Pauli operators

$$G \coloneqq \left\{ a \in \{1, \cdots, 4^n - 1\} \ : \ \left| \frac{\mathrm{Tr}\left(C^\dagger P_a C P_a\right)}{\mathrm{Tr}\left(C^\dagger C\right)} \right| \leq \left(\frac{2^k}{2^n - 1} \right)^{1/3} \right\}. \tag{A.25}$$

We claim that the size of G satisfies

$$|G| \geq \left(1 - \left(\frac{2^k}{2^n - 1} \right)^{1/3} \right) (4^n - 1), \tag{A.26}$$

which can be shown by contradiction: Suppose this does not hold, we would have

$$\begin{aligned} \sum_{a=1}^{4^n-1} \left(\frac{\mathrm{Tr}\left(C^\dagger P_a C P_a\right)}{\mathrm{Tr}\left(C^\dagger C\right)} \right)^2 &\geq \left(\frac{2^k}{2^n - 1} \right)^{2/3} \times (4^n - 1 - |G|) \\ &> \left(\frac{2^k}{2^n - 1} \right)^{2/3} \times \left(\frac{2^k}{2^n - 1} \right)^{1/3} (4^n - 1) \\ &= 2^k (2^n + 1). \end{aligned} \tag{A.27}$$

However, the L.H.S. of the above can be upper bounded as

$$\begin{aligned} \sum_{a=1}^{4^n-1} \left(\frac{\mathrm{Tr}\left(C^\dagger P_a C P_a\right)}{\mathrm{Tr}\left(C^\dagger C\right)} \right)^2 &\leq \sum_{a=1}^{4^n-1} \frac{2^k \, \mathrm{Tr}(C C^\dagger P_a C C^\dagger P_a)}{\mathrm{Tr}^2\left(C^\dagger C\right)} \\ &= 2^k \frac{2^n \, \mathrm{Tr}^2\left(C^\dagger C\right) - \mathrm{Tr}(C^\dagger C C^\dagger C)}{\mathrm{Tr}^2\left(C^\dagger C\right)} \\ &\leq 2^k 2^n, \end{aligned} \tag{A.28}$$

where the first inequality follows from Cauchy-Schwarz (see Eq. (A.15)), and the first equality evaluates the Pauli twirling (see Eq. (A.16)). This leads to contradition, and hence proves the desired lower bound on $|G|$.

Now, we define another subset of n-qubit Pauli operators, conditioned on the measure-

ment outcomes $o_{1:N}$, as follows

$$G^{(o_{1:N})} := \left\{ a \in \{1, \cdots, 4^n - 1\} : \left| \frac{\mathrm{Tr}\left(C_{o_i}^{o_{<i}\dagger} P_a C_{o_i}^{o_{<i}} P_a \right)}{\mathrm{Tr}\left(C_{o_i}^{o_{<i}\dagger} C_{o_i}^{o_{<i}} \right)} \right| \leq \left(\frac{2^k}{2^n - 1} \right)^{1/3}, \ \forall i = 1, \cdots, N \right\}. \tag{A.29}$$

By applying a "union bound" on Eq. (A.25), we immediately have the following lower bound on the size of $G^{(o_{1:N})}$:

$$\left| G^{(o_{1:N})} \right| \geq \left(1 - N \left(\frac{2^k}{2^n - 1} \right)^{1/3} \right) (4^n - 1). \tag{A.30}$$

We are now ready to upper bound $\mathrm{TV}(p_1, p_2)$ from Eq. (A.24). The strategy is to divide the sum over all Pauli operators into $G^{(o_{1:N})}$ and $\mathsf{P}^n \backslash G^{(o_{1:N})}$. All terms within the former group are small thanks to the definition of $G^{(o_{1:N})}$; Terms within the latter group could be large, but the total number of them are small. Combining these two gives a pretty good upper bound on $\mathrm{TV}(p_1, p_2)$. In math,

$$\begin{aligned}
&\mathrm{TV}(p_1, p_2) \\
=& \sum_{\substack{o_{1:N} \text{ s.t.} \\ p_1(o_{1:N}) \geq p_2(o_{1:N})}} \frac{p_1(o_{1:N})}{4^n - 1} \left(\sum_{a \in G^{(o_{1:N})}} + \sum_{\substack{a \in \mathbb{Z}_2^{2n} \backslash G^{(o_{1:N})}, \\ a \neq 0}} \right) \left(1 - \prod_{i=1}^{N} \left(1 + \frac{\mathrm{Tr}\left(C_{o_i}^{o_{<i}\dagger} P_a C_{o_i}^{o_{<i}} P_a \right)}{\mathrm{Tr}\left(C_{o_i}^{o_{<i}\dagger} C_{o_i}^{o_{<i}} \right)} \right) \right) \\
\leq& \sum_{\substack{o_{1:N} \text{ s.t.} \\ p_1(o_{1:N}) \geq p_2(o_{1:N})}} \frac{p_1(o_{1:N})}{4^n - 1} \left(\left| G^{(o_{1:N})} \right| \times \left(1 - \left(1 - \left(\frac{2^k}{2^n - 1} \right)^{1/3} \right)^N \right) + \left(4^n - 1 - \left| G^{(o_{1:N})} \right| \right) \right) \\
\leq& \sum_{\substack{o_{1:N} \text{ s.t.} \\ p_1(o_{1:N}) \geq p_2(o_{1:N})}} p_1(o_{1:N}) \left(N \left(\frac{2^k}{2^n - 1} \right)^{1/3} + N \left(\frac{2^k}{2^n - 1} \right)^{1/3} \right) \\
\leq& 2N \left(\frac{2^k}{2^n - 1} \right)^{1/3}.
\end{aligned} \tag{A.31}$$

The first inequality uses an additional fact that $\left| \mathrm{Tr}(C^\dagger P_a C P_a) \right| / \left| \mathrm{Tr}(C^\dagger C) \right| \leq 1$ to bound the second sum, which follows from the Cauchy-Schwarz inequality. The second inequality

uses the bounds $\left|G^{(o_{1:N})}\right| \leq 4^n - 1$ for the first sum and Eq. (A.30) for the second sum, as well as the fact that $1 - (1-x)^N \leq Nx$ for all $0 \leq x \leq 1$. (Note that $2^k/(2^n - 1) \leq 1$ only if $k \leq n - 1$, but our targeted upper bound trivially holds for $k = n$.) This completes the proof of Lemma 1.2.

□

A lower bound for the most general ancilla-free strategies. The lower bounds in the previous sections work for non-concatenating strategies, where one is not allowed to concatenate (or, coherently access) multiple copies of the unknwon channel before doing a single measurement. The concatenating strategies, as depicted in Fig. 3.1, are however a natural apparatus for many randomized benchmarking protocols, where one effectively concatenate a varying number of noise channels in order to measure a series of exponentially-decaying values, and then extract the parameters of interest via fitting the decay rate. The original purpose of such concatenation in these protocols is to eliminate the effect of state-preparation-and-measurement (SPAM) error. Here, we want to understand whether concatenating strategies also provide a sample complexity advantage. The short answer is no. We will present a sample complexity lower bound of $\Omega(2^{n/3})$ for the most powerful (adaptive, concatenating) ancilla-free ($k = 0$) Pauli channel estimation schemes. This result justifies our claim that ancillary systems are indeed indispensable to overcome the exponential barrier in sample complexity.

To start with, we give a rigorous definition of the most general ancilla-free strategies that we are going to study

Definition 1.3. *Let Λ be an unknwon n-qubit Pauli channel. An adaptive, concatenating, ancilla-free ($k = 0$) estimation protocol is specified by the following parameters. Let N denote the total rounds of measurements. For the ith round, let $\rho^{o<i}$ denote the input state, $\{E_{o_i}^{o<i}\}_{o_i}$ denote the POVM measurement, $M^{o<i}$ denote the length of concatenation, and $\{\mathcal{C}_k^{o<i}\}_{k=1}^{M^{o<i}}$ denote a set of processing channels. The ith measurement outcome is given by*

o_i with probability

$$\Pr(o_i|o_{<i}) = \mathrm{Tr}\left[E_{o_i}^{o_{<i}}\Lambda(\mathcal{C}_{M^{o_{<i}}-1}^{o_{<i}}(\cdots\mathcal{C}_2^{o_{<i}}(\Lambda(\mathcal{C}_1^{o_{<i}}(\Lambda(\rho^{o_{<i}}))))\cdots))\right]. \tag{A.32}$$

All the superscript $o_{<i} \coloneqq [o_1, \cdots, o_{i-1}]$ are used to emphasize the dependence on previous measurement outcomes. The protocol should produce an estimate of Λ via classical processing on the measurement outcomes $o_{1:N}$.

The problem we are interested in is still approximating the Pauli eigenvalues $\boldsymbol{\lambda}$ to small error in l_∞ distance. Note that, the parameter N in the above definition is not exactly the sample complexity but is the total number of measurements conducted. Since one is allowed to concatenate multiple copies of Λ in a single measurement, N is a lower bound for the sample complexity of Λ. Our result is summarized in the following theorem.

Theorem 1.3. *For any adaptive, concatenating, ancilla-free ($k = 0$) protocols that gives an estimate $\widehat{\boldsymbol{\lambda}}$ of the Pauli eigenvalues $\boldsymbol{\lambda}$ of an arbitrary unknown n-qubit Pauli channel Λ such that*

$$|\widehat{\lambda}_a - \lambda_a| < \frac{1}{2}, \quad \forall a \in \mathbb{Z}_2^{2n} \tag{A.33}$$

holds with high probability, the rounds of measurements N (and hence the number of samples of Λ) required is at least $\Omega(2^{n/3})$.

Proof. The proof methods are similar to the proof of Theorem 1.2. Define the following Pauli channels.

$$\begin{cases} \Lambda_{\mathrm{dep}}(\cdot) = \dfrac{1}{2^n} I\,\mathrm{Tr}(\cdot), \\ \Lambda_a(\cdot) = \dfrac{1}{2^n}\left(I\,\mathrm{Tr}(\cdot) + P_a\,\mathrm{Tr}(P_a(\cdot))\right),\ \forall a \in \{1, \cdots, 4^n - 1\}. \end{cases} \tag{A.34}$$

We consider the problem of distinguishing the following two equal-probable hypotheses

1. Given N copies of $\Lambda = \Lambda_{\mathrm{dep}}$.

2. Given N copies of $\Lambda = \Lambda_a$ for a uniformly-randomly picked $a \in \{1, \cdots, 4^n - 1\}$.

An estimation protocol satisfying our assumptions should be able to distinguish these two hypotheses with high probability. Let the probability distribution of the measurement outcomes $o_{1:N}$ under these two hypotheses be p_1 and p_2, respectively. The total variance distance $\mathrm{TV}(p_1, p_2)$ must be at least $\Omega(1)$ for the distinguishing task to succeed with high probability. We will show in the following that $\mathrm{TV}(p_1, p_2) = O(N2^{-n/3})$, which then gives the claimed lower bound of $N = \Omega(2^{n/3})$.

To start with, based on the same argument as in the proof of Lemma 1.2, it suffices to consider pure state input and rank-1 POVM measurements, so we replace $\rho^{o<i}$ and $\{E_{o_i}^{o<i}\}_{o_i}$ in Definition 1.3 with $|A^{o<i}\rangle\langle A^{o<i}|$ and $\{w_{o_i}^{o<i} 2^n |B^{o<i}\rangle\langle B^{o<i}|\}_{o_i}$ respectively, where $|A^{o<i}\rangle, |B_{o_i}^{o<i}\rangle \in \mathbb{C}^{2^n}$ are unit vectors, and $\sum_{o_i} w_{o_i}^{o<i} = 1$ by normalization.

Next, we calculate the distribution of $o_{1:N}$ under the two different hypotheses. The expression for p_1 can be easily obtained, as Λ_{dep} is simply the completely deplorizing channel. We have

$$\begin{aligned}
&p_1(o_{1:N}) \\
&= \prod_{i=1}^{N} w_{o_i}^{o<i} 2^n \langle B_{o_i}^{o<i}| \Lambda_{\mathrm{dep}}(\mathcal{C}_{M^{o<i}-1}^{o<i}(\cdots \mathcal{C}_2^{o<i}(\Lambda_{\mathrm{dep}}(\mathcal{C}_1^{o<i}(\Lambda_{\mathrm{dep}}(|A^{o<i}\rangle\langle A^{o<i}|)))) \cdots)) |B_{o_i}^{o<i}\rangle \\
&= \prod_{i=1}^{N} w_{o_i}^{o<i}.
\end{aligned} \tag{A.35}$$

The expression for p_2 is more complicated. We first define the following recursive expression

$$\xi_a^{o<i}[m] \coloneqq \begin{cases} 2^{-n} \operatorname{Tr}\left(P_a \mathcal{C}_{m-1}(I + P_a \xi_a^{o<i}[m-1])\right), & 2 \le m \le M^{o<i}, \\ \langle A^{o<i}| P_a |A^{o<i}\rangle, & m = 1. \end{cases} \tag{A.36}$$

The expression for p_2 can then be calculated as follows

$$
\begin{aligned}
& p_2(o_{1:N}) \\
&= \mathop{\mathbb{E}}_{a\neq 0} \prod_{i=1}^{N} w_{o_i}^{o_{<i}} 2^n \left\langle B_{o_i}^{o_{<i}} \right| \Lambda_a(\mathcal{C}_{M^{o_{<i}}-1}^{o_{<i}}(\cdots \mathcal{C}_2^{o_{<i}}(\Lambda_a(\mathcal{C}_1^{o_{<i}}(\Lambda_a(|A^{o_{<i}}\rangle\langle A^{o_{<i}}|)))) \cdots)) \left| B_{o_i}^{o_{<i}} \right\rangle \\
&= \mathop{\mathbb{E}}_{a\neq 0} \prod_{i=1}^{N} w_{o_i}^{o_{<i}} \left\langle B_{o_i}^{o_{<i}} \right| \Lambda_a(\mathcal{C}_{M^{o_{<i}}-1}^{o_{<i}}(\cdots \mathcal{C}_2^{o_{<i}}(\Lambda_a(\mathcal{C}_1^{o_{<i}}(I + P_a \xi_a^{o_{<i}}[1]))) \cdots)) \left| B_{o_i}^{o_{<i}} \right\rangle \\
&= \mathop{\mathbb{E}}_{a\neq 0} \prod_{i=1}^{N} w_{o_i}^{o_{<i}} \left\langle B_{o_i}^{o_{<i}} \right| \Lambda_a(\mathcal{C}_{M^{o_{<i}}-1}^{o_{<i}}(\cdots \mathcal{C}_2^{o_{<i}}(I + 2^{-n} P_a \operatorname{Tr}(P_a \mathcal{C}_1^{o_{<i}}(I + P_a \xi_a^{o_{<i}}[1])) \cdots)) \left| B_{o_i}^{o_{<i}} \right\rangle \\
&= \mathop{\mathbb{E}}_{a\neq 0} \prod_{i=1}^{N} w_{o_i}^{o_{<i}} \left\langle B_{o_i}^{o_{<i}} \right| \Lambda_a(\mathcal{C}_{M^{o_{<i}}-1}^{o_{<i}}(\cdots \mathcal{C}_2^{o_{<i}}(I + P_a \xi_a^{o_{<i}}[2]) \cdots)) \left| B_{o_i}^{o_{<i}} \right\rangle \\
&= \cdots \\
&= \mathop{\mathbb{E}}_{a\neq 0} \prod_{i=1}^{N} w_{o_i}^{o_{<i}} \left\langle B_{o_i}^{o_{<i}} \right| I + P_a \xi_a^{o_{<i}}[M^{o_{<i}}] \left| B_{o_i}^{o_{<i}} \right\rangle \\
&= \mathop{\mathbb{E}}_{a\neq 0} \prod_{i=1}^{N} w_{o_i}^{o_{<i}} (1 + \xi_a^{o_{<i}}[M^{o_{<i}}] \left\langle B_{o_i}^{o_{<i}} \right| P_a \left| B_{o_i}^{o_{<i}} \right\rangle).
\end{aligned}
\tag{A.37}
$$

The third line uses the fact that $\mathcal{C}_k^{o_{<i}}$ is trace-preserving. The total variance distance between p_1 and p_2 is then

$$
\begin{aligned}
\mathrm{TV}(p_1, p_2) &= \mathop{\mathbb{E}}_{a\neq 0} \sum_{\substack{o_{1:N} \text{ s.t.} \\ p_1(o_{1:N}) \geq p_2(o_{1:N})}} \left(\prod_{i=1}^{N} w_{o_i}^{o_{<i}} \right) \left(1 - \prod_{i=1}^{N} \left(1 + \xi_a^{o_{<i}}[M^{o_{<i}}] \left\langle B_{o_i}^{o_{<i}} \right| P_a \left| B_{o_i}^{o_{<i}} \right\rangle \right) \right) \\
&= \mathop{\mathbb{E}}_{a\neq 0} \sum_{\substack{o_{1:N} \text{ s.t.} \\ p_1(o_{1:N}) \geq p_2(o_{1:N})}} p_1(o_{1:N}) \left(1 - \prod_{i=1}^{N} \left(1 + \xi_a^{o_{<i}}[M^{o_{<i}}] \left\langle B_{o_i}^{o_{<i}} \right| P_a \left| B_{o_i}^{o_{<i}} \right\rangle \right) \right).
\end{aligned}
\tag{A.38}
$$

We now need a bound of $|\xi_a^{o_{<i}}[M^{o_{<i}}]| \leq 1$, which can be shown by induction. We see

$|\xi_a^{o<i}[1]| \leq 1$ by definition. Suppose $|\xi_a^{o<i}[m-1]| \leq 1$, we have

$$\begin{aligned} |\xi_a^{o<i}[m]| &= \left| \mathrm{Tr}\left(P_a \mathcal{C}_{m-1}^{o<i}\left(\frac{I + P_a \xi_a^{o<i}[m-1]}{2^n} \right) \right) \right| \\ &\leq \|P_a\|_\infty \, \mathrm{Tr}\left| \mathcal{C}_{m-1}^{o<i}\left(\frac{I + P_a \xi_a^{o<i}[m-1]}{2^n} \right) \right| \\ &= \mathrm{Tr}\left(\frac{I + P_a \xi_a^{o<i}[m-1]}{2^n} \right) \\ &= 1. \end{aligned} \tag{A.39}$$

The first line is by the defining recursive expression; The second line is by the tracial matrix Hölder inequality; The third line uses the fact that $\mathcal{C}_{m-1}^{o<i}$ is a positive map, and that $2^{-n}\left(I + P_a \xi_a^{o<i}[m-1]\right)$ is positive semidefinite thanks to the induction hypothesis $|\xi_a^{o<i}[m-1]| \leq 1$. Thus we can remove the modulus within the trace, and also remove $\mathcal{C}_{m-1}^{o<i}$ as it is trace-preserving. By induction, we've shown $|\xi_a^{o<i}[M^{o<i}]| \leq 1$.

The remaining part of bounding $\mathrm{TV}(p_1, p_2)$ is basically the same as in the proof of Lemma 1.2. We repeat it here for completeness. Let $|B\rangle$ be any n-qubit pure state. Consider the following subset of n-qubit Pauli operators

$$G \coloneqq \left\{ a \in \{1, \cdots, 4^n - 1\} \ : \ |\langle B| P_a |B\rangle| \leq \left(\frac{1}{2^n + 1} \right)^{1/3} \right\}. \tag{A.40}$$

We claim that the size of G satisfies

$$|G| \geq \left(1 - \left(\frac{1}{2^n + 1} \right)^{1/3} \right) (4^n - 1), \tag{A.41}$$

which can be shown by contradiction: Suppose this does not hold, we would have

$$\begin{aligned}\sum_{a=1}^{4^n-1} \langle B| P_a |B\rangle^2 &\geq \left(\frac{1}{2^n+1}\right)^{2/3} \times (4^n - 1 - |G|) \\ &> \left(\frac{1}{2^n+1}\right)^{2/3} \times \left(\frac{1}{2^n+1}\right)^{1/3} (4^n - 1) \\ &= 2^n - 1.\end{aligned} \tag{A.42}$$

However, the L.H.S. of the above can be calculated as

$$\sum_{a=1}^{4^n-1} \langle B| P_a |B\rangle^2 = \sum_{a=0}^{4^n-1} \langle B| P_a |B\rangle^2 - 1 = 2^n - 1. \tag{A.43}$$

This leads to contradition, and hence proves the desired lower bound on $|G|$.

Now, we define another subset of n-qubit Pauli operators, conditioned on the measurement outcomes $o_{1:N}$, as follows

$$G^{(o_{1:N})} \coloneqq \left\{ a \in \{1, \cdots, 4^n - 1\} \ : \ \left|\langle B_{o_i}^{o_{<i}}| P_a |B_{o_i}^{o_{<i}}\rangle\right| \leq \left(\frac{1}{2^n+1}\right)^{1/3}, \ \forall i = 1, \cdots, N \right\}. \tag{A.44}$$

By applying a "union bound" on Eq. (A.40), we immediately have the following lower bound on the size of $G^{(o_{1:N})}$:

$$\left|G^{(o_{1:N})}\right| \geq \left(1 - N\left(\frac{1}{2^n+1}\right)^{1/3}\right)(4^n - 1). \tag{A.45}$$

We are now ready to upper bound $\mathrm{TV}(p_1, p_2)$ from Eq. (A.38). The strategy is to divide the sum over all Pauli operators into $G^{(o_{1:N})}$ and $\mathsf{P}^n \backslash G^{(o_{1:N})}$. All terms within the former group are small thanks to the definition of $G^{(o_{1:N})}$; Terms within the latter group could be large, but the total number of them are small. Combining these two gives a pretty good

upper bound on $\mathrm{TV}(p_1, p_2)$. In math,

$$
\begin{aligned}
&\mathrm{TV}(p_1, p_2) \\
=& \sum_{\substack{o_{1:N} \text{ s.t.} \\ p_1(o_{1:N}) \geq p_2(o_{1:N})}} \frac{p_1(o_{1:N})}{4^n - 1} \left(\sum_{a \in G^{(o_{1:N})}} + \sum_{\substack{a \in \mathbb{Z}_2^{2n} \setminus G^{(o_{1:N})}, \\ a \neq 0}} \right) \left(1 - \prod_{i=1}^{N} \left(1 + \xi_a^{o_{<i}} [M^{o_{<i}}] \left\langle B_{o_i}^{o_{<i}} \right| P_a \left| B_{o_i}^{o_{<i}} \right\rangle \right) \right) \\
\leq& \sum_{\substack{o_{1:N} \text{ s.t.} \\ p_1(o_{1:N}) \geq p_2(o_{1:N})}} \frac{p_1(o_{1:N})}{4^n - 1} \left(\left| G^{(o_{1:N})} \right| \times \left(1 - \left(1 - \left(\frac{1}{2^n + 1} \right)^{1/3} \right)^N \right) + \left(4^n - 1 - \left| G^{(o_{1:N})} \right| \right) \right) \\
\leq& \sum_{\substack{o_{1:N} \text{ s.t.} \\ p_1(o_{1:N}) \geq p_2(o_{1:N})}} p_1(o_{1:N}) \left(N \left(\frac{1}{2^n + 1} \right)^{1/3} + N \left(\frac{1}{2^n + 1} \right)^{1/3} \right) \\
\leq& 2N \left(\frac{1}{2^n + 1} \right)^{1/3}.
\end{aligned}
\tag{A.46}
$$

The first inequality applies the bound $|\xi_a^{o_{<i}}[M^{o_{<i}}]| \leq 1$. Besides, the first sum uses the bound from the definition of $G^{(o_{1:N})}$, and the second sum is by $\left| \left\langle B_{o_i}^{o_{<i}} \right| P_a \left| B_{o_i}^{o_{<i}} \right\rangle \right| \leq 1$. The second inequality uses the bounds $\left| G^{(o_{1:N})} \right| \leq 4^n - 1$ for the first sum and Eq. (A.45) for the second sum, as well as the fact that $1 - (1 - x)^N \leq Nx$ for all $0 \leq x \leq 1$. Now we have obtained the claimed bound $\mathrm{TV}(p_1, p_2) = O(N 2^{-n/3})$, and hence complete the proof of Theorem 1.3. □

A lower bound for the most general entangled strategies In this section, we prove a lower bound of $\Omega(n)$ for the fully entangled estimation strategies, which is the most general measurement strategies one can do to learn an unknown channel even with the help of quantum computers, see Fig. 3.1 (a). Note that, since we assume there is an unlimited amount of quantum memory, we can without loss of generality eliminate any intermediate measurements, and only conduct one joint measurement after sequentially processing all N samples of the channel.

Theorem 1.4. *For any fully entangled measurement protocols that give an estimate* $\widehat{\boldsymbol{\lambda}}$ *for*

the Pauli eigenvalues $\boldsymbol{\lambda}$ *of an arbitrary unknown* n*-qubit Pauli channel* Λ *such that*

$$|\widehat{\lambda}_a - \lambda_a| < \frac{1}{2}, \quad \forall a \in \mathbb{Z}_2^{2n} \tag{A.47}$$

holds with high probability, the number of samples of Λ *required is at least* $\Omega(n)$*.*

Proof. We first show that, by using a technique known as *teleportation stretching* [PLOB17, PLLP19], any fully entangled measurement protocols for N copies of an arbitrary Pauli channel Λ can be simulated by a joint measurement on N copies of the Choi state J_Λ which is defined as

$$\begin{aligned} J_\Lambda &\coloneqq \Lambda \otimes \mathbb{1}(|\Psi^+\rangle\langle\Psi^+|) \\ &= \frac{1}{4^n} \sum_{a \in \mathbb{Z}_2^{2n}} \lambda_a P_a \otimes P_a^T. \end{aligned} \tag{A.48}$$

This follows from the existence of a quantum channel $\mathcal{T}$ such that $\Lambda(\rho) = \mathcal{T}(\rho \otimes J_\Lambda)$ holds for all Pauli channels Λ. One possible construction of $\mathcal{T}$ is shown in Fig. A.1 (see [BDSW96]). In word, one first applies a Bell measurement on the input state ρ and half of the Choi state J_Λ. Then, conditioned on the Bell measurement outcome $|\Psi_b\rangle$, one applies a Pauli correction P_b on the other half of J_Λ, which will then be equal to $\Lambda(\rho)$. Indeed, the post-measurement state conditioned on Bell measurement outcome b is

$$\begin{aligned} \rho_b &\propto \langle\Psi_b|_{AB} \, \rho^A \otimes J_\Lambda^{BC} \, |\Psi_b\rangle_{AB} \\ &\propto \sum_{a \in \mathbb{Z}_2^{2n}} \lambda_a \, \langle\Psi_b| \, \rho \otimes P_a \, |\Psi_b\rangle \otimes P_a^T \\ &= \frac{1}{2^n} \sum_{a \in \mathbb{Z}_2^{2n}} \lambda_a (-1)^{\langle a,b \rangle} \operatorname{Tr}(\rho P_a) P_a. \end{aligned} \tag{A.49}$$

After applying the Pauli correction, the state becomes

$$P_b \rho_b P_b = \frac{1}{2^n} \sum_{a \in \mathbb{Z}_2^{2n}} \lambda_a \operatorname{Tr}(\rho P_a) P_a = \Lambda(\rho), \tag{A.50}$$

which justify the relation $\Lambda(\rho) = \mathcal{T}(\rho \otimes J_\Lambda)$.

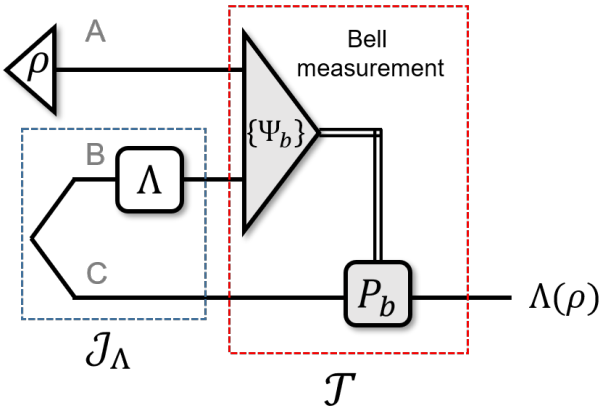


Figure A.1: Construction of the teleportation simulation channel $\mathcal{T}$.

With the help of teleportation stretching, one can reduce any measurement protocols for N copies of Λ to a single POVM measurement on N copies of J_Λ, as shown in Fig. A.2.

Now, recall the communication task defined in Sec. A.2, where Alice and Bob share the following "codebook"

$$(a, s) \in \{1, \cdots, 4^n - 1\} \times \{\pm 1\} \longrightarrow \Lambda_{(a,s)}(\cdot) = \frac{1}{2^n} \left(I \operatorname{Tr}(\cdot) + s P_a \operatorname{Tr}(P_a(\cdot)) \right), \tag{A.51}$$

and Alice randomly picks one possible (a, s) and send N copies of $\Lambda_{(a,s)}$ to Bob. If there exists a fully entangled estimation protocol using N samples and satisfying the assumption of Theorem 1.4, Bob can determine Alice's choice of (a, s) with high probability. According to Fano's inequality, the mutual information between the random variable pair (a, s) and Bob's measurement result o has the following lower bound

$$I((a, s) : o) \geq \Omega(\log 2(4^n - 1)) = \Omega(n). \tag{A.52}$$

One the other hand, since any measurement Bob conducts can be simulated by a measurement on N copies of $J_{\Lambda_{(a,s)}}$, Holevo's theorem [Hol73, Wil13] can be apply to $I((a, s) : o)$.

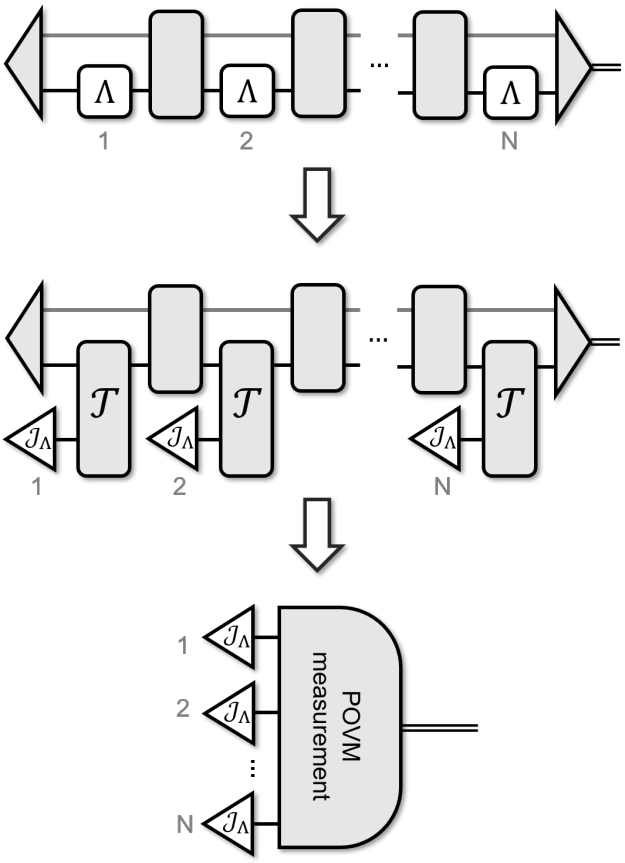


Figure A.2: Teleportation stretching [PLOB17, PLLP19]: simulation of an arbitrary entangled measurement on Λ by a single measurement on J_Λ. One just needs to simulate every application of $\Lambda(\cdot)$ by $\mathcal{T}((\cdot) \otimes \rho)$. The whole measurement protocol then become a joint POVM measurement on N copies of J_Λ with no adaptivity.

We have

$$\begin{aligned} I((a,s):o) &\leq S\left(\mathop{\mathbb{E}}_{(a,s)} J_{\Lambda_{(a,s)}}^{\otimes N}\right) - \mathop{\mathbb{E}}_{(a,s)} S\left(J_{\Lambda_{(a,s)}}^{\otimes N}\right) \\ &= S\left(\mathop{\mathbb{E}}_{(a,s)} J_{\Lambda_{(a,s)}}^{\otimes N}\right) - N \mathop{\mathbb{E}}_{(a,s)} S\left(J_{\Lambda_{(a,s)}}\right) \\ &\leq 2nN - (2n-1)N \\ &= N. \end{aligned} \tag{A.53}$$

In the third line, the first term is a trivial upper bound for the von Neumann entropy on a 2^{2nN}-dimensional Hilbert space. The second term uses the observation that

$$J_{\Lambda_{(a,s)}} = \frac{1}{4^n}(I \otimes I + sP_a \otimes P_a^T) \tag{A.54}$$

is a maximally mixed state on a 2^{2n-1}-dimensional Hilbert space, thus $S(J_{\Lambda_{(a,s)}}) = 2n-1$. This yields the lower bound $N = \Omega(n)$. □

A.3 On the definitions of entanglement-free learning

In this section, we formally define the separable schemes and the classical-memory-assisted schemes for quantum learning. Then, we prove the equivalence between the two schemes. Namely, either one can simulate the other with the same sample complexity.

Definition 1.4. *A **separable scheme** (SEP) for learning properties from N copies of a channel Λ acting on $\mathcal{L}(\mathcal{H}_S)$ is specified by the following elements:*

1. *An arbitrarily large ancillary system $\mathcal{H}_A$ whose dimension can depend on n.*
2. *A collection of processing channels $\{\mathcal{C}_t\}_{t=1}^{N-1}$ acting on $\mathcal{L}(\mathcal{H}_A \otimes \mathcal{H}_S)$ such that each $\mathcal{C}_t$ is separable across A and S. That is,*

$$\mathcal{C}_t = \sum_j \mathcal{A}_{t,j}^A \otimes \mathcal{B}_{t,j}^S, \quad \text{where } \mathcal{A}_{t,j},\ \mathcal{B}_{t,j} \text{ is CPTNI.}$$

3. *An initial state ρ_0 and a final POVM measurement $\{E_k\}_k$ acting on $\mathcal{H}_A \otimes \mathcal{H}_S$ such that both of them are separable across A and S. That is,*

$$\rho_0 = \sum_j \sigma_j^A \otimes \gamma_j^S, \quad E_k = \sum_j M_{k,j}^A \otimes N_{k,j}^S, \qquad \text{for } \sigma_j, \gamma_j, M_{k,j}, N_{k,j} \geq 0.$$

The separable scheme works by inputing ρ_0, interleaving N copies of Λ with the processing channels $\mathcal{C}_t$, and measuring $\{E_k\}_k$ at the end, yielding the following outcome distribution:

$$\Pr[k|\Lambda] = \mathrm{Tr}\left[E_k \left((\mathbb{1}^A \otimes \Lambda^S) \circ \mathcal{C}_{N-1} \circ \cdots \circ (\mathbb{1}^A \otimes \Lambda^S) \circ \mathcal{C}_1 \circ (\mathbb{1}^A \otimes \Lambda^S)\right)(\rho_0)\right].$$

Finally, an algorithm is specified to predict the desired properties based on the observed outcome k.

Since we allow arbitrarily large ancilla for separable scheme, any adaptive control can be taken into account. We note again that separable operations contains LOCC as a subset.

Definition 1.5. *A **classical-memory-assisted scheme** for learning properties from N copies of a channel Λ acting on $\mathcal{L}(\mathcal{H}_S)$ is specified by the following elements:*

1. *N arbitrarily large classical registers $\{A_t\}_{t=0}^{N-1}$.*

2. *A collection of processing channels (or, quantum instruments) $\{\mathcal{C}^{\boldsymbol{o}<t}\}_{t=1}^{N-1}$ defined as*

$$\mathcal{C}^{\boldsymbol{o}<t}(\rho) \coloneqq \sum_{o_t} |o_t\rangle\langle o_t|^{A_t} \otimes \mathcal{C}_{o_t}^{\boldsymbol{o}<t}(\rho)^S, \qquad \text{for } \mathcal{C}_{o_t}^{\boldsymbol{o}<t} \in \mathrm{CPTNI}(\mathcal{L}(\mathcal{H}_S)),$$

   where the superscript of $o_{<t}$ indicates that the t-th processing channels can be chosen adaptively conditioned on the previous t classical registers.

3. *An ensemble of initial states*

$$\rho_0 \coloneqq \sum_{o_0} |o_0\rangle\langle o_0|^{A_0} \otimes \rho_{o_0}^S,$$

where ρ_{o_0} is some sub-normalized quantum state, and a final POVM measurement

$$\{E_{o_N}^{o_{<N}}\}_{o_N}^{S},$$

which can be chosen adaptively conditioned on all previous outcomes.

The scheme works by inputing the initial state, interleaving N copies of Λ with the processing channels, and measuring at the end, yielding the following outcome distribution:

$$\Pr[\boldsymbol{o}|\Lambda] \equiv \Pr[o_{0:N}|\Lambda] = \mathrm{Tr}\left[E_{o_N}^{o_{<N}}\left(\Lambda \circ \mathcal{C}_{o_{N-1}}^{o_{<N-1}} \circ \cdots \circ \mathcal{C}_{o_2}^{o_{<2}} \circ \Lambda \circ \mathcal{C}_{o_1}^{o_0} \circ \Lambda\right)(\rho_{o_0})\right].$$

Finally, an algorithm is specified to predict the desired properties based on the observed outcome $\boldsymbol{o}$.

We remind that classical-memory-assisted schemes can be understood as quantum circuits assisted by mid-circuit measurement and feed-forward control.

We now prove the equivalence between these two schemes. We call a scheme A can be **simulated** by another scheme B if there exists a mapping $\mathcal{S} : \boldsymbol{o}_B \mapsto \boldsymbol{o}_A$ independent of Λ such that $\Pr_A[\boldsymbol{o}_A|\Lambda] = \sum_{\boldsymbol{o}_B:\mathcal{S}(\boldsymbol{o}_B)=\boldsymbol{o}_A} \Pr_B[\boldsymbol{o}_B|\Lambda]$ for all $\boldsymbol{o}_A$ and Λ.

Proposition 1.6. *Any separable scheme can be simulated by a classical-memory-assisted scheme using the same number of samples and vice versa.*

Proof. A classical-memory-assisted scheme can obviously be simulated by a separable scheme using the same number of samples, which is clear from Fig. 1 in the main text. Formally, choose the ancillary system to include all the classical registers, $\mathcal{H}_A = \otimes_{t=0}^{N-1}\mathcal{H}_{A_t}$, and choose the processing channel to be

$$\mathcal{C}_t(\cdot) \coloneqq \sum_{\boldsymbol{o}_{\le t}} \left(\bigotimes_{k=0}^{t-1} |o_k\rangle\langle o_k|(\cdot)|o_k\rangle\langle o_k|^{A_k}\right) \otimes \mathrm{Tr}_{A_t}(\cdot)|o_t\rangle\langle o_t|^{A_t} \otimes \mathbb{1}^{A_{t+1:N-1}} \otimes \mathcal{C}_{o_t}^{\boldsymbol{o}_{<t}}(\cdot)^{S} \quad \text{(A.55)}$$

which are separable channels between A and S. In word, the t-th processing channel picks up the quantum instrument $\mathcal{C}^{\boldsymbol{o}<t}$ according to $A_{0:t-1}$, writes down the outcome to A_t, and does nothing to $A_{t+1:N}$. Alternatively, this channel can be written in the PTM representation as

$$\mathcal{C}_t^{AS} = \sum_{\boldsymbol{o}\leq t} \left(\bigotimes_{k=0}^{t-1} |o_k\rangle\!\rangle\langle\!\langle o_k|^{A_k} \right) \otimes |o_t\rangle\!\rangle\langle\!\langle I|^{A_t} \otimes \mathbb{1}^{A_{t+1:N-1}} \otimes \mathcal{C}_{o_t}^{\boldsymbol{o}<t S}. \tag{A.56}$$

where $|o_k\rangle\!\rangle$ is defined to be the vectorization of the computational basis state $|o_k\rangle\langle o_k|$. Additionally, choose the initial state as

$$\sum_{o_0} |o_0\rangle\langle o_0|^{A_0} \otimes \rho_{\text{junk}}^{A_{1:N-1}} \otimes \rho_{o_0}^{S},$$

where ρ_{junk} can be any quantum state, and the final POVM as

$$\left\{ \left(\bigotimes_{t=0}^{N-1} |o_t\rangle\langle o_t|^{A_t} \right) \otimes \left(E_{o_N}^{o<N} \right)^S \right\}_{\boldsymbol{o}}.$$

One can verify the outcome distribution will be the same for both schemes.

Now we prove the opposite direction. For a generic separable scheme, expand the outcome distribution,

$$\Pr[k|\Lambda] = \sum_{\boldsymbol{j}_{0:N}} \operatorname{Tr}\left[M_{k,j_N}(\mathcal{A}_{N-1,j_{N-1}} \cdots \mathcal{A}_{1,j_1})(\sigma_{j_0})\right] \operatorname{Tr}\left[N_{k,j_N}(\Lambda\mathcal{B}_{N-1,j_{N-1}} \cdots \Lambda\mathcal{B}_{1,j_1}\Lambda)(\gamma_{j_0})\right]. \tag{A.57}$$

A crucial observation is that, the first factor is independent of Λ conditioned on $\boldsymbol{j}_{0:N}$. Therefore, instead of keeping track of the quantum state in the ancillary system, we just need to keep track of the index $\boldsymbol{j}$ using classical registers. Formally, we can design a classical-memory-assisted scheme with the following processing channels

$$\mathcal{C}^{j<t}(\rho^S) \coloneqq \sum_{j_t} |j_t\rangle\langle j_t|^{A_t} \otimes \frac{\operatorname{Tr}\left[(\mathcal{A}_{t,j_t}\mathcal{A}_{t-1,j_{t-1}} \cdots \mathcal{A}_{1,j_1})(\sigma_{j_0})\right]}{\operatorname{Tr}\left[(\mathcal{A}_{t-1,j_{t-1}} \cdots \mathcal{A}_{1,j_1})(\sigma_{j_0})\right]} \mathcal{B}_{t,j_t}(\rho^S). \tag{A.58}$$

One can verify that $\mathcal{C}^{j<t}$ is CPTP. Additionally, choose the initial state to be

$$\rho_0 = \sum_{j_0} |j_0\rangle\langle j_0|^{A_0} \otimes \mathrm{Tr}[\sigma_{j_0}]\gamma_{j_0},$$

and the final POVM measurement to be

$$E^{j<N}_{k,j_N} := \frac{\mathrm{Tr}\left[M_{k,j_N}(\mathcal{A}_{N-1,j_{N-1}} \cdots \mathcal{A}_{1,j_1})(\sigma_{j_0})\right]}{\mathrm{Tr}\left[(\mathcal{A}_{N-1,j_{N-1}} \cdots \mathcal{A}_{1,j_1})(\sigma_{j_0})\right]} \cdot N_{k,j_N}. \tag{A.59}$$

The outcome distribution can then be computed as

$$\Pr[k, j_{0:N}|\Lambda] = \mathrm{Tr}\left[M_{k,j_N}(\mathcal{A}_{N-1,j_{N-1}} \cdots \mathcal{A}_{1,j_1})(\sigma_{j_0})\right] \mathrm{Tr}\left[N_{k,j_N}(\Lambda\mathcal{B}_{N-1,j_{N-1}} \cdots \Lambda\mathcal{B}_{1,j_1}\Lambda)(\gamma_{j_0})\right]. \tag{A.60}$$

By tracing out the classical register $A_{0:N}$, we retrieve the distribution of the separable scheme. This means any separable scheme can be simulated by a classical-memory-assisted scheme using the same number of samples, which completes our proof. □

Sample complexity vs. Number of measurements. Since the separable schemes and classical-memory-assisted schemes are equivalent in terms of sample complexity N_{samp}, we will call both of them entanglement-free schemes. However, for the classical-memory-assisted schemes, there is an alternative figure of merit that is practically interesting, which we call *the number of measurements*, defined as follows

Definition 1.7. *The number of measurements, N_{meas}, of a classical-memory-assisted scheme is defined to be the number of non-trivial quantum instruments used in the worst case (where the final POVM measurement also counts). Here, a quantum instrument $\mathcal{C}^{S\to AS} = \sum_t |t\rangle\langle t|^A \otimes \mathcal{C}^S_t$ is called trivial if every $\mathcal{C}^S_t$ is proportional to some* CPTP *map.*

The reason why we call such quantum instruments trivial is that they are not extracting any information from the input, hence not doing a measurement. Indeed, the probability of seeing

$|t\rangle$ is independent of the input state for such quantum instruments. We also emphasize that, for different measurement outcome sequence there can be different numbers of non-trivial quantum instruments, and our definition focus on the worst case.

By definition, $N_{\text{samp}} \geq N_{\text{meas}}$, as each Λ is followed by one quantum instrument (or the final POVM measurement). The lower bound for the latter thus implies one for the former. In the following, when we talk about the number of measurements of an entanglement-free schemes, we are specifically referring to the classical-memory-assisted schemes.

A.4 Proof of Theorem 3.3

In this section, we present our main technical results, restated in Theorem 1.5.

Theorem 1.5. *If there exists a classical-memory-assisted scheme that, for any n-qubit Pauli channel Λ, outputs an estimator $\widehat{\lambda}_a$ such that $|\widehat{\lambda}_a - \lambda_a| \leq \varepsilon \leq 1/6$ with probability at least $2/3$ for any $a \in \mathsf{P}^n$, after making N rounds of measurement, then $N = \Omega(2^n/\varepsilon^2)$.*

Proof. We first prove a lower bound for the sample complexity, and then strengthen it to the number of measurements. Define the following set of Pauli channels

$$\begin{aligned} \Lambda_{a,\pm} &\coloneqq |\sigma_0\rangle\!\rangle\langle\!\langle\sigma_0| \pm \varepsilon_0 |\sigma_a\rangle\!\rangle\langle\!\langle\sigma_a|, \quad P_a \neq I; \\ \Lambda_0 &\coloneqq |\sigma_0\rangle\!\rangle\langle\!\langle\sigma_0|. \end{aligned} \tag{A.61}$$

Here we will set $\varepsilon_0 \leq 1/3$. Λ_0 is known as the completely depolarizing channel. To see $\Lambda_{a,\pm}$ is CPTP, we just need to check the non-negativity of the associated Pauli error rates

$$p_c = \frac{1}{4^n} \sum_b (-1)^{\langle c,b\rangle} \lambda_b = \frac{1}{4^n}(1 \pm \varepsilon_0) \geq 0, \quad \forall c \in \mathsf{P}^n. \tag{A.62}$$

A learning algorithm satisfying Theorem 1.5 with $\varepsilon = \varepsilon_0/2$ is able to distinguish any one of these Pauli channels with high success probability. Now, we define a partially-revealed

hypothesis-testing task similar to Ref. [HBC+22]: A referee first samples an $a \in \mathsf{P}^n \backslash I$ and $s = \pm 1$ according to uniform distribution. Then conducts one of the following actions with equal probability:

1. Sending N copies of Λ_0 to the player.

2. Sending N copies of $\Lambda_{a,s}$ to the player.

The player is then asked to measure the N copies of channels with any schemes. After the measurement has been completed, the referee reveals the value of a to the player. The player is now asked to guess which action the referee has taken, based on the obtained measurement outcome. Crucially, the player must completes all quantum measurement before the reveal of a, and can only do classical post-processing after that.

Suppose there exists a separable scheme satisfying the assumption of Theorem 1.5. Then the player can win the game with probability 2/3 for any (a, s): Just by querying λ_a after the referee revealed the value of a, the player can distinguish among $\{\Lambda_{a,-}, \Lambda_0, \Lambda_{a,+}\}$ with 2/3 chance. Denote the measurement outcome distribution for Λ_0, $\Lambda_{a,\pm}$ as p_0, $p_{a,\pm}$, respectively. Based on Le Cam's two-point method [LeC73], we must have

$$\mathbb{E}_{a\neq 0}\mathrm{TVD}(p_0,\ \mathbb{E}_{s=\pm 1}[p_{a,s}]) \geq (1 - 2 * \frac{1}{3}) = \frac{1}{3}. \tag{A.63}$$

Now we compute the L.H.S., *i.e.*, the average TVD.

$$\mathbb{E}_{a\neq 0}\mathrm{TVD}(p_0,\ \mathbb{E}_{s=\pm 1}[p_{a,s}]) = \mathbb{E}_{a\neq 0} \sum_{\boldsymbol{o}} \max\left\{0, p_0(\boldsymbol{o}) - \mathbb{E}_s p_{a,s}(\boldsymbol{o})\right\}. \tag{A.64}$$

We need some additional notations on the classical-memory-assisted scheme. Focusing on the t-th step, for every CP trace-non-increasing map $\mathcal{C}_{o_t}^{\boldsymbol{o}<t}$, choosing any Kraus representation $\mathcal{C}_{o_t}^{\boldsymbol{o}<t}(\cdot) = \sum_j K_j(\cdot)K_j^\dagger$ [Cho75], we define $E_{o_t}^{(\boldsymbol{o}<t)} := \sum_j K_j^\dagger K_j$ as the associated POVM

element. Indeed,

$$\mathrm{Tr}[E_{o_t}^{(\boldsymbol{o}_{<t})}\rho] = \mathrm{Tr}[\mathcal{C}_{o_t}^{\boldsymbol{o}_{<t}}(\rho)]. \tag{A.65}$$

On the other hand, we denote the state fed into the t-th $\Lambda_{a,s}$ as $\rho_{a,s}^{(\boldsymbol{o}_{<t})}$, which satisfies the following recurrence relation by definition

$$\rho_{a,s}^{(\boldsymbol{o}_{<t+1})} := \frac{\mathcal{C}_{o_t}^{\boldsymbol{o}_{<t}} \circ \Lambda(\rho_{a,s}^{(\boldsymbol{o}_{<t})})}{\mathrm{Tr}[\mathcal{C}_{o_t}^{\boldsymbol{o}_{<t}} \circ \Lambda(\rho_{a,s}^{(\boldsymbol{o}_{<t})})]}, \tag{A.66}$$

and the initial condition $\rho_{a,s}^{(\boldsymbol{o}_{<1})} \equiv \rho_{\mathrm{init}}$. Here we assume ρ_{init} to be fixed rather than drawing from an ensemble, without loss of generality, as the total variation distance is joint convex.

Now, focus on the probability distribution difference,

$$p_0(\boldsymbol{o}) - \mathbb{E}_s p_{a,s}(\boldsymbol{o}) \tag{A.67}$$

$$= \prod_{t=1}^{N} \langle\!\langle E_{o_t}^{(\boldsymbol{o}_{<t})} | \Lambda_0 | \rho_0^{(\boldsymbol{o}_{<t})} \rangle\!\rangle - \mathbb{E}_s \prod_{t=1}^{N} \langle\!\langle E_{o_t}^{(\boldsymbol{o}_{<t})} | \Lambda_{a,s} | \rho_{a,s}^{(\boldsymbol{o}_{<t})} \rangle\!\rangle \tag{A.68}$$

$$= \prod_{t=1}^{N} \frac{1}{2^n} \langle\!\langle E_{o_t}^{(\boldsymbol{o}_{<t})} | I \rangle\!\rangle \langle\!\langle I | \rho_0^{(\boldsymbol{o}_{<t})} \rangle\!\rangle - \mathbb{E}_s \prod_{t=1}^{N} \frac{1}{2^n} \left(\langle\!\langle E_{o_t}^{(\boldsymbol{o}_{<t})} | I \rangle\!\rangle \langle\!\langle I | \rho_{a,s}^{(\boldsymbol{o}_{<t})} \rangle\!\rangle + s\varepsilon_0 \langle\!\langle E_{o_t}^{(\boldsymbol{o}_{<t})} | P_a \rangle\!\rangle \langle\!\langle P_a | \rho_{a,s}^{(\boldsymbol{o}_{<t})} \rangle\!\rangle \right) \tag{A.69}$$

$$= \left(\prod_{t=1}^{N} \frac{1}{2^n} \mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})}) \right) \left(1 - \mathbb{E}_s \left(\prod_{t=1}^{N} \left(1 + s\varepsilon_0 \frac{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})} P_a)\, \mathrm{Tr}(P_a \rho_{a,s}^{(\boldsymbol{o}_{<t})})}{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})})} \right) \right) \right) \tag{A.70}$$

$$= p_0(\boldsymbol{o}) \left(1 - \mathbb{E}_s \left(\prod_{t=1}^{N} \left(1 + s\varepsilon_0 \frac{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})} P_a)\, \mathrm{Tr}(P_a \rho_{a,s}^{(\boldsymbol{o}_{<t})})}{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})})} \right) \right) \right) \tag{A.71}$$

Here $\rho_{a,s}^{(\boldsymbol{o}_{<t})}$ is the post-measurement state after the $(t-1)$th measurement, conditioned on

the outcome being $\boldsymbol{o}_{1:t-1}$. Now we try to bound the following term,

$$\mathbb{E}_s\left(\prod_{t=1}^{N}\left(1+s\varepsilon_0\frac{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)\,\mathrm{Tr}(P_a\rho_{a,s}^{(\boldsymbol{o}_{<t})})}{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})})}\right)\right) \tag{A.72}$$

$$=\mathbb{E}_s\exp\left(\sum_{t=1}^{N}\log\left(1+s\varepsilon_0\frac{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)\,\mathrm{Tr}(P_a\rho_{a,s}^{(\boldsymbol{o}_{<t})})}{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})})}\right)\right) \tag{A.73}$$

$$\geq\exp\left(\frac{1}{2}\sum_{t=1}^{N}\sum_{s=\pm1}\log\left(1+s\varepsilon_0\frac{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)\,\mathrm{Tr}(P_a\rho_{a,s}^{(\boldsymbol{o}_{<t})})}{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})})}\right)\right)=:\bigstar, \tag{A.74}$$

where the second line uses the fact that each term in the product is non-negative, and the third line uses Jensen's inequality. Now we define

$$\bar{\rho}_a^{(\boldsymbol{o}_{<t})}=\frac{1}{2}(\rho_{a,+}^{(\boldsymbol{o}_{<t})}+\rho_{a,-}^{(\boldsymbol{o}_{<t})}),\quad \Delta_a^{(\boldsymbol{o}_{<t})}=\frac{1}{2}(\rho_{a,+}^{(\boldsymbol{o}_{<t})}-\rho_{a,-}^{(\boldsymbol{o}_{<t})}). \tag{A.75}$$

Then above inequality can then be expressed as

$$\bigstar \tag{A.76}$$

$$=\exp\left(\frac{1}{2}\sum_{t=1}^{N}\log\left[\left(1+\varepsilon_0\frac{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)\,\mathrm{Tr}(P_a\rho_{a,+}^{(\boldsymbol{o}_{<t})})}{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})})}\right)\left(1-\varepsilon_0\frac{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)\,\mathrm{Tr}(P_a\rho_{a,-}^{(\boldsymbol{o}_{<t})})}{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})})}\right)\right]\right) \tag{A.77}$$

$$=\exp\left(\frac{1}{2}\sum_{t=1}^{N}\log\left[1-\left(\varepsilon_0^2\frac{\mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)}{\mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})})}\left(\mathrm{Tr}^2(P_a\bar{\rho}_a^{(\boldsymbol{o}_{<t})})-\mathrm{Tr}^2(P_a\Delta_a^{(\boldsymbol{o}_{<t})})\right)-\right.\right.\right.$$
$$\left.\left.\left.2\varepsilon_0\frac{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)\,\mathrm{Tr}(P_a\Delta_a^{(\boldsymbol{o}_{<t})})}{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})})}\right)\right]\right) \tag{A.78}$$

$$\geq\exp\left(\frac{1}{2}\sum_{t=1}^{N}\log\left[1-\left(\varepsilon_0^2\frac{\mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)}{\mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})})}\,\mathrm{Tr}^2(P_a\bar{\rho}_a^{(\boldsymbol{o}_{<t})})+2\varepsilon_0\left|\frac{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)\,\mathrm{Tr}(P_a\Delta_a^{(\boldsymbol{o}_{<t})})}{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})})}\right|\right)\right]\right) \tag{A.79}$$

$$\geq\exp\left(-\sum_{t=1}^{N}\left(\varepsilon_0^2\frac{\mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)}{\mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})})}\,\mathrm{Tr}^2(P_a\bar{\rho}_a^{(\boldsymbol{o}_{<t})})+2\varepsilon_0\left|\frac{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)\,\mathrm{Tr}(P_a\Delta_a^{(\boldsymbol{o}_{<t})})}{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})})}\right|\right)\right) \tag{A.80}$$

$$\geq 1-\sum_{t=1}^{N}\left(\varepsilon_0^2\frac{\mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)}{\mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})})}\,\mathrm{Tr}^2(P_a\bar{\rho}_a^{(\boldsymbol{o}_{<t})})+2\varepsilon_0\left|\frac{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)\,\mathrm{Tr}(P_a\Delta_a^{(\boldsymbol{o}_{<t})})}{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})})}\right|\right). \tag{A.81}$$

The fourth line uses the fact that $\log(1-x) \geq -2x$ for $x \in [0, 0.79]$ and that the terms inside the inner braket is upper bounded by $\varepsilon_0^2 + 2\varepsilon_0 \leq 0.78$. The last line uses $\exp(-x) \geq 1 - x$ for any x. Now substitute this back to the expression of avergae TVD.

$$\mathbb{E}_{a\neq 0}\mathrm{TVD}(p_0,\ \mathbb{E}_{s=\pm 1}[p_{a,s}]) \tag{A.82}$$

$$\leq \mathbb{E}_{a\neq 0}\sum_{\boldsymbol{o}}\max\left\{0, p_0(\boldsymbol{o})\sum_{t=1}^{N}\left(\varepsilon_0^2\frac{\mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)}{\mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})})}\mathrm{Tr}^2(P_a\bar{\rho}_a^{(\boldsymbol{o}_{<t})}) + 2\varepsilon_0\left|\frac{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)\,\mathrm{Tr}(P_a\Delta_a^{(\boldsymbol{o}_{<t})})}{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})})}\right|\right)\right\} \tag{A.83}$$

$$= \mathbb{E}_{a\neq 0}\sum_{\boldsymbol{o}}p_0(\boldsymbol{o})\sum_{t=1}^{N}\left(\varepsilon_0^2\frac{\mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)}{\mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})})}\mathrm{Tr}^2(P_a\bar{\rho}_a^{(\boldsymbol{o}_{<t})}) + 2\varepsilon_0\left|\frac{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)\,\mathrm{Tr}(P_a\Delta_a^{(\boldsymbol{o}_{<t})})}{\mathrm{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})})}\right|\right). \tag{A.84}$$

The last line is because the second argument of max is now nonnegative. Now we bound the two terms from above. For the first term, note that

$$\mathbb{E}_{a\neq 0}\frac{\mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)}{\mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})})}\mathrm{Tr}^2(P_a\bar{\rho}_a^{(\boldsymbol{o}_{<t})}) \leq \mathbb{E}_{a\neq 0}\frac{\mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)}{\mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})})} \tag{A.85}$$

$$= \frac{1}{4^n-1}\frac{\sum_a \mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a) - \mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})})}{\mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})})} \tag{A.86}$$

$$= \frac{1}{4^n-1}\frac{2^n\,\mathrm{Tr}({E_{o_t}^{(\boldsymbol{o}_{<t})}}^2) - \mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})})}{\mathrm{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})})} \tag{A.87}$$

$$\leq \frac{2^n}{4^n-1}. \tag{A.88}$$

The third line is Pauli twirling. The last line uses the fact that $E_{o_t}^{(\boldsymbol{o}_{<t})}$ is positive semi-definite and that l_2-norm is upper bounded by l_1-norm.

Here we pause to remark that, if channel concatenation were not allowed, the t-th input state $\rho_{a,s}^{\boldsymbol{o}_{<t}}$ would be independent of (a, s) conditioned on $\boldsymbol{o}_{<t}$, and thus $\Delta_a^{(\boldsymbol{o}_{<t})} = 0$. In that case, we would only have the first term, and the proof would already be completed, similar to [HBC+22, CCHL22]. The major challenge in our setting is how to address the second term.

For the second term, note that

$$\mathbb{E}_{a\neq 0}\left|\frac{\operatorname{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)\operatorname{Tr}(P_a\Delta_a^{(\boldsymbol{o}_{<t})})}{\operatorname{Tr}(E_{o_t}^{(\boldsymbol{o}_{<t})})}\right| \leq \sqrt{\mathbb{E}_{a\neq 0}\frac{\operatorname{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})}P_a)}{\operatorname{Tr}^2(E_{o_t}^{(\boldsymbol{o}_{<t})})}}\sqrt{\mathbb{E}_{a\neq 0}\operatorname{Tr}^2(P_a\Delta_a^{(\boldsymbol{o}_{<t})})} \tag{A.89}$$

$$\leq \sqrt{\frac{2^n}{4^n-1}}\sqrt{\mathbb{E}_{a\neq 0}\operatorname{Tr}^2(P_a\Delta_a^{(\boldsymbol{o}_{<t})})}. \tag{A.90}$$

The first line is Cauchy-Schwarz. The second line is the same as the derivation for the first term. The remaining of the proof is to upper bound the following

$$\mathbb{E}_{a\neq 0}\operatorname{Tr}^2(P_a\Delta_a^{(\boldsymbol{o}_{<t})}) = \mathbb{E}_{a\neq 0}\operatorname{Tr}^2(P_a(\rho_{a,+}^{(\boldsymbol{o}_{<t})}-\rho_{a,-}^{(\boldsymbol{o}_{<t})})/2) = \mathbb{E}_{a\neq 0}\left(\left(\mu_{a,+}^{(\boldsymbol{o}_{<t})}-\mu_{a,-}^{(\boldsymbol{o}_{<t})}\right)/2\right)^2. \tag{A.91}$$

where we've defined $\mu_{a,\pm}^{(\boldsymbol{o}_{<t})} \coloneqq \operatorname{Tr}(P_a\rho_{a,\pm}^{(\boldsymbol{o}_{<t})})$. We have the following recurrence relation

$$\rho_{a,\pm}^{(\boldsymbol{o}_{<t+1})} = \frac{(\mathcal{C}_{o_t}^{(\boldsymbol{o}_{<t})}\circ\Lambda_{a,\pm})(\rho_{a,\pm}^{(\boldsymbol{o}_{<t})})}{\operatorname{Tr}\left[(\mathcal{C}_{o_t}^{(\boldsymbol{o}_{<t})}\circ\Lambda_{a,\pm})(\rho_{a,\pm}^{(\boldsymbol{o}_{<t})})\right]} = \frac{\mathcal{C}_{o_t}^{(\boldsymbol{o}_{<t})}(I\pm\varepsilon_0\mu_{a,\pm}^{(\boldsymbol{o}_{<t})}P_a)}{\operatorname{Tr}\left[\mathcal{C}_{o_t}^{(\boldsymbol{o}_{<t})}(I\pm\varepsilon_0\mu_{a,\pm}^{(\boldsymbol{o}_{<t})}P_a)\right]}, \tag{A.92}$$

Now, let $c_{a,b}^{(\boldsymbol{o}_{\leq t})} \coloneqq \operatorname{Tr}[P_a\mathcal{C}_{o_t}^{(\boldsymbol{o}_{<t})}(P_b)]/2^n$, which is the Pauli transfer matrix representation of $\mathcal{C}_{o_t}^{(\boldsymbol{o}_{<t})}$. Taking expectation value of P_a on both side of the above equation, we get

$$\mu_{a,\pm}^{(\boldsymbol{o}_{<t+1})} = \frac{c_{a,0}^{(\boldsymbol{o}_{\leq t})}\pm\varepsilon_0\mu_{a,\pm}^{(\boldsymbol{o}_{<t})}c_{a,a}^{(\boldsymbol{o}_{\leq t})}}{c_{0,0}^{(\boldsymbol{o}_{\leq t})}\pm\varepsilon_0\mu_{a,\pm}^{(\boldsymbol{o}_{<t})}c_{0,a}^{(\boldsymbol{o}_{\leq t})}}. \tag{A.93}$$

For clarity, we omit the superscript of $c_{a,b}$ and denote $\mu_{a,\pm}^{(\boldsymbol{o}_{<t})} \equiv \mu_{a,\pm}$, $\mu_{a,\pm}^{(\boldsymbol{o}_{<t+1})} \equiv \mu'_{a,\pm}$.

The difference between $\mu'_{a,\pm}$ satisfies

$$\mu'_{a,+} - \mu'_{a,-} \tag{A.94}$$

$$= \left(\frac{c_{a,0}}{c_{0,0} + \varepsilon_0 \mu_{a,+} c_{0,a}} - \frac{c_{a,0}}{c_{0,0} - \varepsilon_0 \mu_{a,-} c_{0,a}}\right) + \varepsilon_0 \left(\frac{\mu_{a,+} c_{a,a}}{c_{0,0} + \varepsilon_0 \mu_{a,+} c_{0,a}} + \frac{\mu_{a,-} c_{a,a}}{c_{0,0} - \varepsilon_0 \mu_{a,-} c_{0,a}}\right) \tag{A.95}$$

$$\equiv (X) + (Y), \tag{A.96}$$

where we denote the first and second braket as X, Y, respectively. Now we have that

$$|X| \leq \frac{|c_{a,0}|}{(1-\varepsilon_0)c_{0,0}} - \frac{|c_{a,0}|}{(1+\varepsilon_0)c_{0,0}} = \frac{2\varepsilon_0}{1-\varepsilon_0^2}\frac{|c_{a,0}|}{c_{0,0}}. \tag{A.97}$$

as $|\mu_{a,\pm}| \leq 1$ and $|c_{0,a}| \leq c_{0,0}$. To see the later, notice that every completely-positive map has a Kraus representation [Cho75], thus

$$|c_{a,b}| \equiv |\mathrm{Tr}[P_a \mathcal{C}_{o_t}(P_b/2^n)]| \tag{A.98}$$

$$\leq \sum_j \left|\mathrm{Tr}[P_a K_j P_b K_j^\dagger]/2^n\right| \tag{A.99}$$

$$\leq \sum_j \sqrt{\mathrm{Tr}[P_a K_j K_j^\dagger P_a^\dagger]}\sqrt{\mathrm{Tr}[P_b K_j^\dagger K_j P_b^\dagger]}/2^n \tag{A.100}$$

$$= \sum_j \mathrm{Tr}[K_j K_j^\dagger]/2^n = \mathrm{Tr}[\mathcal{C}_{o_t}(I/2^n)] \equiv c_{0,0}, \tag{A.101}$$

where the third line is Cauchy-Schwarz. For the second term,

$$|Y| \leq \frac{\varepsilon_0(|\mu_{a,+}| + |\mu_{a,-}|)|c_{a,a}|}{(1-\varepsilon_0)c_{0,0}} \leq \frac{\varepsilon_0}{1-\varepsilon_0}(|\mu_{a,+}| + |\mu_{a,-}|), \tag{A.102}$$

as $|c_{a,a}| \leq c_{0,0}$ which has been proved. Therefore,

$$\mathbb{E}_{a\neq 0}\,\mathrm{Tr}^2(P_a\Delta_a^{(\boldsymbol{o}_{<t})}) = \mathbb{E}_{a\neq 0}((X+Y)/2)^2 \tag{A.103}$$

$$\leq \left(\mathbb{E}_{a\neq 0}X^2 + \mathbb{E}_{a\neq 0}Y^2\right)/2 \tag{A.104}$$

$$\leq \frac{1}{2}(\frac{2\varepsilon_0}{1-\varepsilon_0^2})^2\mathbb{E}_{a\neq 0}\frac{c_{a,0}^2}{c_{0,0}^2} + \frac{1}{2}(\frac{\varepsilon_0}{1-\varepsilon_0})^2\mathbb{E}_{a\neq 0}(|\mu_{a,+}| + |\mu_{a,-}|)^2 \tag{A.105}$$

$$\leq \frac{1}{2}(\frac{2\varepsilon_0}{1-\varepsilon_0^2})^2\frac{2^n}{4^n-1} + \frac{1}{2}(\frac{\varepsilon_0}{1-\varepsilon_0})^2\mathbb{E}_{a\neq 0}(2\mu_{a,+}^2 + 2\mu_{a,-}^2). \tag{A.106}$$

We use $(a+b)^2 \leq 2a^2 + 2b^2$ in the second and last line. For the last line we also use

$$\mathbb{E}_{a\neq 0}\frac{c_{a,0}^2}{c_{0,0}^2} \equiv \frac{\mathbb{E}_{a\neq 0}\,\mathrm{Tr}^2(P_a\mathcal{C}_{o_t}(I))}{\mathrm{Tr}^2(\mathcal{C}_{o_t}(I))} \leq \frac{2^n}{4^n-1}\frac{\mathrm{Tr}\left((\mathcal{C}_{o_t}(I))^2\right)}{\mathrm{Tr}^2(\mathcal{C}_{o_t}(I))} \leq \frac{2^n}{4^n-1}, \tag{A.107}$$

The first and second inequalities are Pauli twirling and monotonicity of l_p norm, respectively. Now we just need to bound $\mathbb{E}_{a\neq 0}\mu_{a,\pm}^2$. Again, by the recurrence relation Eq. (A.93), we have

$$\mathbb{E}_{a\neq 0}\left(\mu'_{a,\pm}\right)^2 = \mathbb{E}_{a\neq 0}\left(\frac{c_{a,0} \pm \varepsilon_0\mu_{a,\pm}c_{a,a}}{c_{0,0} \pm \varepsilon_0\mu_{a,\pm}c_{0,a}}\right)^2 \tag{A.108}$$

$$\leq \frac{\mathbb{E}_{a\neq 0}\left(c_{a,0} \pm \varepsilon_0\mu_{a,\pm}c_{a,a}\right)^2}{(1-\varepsilon_0)^2c_{0,0}^2} \tag{A.109}$$

$$\leq \frac{2\mathbb{E}_{a\neq 0}c_{a,0}^2 + 2\varepsilon_0^2\mathbb{E}_{a\neq 0}\mu_{a,\pm}^2c_{a,a}^2}{(1-\varepsilon_0)^2c_{0,0}^2} \tag{A.110}$$

$$\leq \frac{2}{(1-\varepsilon_0)^2}\frac{2^n}{4^n-1} + \frac{2\varepsilon_0^2}{(1-\varepsilon_0)^2}\mathbb{E}_{a\neq 0}\mu_{a,\pm}^2. \tag{A.111}$$

We also have the initial condition (note that ρ_{init} has no dependence on a)

$$\mathbb{E}_{a\neq 0}\,\mathrm{Tr}^2[P_a\rho_{\mathrm{init}}] \leq \frac{2^n}{4^n-1} < \frac{2}{(1-\varepsilon_0)^2}\frac{2^n}{4^n-1}. \tag{A.112}$$

By induction, we have the following bound,

$$\mathbb{E}_{a\neq 0}\operatorname{Tr}^2[P_a\rho_{a,\pm}^{(\boldsymbol{o}_{<t})}] \equiv \mathbb{E}_{a\neq 0}(\mu_{a,\pm}^{(\boldsymbol{o}_{<t})})^2 \leq \frac{1-\left(\frac{2\varepsilon_0^2}{(1-\varepsilon_0)^2}\right)^{t-1}}{1-\frac{2\varepsilon_0^2}{(1-\varepsilon_0)^2}}\frac{2\cdot 2^n}{(1-\varepsilon_0)^2(4^n-1)} \tag{A.113}$$

$$\leq \frac{2}{1-2\varepsilon_0-\varepsilon_0^2}\cdot\frac{2^n}{4^n-1}, \tag{A.114}$$

which holds since $2\varepsilon_0^2/(1-\varepsilon_0)^2 \leq 1$ for $\varepsilon_0 \leq 1/3$. Substitute this back to Eq. (A.106),

$$\mathbb{E}_{a\neq 0}\operatorname{Tr}^2(P_a\Delta_a^{(\boldsymbol{o}_{<t})}) \leq \frac{1}{2}(\frac{2\varepsilon_0}{1-\varepsilon_0^2})^2\frac{2^n}{4^n-1} + \frac{1}{2}(\frac{\varepsilon_0}{1-\varepsilon_0})^2\frac{8}{1-2\varepsilon_0-\varepsilon_0^2}\cdot\frac{2^n}{4^n-1} \tag{A.115}$$

$$= \frac{1}{2}\left((\frac{2}{1-\varepsilon_0^2})^2 + \frac{1}{(1-\varepsilon_0)^2}\frac{8}{1-2\varepsilon_0-\varepsilon_0^2}\right)\cdot\frac{\varepsilon_0^2 2^n}{4^n-1} \tag{A.116}$$

$$=: f(\varepsilon_0)\cdot\frac{\varepsilon_0^2 2^n}{4^n-1}, \tag{A.117}$$

where $f(\varepsilon_0) = O(1)$ for $\varepsilon_0 \leq 1/3$. Substitute this back to Eq. (A.89) and Eq. (A.84), we got the following TVD bound

$$\begin{aligned}\mathbb{E}_{a\neq 0}\mathrm{TVD}(p_0, \mathbb{E}_{s=\pm 1}[p_{a,s}]) &\leq \sum_{\boldsymbol{o}} p_0(\boldsymbol{o})\sum_{t=1}^{N}\left(\varepsilon_0^2\frac{2^n}{4^n-1} + 2\varepsilon_0^2\frac{2^n}{4^n-1}\sqrt{f(\varepsilon_0)}\right)\\ &= N\left(\varepsilon_0^2\frac{2^n}{4^n-1} + 2\varepsilon_0^2\frac{2^n}{4^n-1}\sqrt{f(\varepsilon_0)}\right),\end{aligned} \tag{A.118}$$

which, combined with Eq. (A.63), gives the following sample complexity lower bound

$$N \geq \frac{1}{3(1+2\sqrt{f(\varepsilon_0)})}\frac{4^n-1}{2^n\varepsilon_0^2} = \frac{1}{12(1+2\sqrt{f(2\varepsilon)})}\frac{4^n-1}{2^n\varepsilon^2}. \tag{A.119}$$

Finally, note that $f(2\varepsilon) < 44$ for $\varepsilon \in [0, 1/6]$, thus we have

$$N \geq 0.005\cdot\frac{4^n-1}{2^n\varepsilon^2} = \Omega(2^n/\varepsilon^2). \tag{A.120}$$

Now we explain how this can be strengthened to a lower bound for N_{meas} (i.e., the number of measurements). Suppose the quantum instrument $\mathcal{C}^{\boldsymbol{o}<t}$ is trivial in the sense of Definition 1.7. Then, the effective POVM elements $E_{o_t}^{\boldsymbol{o}<t}$ for all o_t will be proportional to I, according to the discussion above Eq. (A.65). Now, for any fixed $\boldsymbol{o}$ in Eq. (A.84), the t-th term in the sum will contribute 0 if $E_{o_t}^{\boldsymbol{o}<t}$ is proportional to I, as $\mathrm{Tr}(I \cdot P_a) = 0$ for all $a \neq 0$. Therefore, there are at most N_{meas} non-zero terms in the sum, and we have a uniform upper bound for each of them. The upper bound on TVD, Eq. (A.118), can thus be strengthened to

$$\mathbb{E}_{a\neq 0}\mathrm{TVD}(p_0, \mathbb{E}_{s=\pm 1}[p_{a,s}]) \leq N_{\text{meas}} \left(\varepsilon_0^2 \frac{2^n}{4^n - 1} + 2\varepsilon_0^2 \frac{2^n}{4^n - 1} \sqrt{f(\varepsilon_0)} \right) \tag{A.121}$$

Therefore, the same lower bound for N holds for N_{meas}. This completes the proof of Theorem 1.5. □

APPENDIX B

APPENDICES TO CHAPTER 4

B.1 Notations and Preliminaries

This section provides some basic definitions and preliminaries. For more symbols and notations defined later in the paper, see Table B.1.

Bit string and index set. For a positive integer n, define $[n] \coloneqq \{i \in \mathbb{Z} : 1 \leq i \leq n\}$. We will often use the set of n-bit strings $\mathbb{Z}_2^n$ and the collection of index sets $2^{[n]} \coloneqq \{\nu : \nu \subseteq [n]\}$. Define the following bijection between $\mathbb{Z}_2^n$ and $2^{[n]}$: $s \in \mathbb{Z}_2^n \leftrightarrow \{j \in [n] : s_j = 1\} \in 2^{[n]}$. Throughout the paper, we often do not distinguish a bit string with an index set, as they can always be identified with each other via the aforementioned bijection. We use $|s|$ to denote the Hamming weight of a bit string s (or its cardinality if understood as an index set). We use 0_n to denote the all-zero bit string (which is $\varnothing$ if understood as an index set). We use 1_j to denote the bit string with its jth entry being 1 and other entries being 0 (which is $\{j\}$ if understood as an index set).

We will often use an index set (or, equivalently, a bit string) as the subscript for a string of n objects (e.g., another n-bit string or an n-qubit Pauli operator) to denote a subsequence from the selected indexes. For example, $XYZIZ_{\{1,3\}} = XYZIZ_{10100} = XZ$.

Pauli group. We follow the notations of [FW20, CZSJ22]. Let $\mathsf{P}^n = \{I, X, Y, Z\}^{\otimes n}$ be the Pauli group modulo a global phase. P^n is an Abelian group isomorphic to $\mathbb{Z}_2^{2n}$. Specifically, given a $2n$-bit string $a = (a_x, a_z) = a_{x,1} \cdots a_{x,n} a_{z,1} \cdots a_{z,n}$, the corresponding Pauli operator is given by

$$P_a = \bigotimes_{j=1}^{n} i^{a_{x,j} a_{z,j}} X^{a_{x,j}} Z^{a_{z,j}}, \tag{B.1}$$

Symbol	Definition or meaning	Reference
X	Parameter space	Def. 2.2
$\boldsymbol{e}_i$	Standard basis for parameter space X	Def. 2.2
$X^{S/M/G}$	Subspace of state prep./meas./gate parameters	Def. 2.2
L	Learnable space	Lemma 2.5
T	Gauge space	Lemma 2.7
E	Edge space of the pattern transfer graph	Def. 2.10
Z	Cycle space	Def. 2.11
U	Cut space	Def. 2.12
X_R	Reduced parameter space	Def. 2.16
$\boldsymbol{\vartheta}_i$	Standard basis for reduced parameter space X_R	Def. 2.16
$\mathcal{Q}$	Embedding map from reduced to complete model	Def. 2.16
L_R	Reduced learnable space	Def. 2.17
T_R	Reduced gauge space	Def. 2.18
L_R^G	Reduced learnable subspace for gate parameters	Def. 2.21
$\mathfrak{d}_\nu$	Subsystem depolarizing gauge (SDG)	Def. 2.26
$\pi^{S/M/G}$	Projection to space of state prep./meas./gate parameters	Proof of Theorem 2.5
Ω	Factor set of a Markov Random Field	Sec. B.5.1
Ω_*	Subset of all maximal subsets from Ω	Sec. B.5.1
$\Xi_{\mathcal{G}}$	Extended support map	Lemma 2.33

Table B.1: Table of important symbols.

where the phase is chosen to ensure Hermiticity. We further define $P_0 \coloneqq I^{\otimes n} \equiv I_n$ as a shorthand for the n-qubit identity operator.

Besides the standard addition and inner product defined over $\mathbb{Z}_2^{2n}$, we define the sympletic inner product over P^n as

$$\langle a, b\rangle = \sum_{j=1}^{n} (a_{x,j} b_{z,j} + a_{z,j} b_{x,j}) \mod 2. \tag{B.2}$$

One can verify that $\langle a, b\rangle = 0$ if P_a, P_b commute and $\langle a, b\rangle = 1$ otherwise. We sometimes use the label a and the Pauli operator P_a interchangeably when there is no confusion.

Define $\mathrm{pt} : \mathsf{P}^n \mapsto \mathbb{Z}_2^n$ that maps the ith single-qubit Pauli operator to 0 if it is I, or 1 if it is from $\{X, Y, Z\}$. We call $\mathrm{pt}(a)$ the *pattern* of P_a. Alternatively, $\mathrm{pt}(a)$ represents the support of P_a if understood as an index set. The Hamming weight of $\mathrm{pt}(a)$ is called the

weight of P_a and is denoted by $w(a)$ or $|a|$.

A k-qubit Pauli P acting on a subset of qubits specified by an index set ν is denoted by P_ν. Let $W \in \{I, X, Y, Z\}$ be a single-qubit Pauli operator and $\nu \in 2^{[n]}$, we define $W^\nu = \bigotimes_{j\in\nu} W_j \bigotimes_{j\notin\nu} I_j$.

Pauli channel. An n-qubit Pauli channel Λ has the following two equivalent forms

$$\Lambda(\rho) = \sum_{a\in\mathsf{P}^n} p_a P_a \rho P_a = \sum_{b\in\mathsf{P}^n} \frac{1}{2^n} \lambda_b P_b \operatorname{Tr}[P_b \rho]. \tag{B.3}$$

where $\{p_a\}_a$ is called the *Pauli error rates*, while $\{\lambda_b\}_b$ is called the *Pauli eigenvalues* or *Pauli fidelities* (though the latter is potentially misleading since they can in general be negative). These two sets of parameters are related by the Walsh-Hadamard transform [FW20],

$$\lambda_b = \sum_{a\in\mathsf{P}^n} p_a (-1)^{\langle a,b\rangle}, \quad p_a = \frac{1}{4^n} \sum_{b\in\mathsf{P}^n} \lambda_b (-1)^{\langle a,b\rangle}. \tag{B.4}$$

Note that for Λ to be a quantum channel, the trace-preserving condition is equivalent to $\sum_a p_a \equiv \lambda_0 = 1$, while the complete positivity is equivalent to $p_a \geq 0, \forall a$. For Λ that has the form of Eq. (B.3) but does not satisfy these two conditions, we refer to it as a *Pauli diagonal map*.

We will also use the Pauli-Liouville representation. A Hermitian operator O acting on a 2^n-dimensional Hilbert space can be vectorized as a 4^n-dimensional real vector. We denote this vectorization of O as $|O\rangle\!\rangle$ and the corresponding Hermitian conjugate as $\langle\!\langle O|$, and define the inner product as $\langle\!\langle A|B\rangle\!\rangle \coloneqq \operatorname{Tr}(A^\dagger B)$, which is the Hilbert-Schmidt inner product in the original space. We also introduce the *normalized Pauli operators* $\{\sigma_a \coloneqq P_a/\sqrt{2^n},\ a \in \mathbb{Z}_2^{2n}\}$, whose vectorization $\{|\sigma_a\rangle\!\rangle\}_a$ forms an orthonormal basis for the vectorized space. The action of a quantum channel $\mathcal{E}$ becomes matrix multiplication in this representation, i.e, $|\mathcal{E}(\rho)\rangle\!\rangle = \mathcal{E}^{\mathrm{PL}}|\rho\rangle\!\rangle$, where $\mathcal{E}^{\mathrm{PL}}$ is the corresponding Pauli-Liouville operator and is usually

simply denoted by $\mathcal{E}$ when there is no confusion. In the Pauli-Liouville representation, a Pauli channel is given by

$$\Lambda = \sum_{a \in \mathbb{Z}_2^{2n}} \lambda_a |\sigma_a\rangle\rangle\langle\langle\sigma_a|.$$

which justifies the name of Pauli eigenvalues for $\{\lambda_a\}$.

Clifford group. The n-qubit Clifford group C^n is the group of n-qubit unitary operators that preserve the Pauli group upon conjugation. Mathematically, $\forall C \in \mathsf{C}^n$, $\forall P_a \in \mathsf{P}^n$, $CP_aC^\dagger \in \mathsf{P}^n \times \{\pm\}$. We often denote the unitary channel of a Clifford gate using the same calligraphic letter, e.g., $\mathcal{C}(\rho) \coloneqq C\rho C^\dagger$. We also define, with slight abuse of notation, the mapping $\mathcal{C} : \mathbb{Z}_2^{2n} \mapsto \mathbb{Z}_2^{2n}$ via $\mathcal{C}(P_a) \equiv CP_aC^\dagger = \pm P_{\mathcal{C}(a)}$. The Pauli-Liouville operator of a Clifford $\mathcal{C}$ has the following form,

$$\mathcal{C} = \sum_a (\pm)|\sigma_{\mathcal{C}(a)}\rangle\rangle\langle\langle\sigma_a|. \tag{B.5}$$

Miscellaneous. For a real linear space X, the dual space of it is denoted by X'. For a linear map $\mathcal{Q}$, its dual map is denoted by $\mathcal{Q}^\mathsf{T}$. The image and kernel of a map $\mathcal{Q}$ are denoted by $\mathrm{Im}\mathcal{Q}$ and $\mathrm{Ker}\mathcal{Q}$, respectively. $\mathbb{1}[X]$ is the indicator function that takes 1 if the statement X is true and 0 elsewise. All log represents natural logarithm.

B.2 Learnability of gate set Pauli noise

B.2.1 Basic definitions

In this chapter, we focus on gate sets consisting of arbitrary single-qubit gates and a set of multi-qubit Clifford gates. We assume time-stationary and Markovian noise that has been twirled into Pauli channels. More formally, we have the following,

Definition 2.1 (Complete Pauli noise model)**.** *Given a set of n-qubit Clifford layers $\mathfrak{G}$,*

consider the following gate set,

1. *Initialization:* $\rho_0 = |0\rangle\langle 0|^{\otimes n} \mapsto \tilde{\rho}_0 = \Lambda^S(\rho_0)$, $\Lambda^S = \sum_{a\in\mathsf{P}^n} \lambda^S_{\mathrm{pt}(a)} |\sigma_a\rangle\!\rangle\langle\!\langle\sigma_a|$.
2. *Measurement:* $\{E_j = |j\rangle\langle j|\} \mapsto \{\tilde{E}_j = \Lambda^M(|j\rangle\langle j|)\}$, $\Lambda^M = \sum_{a\in\mathsf{P}^n} \lambda^M_{\mathrm{pt}(a)} |\sigma_a\rangle\!\rangle\langle\!\langle\sigma_a|$.
3. *Single-qubit layer:* $\bigotimes_{i=1}^n U_i \mapsto \bigotimes_{i=1}^n U_i$, $\forall U_i \in \mathrm{SU}(2)$.
4. *Multi-qubit Clifford layer:* $\mathcal{G} \mapsto \tilde{\mathcal{G}} = \mathcal{G} \circ \Lambda^{\mathcal{G}}$, $\forall \mathcal{G} \in \mathfrak{G}$, $\Lambda^{\mathcal{G}} = \sum_{a\in\mathsf{P}^n} \lambda^{\mathcal{G}}_a |\sigma_a\rangle\!\rangle\langle\!\langle\sigma_a|$.

Furthermore, we assume all $\lambda^S_{\mathrm{pt}(a)}, \lambda^M_{\mathrm{pt}(a)}, \lambda^{\mathcal{G}}_a$ *are strictly positive.*

In reality, such a model can always be imposed by applying appropriate Pauli twirling sequence, a technique known as randomized compiling [WE16, HNM+21]. The assumption that single-qubit layer being noiseless can be relaxed to that they have gate-independent noise, in which case such noise can be thought as absorbed into the multi-qubit layer. Note that assuming SPAM noise channel to be a generalized depolarizing channels (i.e., λ_a only depends on $\mathrm{pt}(a)$) is without loss of generality as they only act on Z basis eigenstates.

Definition 2.2 (Parameter space)**.** *Let* $x_a^{S/M/\mathcal{G}} \coloneqq -\log \lambda_a^{S/M/\mathcal{G}}$ *for all* $a \neq \mathbf{0}$ *for SPAM parameters and all* $a \neq I_n$ *for all gates* $\mathcal{G}$*'s parameters. Let* $\boldsymbol{x}$ *be a formal vector defined as*

$$\boldsymbol{x} \coloneqq \sum_{u\neq\mathbf{0}} x_u^S \boldsymbol{e}_u^S + \sum_{u\neq\mathbf{0}} x_u^M \boldsymbol{e}_u^M + \sum_{\mathcal{G}\in\mathfrak{G}} \sum_{a\neq I_n} x_a^{\mathcal{G}} \boldsymbol{e}_a^{\mathcal{G}} \tag{B.6}$$

where $\{\boldsymbol{e}_u^S, \boldsymbol{e}_u^M, \boldsymbol{e}_a^{\mathcal{G}} : u \neq 0_n, a \neq I_n, \mathcal{G} \in \mathfrak{G}\}$ *forms an orthonormal basis for a real linear space* X*, called the parameter space. Furthermore, define* $X^{S/M} \coloneqq \mathrm{span}\{\boldsymbol{e}_u^{S/M} : u \neq 0_n\}$, $X^G \coloneqq \mathrm{span}\{\boldsymbol{e}_a^{\mathcal{G}} : a \neq I_n, \mathcal{G} \in \mathfrak{G}\}$.

Note that, we do not include the Pauli fidelity corresponding to identity, as they are always 1 for the channels to be trace-preserving. Clearly, any realization of the noise parameters uniquely corresponds to a vector $\boldsymbol{x}$ in X, but not all $\boldsymbol{x} \in X$ corresponds to a physical noise model as the channels might not be completely positive.

Definition 2.3 (Experiment)**.** *An experiment is a map $F : X \mapsto \mathbb{R}^{2^n}$ specified by a quantum circuit $\mathcal{C}$ (which is a sequence of finitely many single-qubit layers and multi-qubit Clifford layers within $\mathfrak{G}$), defined as $F(\boldsymbol{x})_j = \mathrm{Tr}[\widetilde{E}_j \widetilde{\mathcal{C}}(\widetilde{\rho}_0)]$. Here the noisy initial state $\widetilde{\rho}_0$, noisy circuit $\widetilde{\mathcal{C}}$, and noisy measurement, $\widetilde{E_j}$ depends on $\boldsymbol{x}$.*

When $\boldsymbol{x}$ represents a physical noise model, any experiment $F(\boldsymbol{x})$ yields a Born's rule probability distribution. However, for mathematical convenience, we naturally extend the domain of F to all over X.

Definition 2.4 (Learnable function)**.** *A real-valued function f on X is **learnable** if there exists a set of experiments $\{F_j\}_{j=1}^m$ such that f is a function of $\{F_j\}_{j=1}^m$. That is, there exists some $\hat{f}$ such that $f(\boldsymbol{x}) = \hat{f}(F_1(\boldsymbol{x}), \cdots, F_m(\boldsymbol{x}))$ for all $\boldsymbol{x} \in X$.*

We comment that the learnable function we define here is also known as identifiable function in the literature of identifiability analysis [Hsi83]. Here, for consistency, we will stick to the term "learnable" which has been used in [CLO$^+$23, HFP22].

We will be specifically interested in studying the learnability of linear functions (and we will later explain why that suffices). Note that, there exists an identification between X and its dual space $X' \equiv \mathcal{L}(X, \mathbb{R})$ using the standard inner product, given by $\boldsymbol{x} \leftrightarrow f_{\boldsymbol{x}}(\boldsymbol{y}) \coloneqq \boldsymbol{x} \cdot \boldsymbol{y}$. One can thus think of a linear function on X as a co-vector and denote it with a boldface letter.

Lemma 2.5. *The learnable linear functions form a subspace of X', called the **learnable space**, denoted by L.*

Definition 2.6 (Gauge vector)**.** *A vector $\boldsymbol{y} \in X$ is a **gauge** vector if for any $\boldsymbol{x} \in X$, $F(\boldsymbol{x}) = F(\boldsymbol{x} + \eta \boldsymbol{y})$ for any experiment F and $\eta \in \mathbb{R}$.*

Lemma 2.7. *The gauge vectors form a linear subspace of X called the **gauge space**, denoted by T.*

Proof. By definition, $\boldsymbol{y} \in X$ being a gauge vector implies $\eta\boldsymbol{y}$ being a gauge vector for all $\eta \in \mathbb{R}$. Let $\boldsymbol{y}_1, \boldsymbol{y}_2 \in X$ be any gauge vectors. By definition, for any $\eta \in \mathbb{R}$ and any observation F,

$$F(\boldsymbol{x}) = F(\boldsymbol{x} + \eta\boldsymbol{y}_1) = F(\boldsymbol{x} + \eta\boldsymbol{y}_1 + \eta\boldsymbol{y}_2). \tag{B.7}$$

Therefore, $\boldsymbol{y}_1 + \boldsymbol{y}_2$ is also a gauge vector. This proves the gauge vectors forms a linear subspace. □

Lemma 2.8. *$L \perp T$. That is, for any $\boldsymbol{f} \in L$ and $\boldsymbol{y} \in T$ we have $\boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{f} \cdot \boldsymbol{y} = 0$.*

Proof. $\boldsymbol{y} \in T$ implies $F(\boldsymbol{x}) = F(\boldsymbol{x} + \boldsymbol{y})$ for any experiment F and any $\boldsymbol{x} \in X$. $\boldsymbol{f} \in L$ implies there is a set of experiments $\{F_1, \cdots, F_m\}$ and a function $\hat{f}$ such that $\boldsymbol{f}(\boldsymbol{x}) = \hat{f}(F_1(\boldsymbol{x}), \cdots, F_m(\boldsymbol{x}))$. Therefore,

$$\boldsymbol{f}(\boldsymbol{x} + \boldsymbol{y}) = \hat{f}(\{F_j(\boldsymbol{x} + \boldsymbol{y})\}_{j=1}^m) = \hat{f}(\{F_j(\boldsymbol{x})\}_{j=1}^m) = \boldsymbol{f}(\boldsymbol{x}) \tag{B.8}$$

for any $\boldsymbol{x}$. Thus $\boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{f}(\boldsymbol{x} + \boldsymbol{y}) - \boldsymbol{f}(\boldsymbol{x}) = 0$ by linearity. This completes the proof. □

B.2.2 Learnability of complete noise model

In this section, we study the learnability of the complete Pauli noise model. The results presented here are an extension to those in [CLO$^+$23]. However, we will use a quite different proof technique that allows us to generalize the results to noise model with arbitrary constraints in the next section. The following graph is originally defined in [CLO$^+$23], but we twist the definition a bit to include a "SPAM node".

Definition 2.9 (pattern transfer graph with SPAM node (PTG))**.** *The pattern transfer graph associated with an n-qubit Pauli noise model is a directed graph $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ constructed as follows:*

- V*: There are $2^n - 1$ nodes v_s for all $s \in \{0,1\}^n \backslash \{0^n\}$ and 1 additional root node v_R.*

- E:

  1. *For each* $a \in \mathsf{P}^n \backslash \{I\}$ *and* $\mathcal{G} \in \mathfrak{G}$*, construct an edge* $v_{\mathrm{pt}(P_a)} \to v_{\mathrm{pt}(\mathcal{G}(P_a))}$ *labeled by* $\boldsymbol{e}_a^{\mathcal{G}}$.
  2. *For each* $u \in \{0,1\}^n \backslash \{0^n\}$*, construct an edge* $v_R \to v_u$ *labeled by* $\boldsymbol{e}_u^S$.
  3. *For each* $u \in \{0,1\}^n \backslash \{0^n\}$*, construct an edge* $v_u \to v_R$ *labeled by* $\boldsymbol{e}_u^M$.

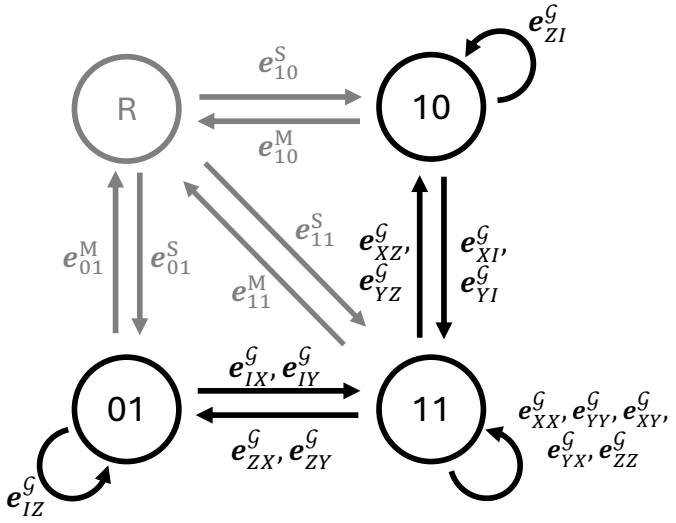


Figure B.1: The PTG for $n = 2$, $\mathfrak{G} = \{\mathrm{CZ}\}$. Here we define $\mathcal{G} = \mathrm{CZ}$ for notational simplicity. The black edges corresponds to gate noise parameters, while the gray edges corresponds to SPAM noise parameters. Multiple edges are shown as one edge with multiple labels.

The following definition of algebraic graph theoretic terminology follows [CLO+23].

Definition 2.10. *The* edge space E *is a real linear space spanned by a standard basis* $\{\boldsymbol{e}\}$ *where each* $\boldsymbol{e}$ *is a formal vector corresponding to an edge* $e \in \mathsf{E}$.

Let the edge space be the real linear space that takes all the edges as a basis, denoted by E. Clearly $E \cong X \cong X'$ up to the natural isomorphism. There are two subspaces of E of special interest, which are defined as follows:

A *path* is an alternating sequence of vertices and edges $z = (v_{i_0}, e_{i_1}, v_{i_1}, e_{i_2}, \cdots, e_{i_q}, v_{i_q})$ such that each edge satisfies $e_{i_k} = (v_{i_{k-1}}, v_{i_k})$. A *cycle* is a closed path, which means

$v_{i_0} = v_{i_q}$.[1] A cycle is *simple* if it passes through each unique edge at most once. For a simple cycle z in the PTG, we assign a vector $\boldsymbol{x}_z \in E$ as follows

$$\boldsymbol{x}_z[e] = \begin{cases} +1, & e \in z. \\ 0, & e \notin z. \end{cases} \tag{B.9}$$

Definition 2.11. *The **cycle space** Z is a linear subspace of E spanned by all simple cycles in* G.

A graph is *strongly-connected* if for any two vertices, there exists a cycle containing both of them. The PTG we considered is strongly-connected by construction, as v_R is directed connected to any other nodes.

Given a partition of vertices $\mathsf{V} = \mathsf{V}_0 \cup \mathsf{V}_0^c$, a *cut* is the set of all edges $e = (u, v)$ such that one of u, v belongs to V_0 and the other belongs to V_0^c. For each cut p we assign a vector $\boldsymbol{y}_p \in E$ as follows

$$\boldsymbol{y}_p(e) = \begin{cases} +1, & e \in p,\ e \text{ goes from } \mathsf{V}_0 \text{ to } \mathsf{V}_0^c. \\ -1, & e \in p,\ e \text{ goes from } \mathsf{V}_0^c \text{ to } \mathsf{V}_0. \\ 0, & e \notin p. \end{cases} \tag{B.10}$$

Definition 2.12. *The **cut space** U is a linear subspace of E spanned by all cuts in* G.

Below is an important fact about the cycle space Z and cut space U.

Lemma 2.13 ([GLS03])**.** *$E = Z \oplus^{\perp} U$. Here $\oplus^{\perp}$ means an orthogonal direct sum.*

The following theorem is a crucial ingredient to understand the learnability of the complete Pauli noise model.

1. In the literature of algebraic graph theory, people sometimes allow reverse edges (i.e., $e_{i_k} = (v_{i_k}, v_{i_{k-1}})$) in the definition of a cycle and refer to our definition as an *oriented cycle* or a *circuit*. Since all graphs considered in this chapter are either strongly-connected or a union of strongly-connected components, and it has been shown that oriented cycles can linearly represent any cycle for such graphs [GLS03], working with our current definition would not cause any inconsistency.

Theorem 2.1 (Linearity of experiments)**.** *Define a cycle to be* ***rooted*** *if it contains* v_R *exactly once.*

(a) For any experiment F*, there exists a finite set of linear functions* $\{\boldsymbol{f}_j\}_{j=1}^M$ *where each* $\boldsymbol{f}_j$ *corresponds to a rooted cycle, such that* F *is a function of* $\{\boldsymbol{f}_j\}$*. That is, there exists a function* $\widehat{F}$ *such that* $F(\boldsymbol{x}) = \widehat{F}(\boldsymbol{f}_1(\boldsymbol{x}), \cdots, \boldsymbol{f}_M(\boldsymbol{x}))$ *for all* $\boldsymbol{x} \in X$*.*

(b) For any linear function $\boldsymbol{f}_j$ *corresponding to a rooted cycle, there exists an experiment* F *such that* $\boldsymbol{f}_j$ *is a function of* F*. That is, there exists a function* $\widehat{f}_j$ *such that* $\boldsymbol{f}_j(\boldsymbol{x}) = \widehat{f}_j(F(\boldsymbol{x}))$ *for all* $\boldsymbol{x} \in X$*.*

This is detailed in [CLO+23]. For completeness, we give a proof in SM Sec. B.7.1. Intuitively, Theorem 2.1(b) says that any rooted cycle can be learned from at least one experiments. In fact, the experiment involves preparing certain noisy Pauli eigenstate, applying a sequence of perfect single-qubit and noisy multi-qubit Clifford gates, and measuring certain noisy Pauli observable. The expectation value will have the form of $\lambda_{b_0}^S \lambda_{P_1}^{\mathcal{G}_1} \cdots \lambda_{P_m}^{\mathcal{G}_m} \lambda_{b_m}^M$, taking a log of which yields the desired rooted cycle; Theorem 2.1(a) says that rooted cycles are complete, in the sense that any experiment (possibly involving non-stabilizer states, measurements, and non-Clifford single-qubit gates) can be predicted using linear combinations of rooted cycles. Therefore, it is sufficient to characterize the learnability of linear functions and not to worry about the non-linear ones.

To understand the full picture of learnable and gauge degrees of freedom, we introduce the following definition

Definition 2.14 (Observation map)**.** *Fix an arbitrary basis* $\{\boldsymbol{f}_i\}_{i=1}^{|Z|}$ *for* Z *such that each* $\boldsymbol{f}_i$ *represents a rooted cycle. Define the* ***observation*** *map* $\mathcal{P} : X \mapsto \mathbb{R}^{|Z|}$ *via*

$$\mathcal{P}(\boldsymbol{x}) = [\boldsymbol{f}_1(\boldsymbol{x}), \boldsymbol{f}_2(\boldsymbol{x}), \cdots, \boldsymbol{f}_{|Z|}(\boldsymbol{x})]^{\mathrm{T}}. \tag{B.11}$$

Note that, thanks to the strong connectivity of the PTG, a cycle basis consisting of only rooted cycles always exist. One construction is given below.

Lemma 2.15 (A rooted cycle basis)**.** *The following set of vectors forms a basis for Z,*

$$\mathsf{B} \coloneqq \{\boldsymbol{e}_u^S + \boldsymbol{e}_u^M : u \neq 0_n\} \cup \{\boldsymbol{e}_{\mathrm{pt}(a)}^S + \boldsymbol{e}_a^{\mathcal{G}} + \boldsymbol{e}_{\mathrm{pt}(\mathcal{G}(a))}^M : \mathcal{G} \in \mathfrak{G}, a \neq I_n\}. \tag{B.12}$$

Proof. To begin with, every vector in B is obviously a rooted cycle. Since the PTG is strongly connected by construction, the dimension of T is $2^n - 1$ [GLS03], thus

$$|Z| = |E| - |T| = 2^n - 1 + |\mathfrak{G}|(4^n - 1), \tag{B.13}$$

which is the same as the cardinality of B. Therefore, it suffices to show the linear independence for B. Consider the equation $\sum_{i:\boldsymbol{e}_i \in \mathsf{B}} \alpha_i \boldsymbol{e}_i = \boldsymbol{0}$. Note that each $\boldsymbol{e}_a^{\mathcal{G}}$ appears in exactly one vector from the second part of B, for which the corresponding α_i must be all zero, but the first part is also obviously linearly independent, which means all α_i must be zero, proving the linear independence. □

Theorem 2.2. *Up to identification between X and X' using the standard basis,*

$$\begin{array}{ccc} \mathrm{Im}\mathcal{P}^{\mathsf{T}} & & \mathrm{Ker}\mathcal{P} \\ \| & & \| \\ X = L & \oplus^{\perp} & T \\ \| & \| & \| \\ E = Z & \oplus^{\perp} & U \end{array} . \tag{B.14}$$

In words, the cycle space Z is equal to the learnable space L, while the cut space U is equal to the gauge space T. Besides, the two spaces are orthogonal complement to each other. Finally, the learnable space L is also equal to the image for the dual of the observation map

$\mathcal{P}$, *while the gauge space* T *is the kernel of* $\mathcal{P}$.

Remark: while $E = Z \oplus^{\perp} U$ is a graph-theoretic fact, we don't know a priori whether $X = L \oplus T$ holds under our definition.

Proof. We first show that $T = \mathrm{Ker}(\mathcal{P})$, *i.e.,* $\boldsymbol{y} \in T \Leftrightarrow \mathcal{P}(\boldsymbol{y}) = \mathbf{0}$. Indeed, Lemma 2.1(a) implies that $F(\boldsymbol{x}) = \widehat{F}(\mathcal{P}(\boldsymbol{x}))$ for some function $\widehat{F}$. Thus, for any $\boldsymbol{y} \in \mathrm{Ker}(\mathcal{P})$ and $\eta \in \mathbb{R}$,

$$F(\boldsymbol{x} + \eta\boldsymbol{y}) = \widehat{F}(\mathcal{P}(\boldsymbol{x} + \eta\boldsymbol{y})) = \widehat{F}(\mathcal{P}(\boldsymbol{x})) = F(\boldsymbol{x}). \tag{B.15}$$

The second equality uses the linearity of $\mathcal{P}$. Therefore, $\mathrm{Ker}(\mathcal{P}) \subseteq T$. On the other hand, for any $\boldsymbol{y} \notin \mathrm{Ker}(\mathcal{P})$, there is at least one $\boldsymbol{f}_i$ such that $\boldsymbol{f}_i(\boldsymbol{y}) \neq 0$. Now, Lemma 2.1(b) implies that there exists an experiment F_i and a function $\hat{f}_i$ such that $\boldsymbol{f}_i(\boldsymbol{x}) = \hat{f}_i(F_i(\boldsymbol{x}))$ for all $\boldsymbol{x}$. Thus,

$$\begin{aligned} \boldsymbol{f}_i(\boldsymbol{y}) \neq 0 &\Rightarrow \boldsymbol{f}_i(\boldsymbol{x}) \neq \boldsymbol{f}_i(\boldsymbol{x} + \boldsymbol{y}) \\ &\Rightarrow \hat{f}_i(F(\boldsymbol{x})) \neq \hat{f}_i(F(\boldsymbol{x} + \boldsymbol{y})) \\ &\Rightarrow F(\boldsymbol{x}) \neq F(\boldsymbol{x} + \boldsymbol{y}) \\ &\Rightarrow \boldsymbol{y} \notin T. \end{aligned} \tag{B.16}$$

Therefore, $T \subseteq \mathrm{Ker}\mathcal{P}$. This proves the claim that $T = \mathrm{Ker}P$.

We then show that $L = \mathrm{Im}(\mathcal{P}^{\mathsf{T}})$. For any $\boldsymbol{f} \in \mathrm{Im}(\mathcal{P}^{\mathsf{T}})$, there is some $\boldsymbol{h} \in \mathcal{L}(\mathbb{R}^{|Z|}, \mathbb{R})$ such that $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{h}(\mathcal{P}(\boldsymbol{x})),\ \forall \boldsymbol{x} \in X$. Lemma 2.1(b) implies that each entry of $\mathcal{P}(\boldsymbol{x})$ is learnable, thus $\boldsymbol{f}$ is learnable. This means $\mathrm{Im}(\mathcal{P}^{\mathsf{T}}) \subseteq L$; On the other hand, $L \perp T = \mathrm{Ker}(\mathcal{P}) \Rightarrow L \subseteq \mathrm{Im}(\mathcal{P}^{\mathsf{T}})$ as the kernel is the orthogonal complement of the adjoint image. This proves the claim that $L = \mathrm{Im}(\mathcal{P}^{\mathsf{T}})$.

Finally, it is not hard to see that $\mathrm{Ker}(\mathcal{P}) = U$ thanks to the duality between cycle space and cut space [GLS03], and hence $\mathrm{Im}(\mathcal{P}^{\mathsf{T}}) = Z$. This completes the proof of Theorem 2.2. □

B.2.3 Learnability of reduced noise model

Since the complete noise model contains exponentially many parameters, we often need to make additional assumptions in practice to reduce its complexity. Here, we want to understand the learnability of such reduced noise model.

Definition 2.16. *A reduced Pauli noise model is defined by a tuple $(X_R, \mathcal{Q})$, where*

- *X_R is a real linear space called the **reduced parameter space**. A realization of the reduced noise parameters $\{r_i\}_{i=1}^{|X_R|}$ corresponds to a vector $\boldsymbol{r} \in X_R$ via $\boldsymbol{r} = \sum_i r_i \boldsymbol{\vartheta}_i$ where $\{\boldsymbol{\vartheta}_i\}$ is a standard basis for X_R.*
- *$\mathcal{Q} : X_R \mapsto X$ is an injective linear map called the **embedding map** such that for any $\boldsymbol{r}$, the corresponding realization of the complete noise parameters is given by $\boldsymbol{x} \coloneqq \mathcal{Q}(\boldsymbol{r})$.*

Note that we require $\mathcal{Q}$ to be linear and injective. We will see in the following sections that many physically-motivated reduced noise models (e.g., spatially local noise) can indeed be described by a linear embedding map. The requirement of injectivity, equivalent to $\mathrm{Ker}(\mathcal{Q}) = \{\mathbf{0}\}$, is for mathematical convenience. In practice, if the embedding map is not injective, one can always construct another reduced model with fewer parameters and admits an injective embedding map.

Any experiment F as in Definition 2.3 induces a function $F \circ \mathcal{Q}$ acting on X_R. The definitions and properties of learnable functions and gauge vectors then carry over to the reduced model, by replacing X by X_R in Definitions 2.4, 2.6. Formally, we have

Definition 2.17 (Reduced learnable space)**.** *A real-valued function f on X_R is **learnable** if there exists a finite set of experiments $\{F_j\}_{j=1}^m$ such that f is a function of $\{F_j \circ \mathcal{Q}\}_{j=1}^m$. That is, there exists some $\hat{f}$ such that $f(\boldsymbol{r}) = \hat{f}(F_1(\mathcal{Q}(\boldsymbol{r})), \cdots, F_m(\mathcal{Q}(\boldsymbol{r})))$ for all $\boldsymbol{r} \in X_R$. The linear subspace of X_R' spanned by all linear learnable functions is called the **reduced learnable space**, denoted by L_R.*

Definition 2.18 (Reduced gauge space)**.** *A vector $\boldsymbol{y} \in X_R$ is a* ***gauge*** *vector if for any $\boldsymbol{r} \in X_R$, $F(\mathcal{Q}(\boldsymbol{r})) = F(\mathcal{Q}(\boldsymbol{r} + \eta\boldsymbol{y}))$ for any experiment F and $\eta \in \mathbb{R}$. The linear subspace of X_R spanned by all the gauge vectors is called the* ***reduced gauge space****, denoted by T_R.*

The characterization of learnability naturally extends to the reduced model, as stated in the next theorem.

Theorem 2.3. *$L_R = \mathrm{Im}(\mathcal{Q}^{\mathsf{T}} \circ \mathcal{P}^{\mathsf{T}})$, $T_R = \mathrm{Ker}(\mathcal{P} \circ \mathcal{Q})$.*

We provide a proof for Theorem 2.3 in Sec. B.7.2. Basically, all steps in the proof of Theorem 2.2 naturally carry over. The following corollary characterizes the relation of the learnable/gauge spaces between the complete and the reduced noise model.

Corollary 2.19. *$L_R = \mathcal{Q}^{\mathsf{T}}(L)$, $T_R = \mathcal{Q}^{-1}(T)$.*

Here $\mathcal{Q}^{-1}(T) \coloneqq \{\boldsymbol{x} \in X_R \mid \mathcal{Q}(x) \in T\}$ is the preimage of T under $\mathcal{Q}$. Alternatively, $T_R = \mathcal{Q}^{-1}(T \cap \mathrm{Im}(\mathcal{Q}))$ where $\mathcal{Q}^{-1}$ is viewed as the inverse of Q restricted to its image (recall that $\mathcal{Q}$ is assumed to be injective).

Proof. For any $\boldsymbol{f} \in X_R'$, $\boldsymbol{f} \in L_R \Leftrightarrow \exists \boldsymbol{g} \in \mathcal{L}(\mathbb{R}^{|Z|}, \mathbb{R})$ *s.t.* $\boldsymbol{f} = \mathcal{Q}^{\mathsf{T}}(\mathcal{P}^{\mathsf{T}}(g)) \Leftrightarrow \boldsymbol{f} \in \mathcal{Q}^{\mathsf{T}}(L)$. Thus, $L_R = \mathcal{Q}^{\mathsf{T}}(L)$.

Let us consider $\mathcal{Q}(T_R)$. Theorems 2.2, 2.3 imply that $\mathcal{Q}(T_R) \subseteq T$. Also, $\mathcal{Q}(T_R) \subseteq \mathrm{Im}(Q)$ trivially holds. On the other hand, for any $\boldsymbol{y} \in T \cap \mathrm{Im}(\mathcal{Q})$, $\mathcal{P}(\boldsymbol{y}) = 0$, and there exists some $\boldsymbol{x} \in X_R$ such that $\boldsymbol{y} = \mathcal{Q}(\boldsymbol{x})$, but this implies $\mathcal{P}(\mathcal{Q}(\boldsymbol{x})) = 0$ and thus $\boldsymbol{x} \in T_R$, $\boldsymbol{y} \in \mathcal{Q}(T_R)$. We conclude that $\mathcal{Q}(T_R) = T \cap \mathrm{Im}(\mathcal{Q})$. Finally, note that $\mathcal{Q}$ is injective by assumption and is thus invertible when restricted to its image. We conclude that $T_R = \mathcal{Q}^{-1}(T \cap \mathrm{Im}(\mathcal{Q}))$. □

B.3 Algorithms for learning gate set Pauli noise

We have established a theory for characterizing the learnable/gauge degrees of freedom for any complete or reduced Pauli noise models. The next question to be asked is how to

design experiments to learn the learnable degrees of freedom, and how efficiently one can learn them. In this section, we first introduce a general algorithm for learning Pauli noise models, which leads to efficient experimental designs. We then discuss the issue of learning noise parameters to additive/multiplicative precision. The algorithms discussed here will be applied in the next section to address many practically interesting examples of Pauli noise models.

From this section, we focus on Clifford circuits as they are sufficient for learning Pauli noise. Also, instead of describing an experiment with its measurement probability distribution, we will talk about expectation values of measuring certain Pauli observable. Since we assume noiseless single-qubit layers, measuring Pauli P_a effectively picks up $\lambda^M_{\mathrm{pt}(a)}$ from the measurement noise channel. We thus define the effective noisy observable as $\tilde{P}_a \coloneqq \lambda^M_{\mathrm{pt}(a)} P_a$. Similarly, we use ρ_a to denote the $+1$ eigenstate of P_a obtained by applying an appropriate single-qubit Clifford layer to $\rho_0 \equiv |0\rangle\langle 0|^{\otimes n}$.[2] The noisy realization of ρ_a is denoted as $\tilde{\rho}_a$ and satisfies $\mathrm{Tr}[P_a \tilde{\rho}_a] = \lambda^S_{\mathrm{pt}(a)}$. Finally, we use the tuple $(\rho_a, \mathcal{C}, P_b)$ to denote an experiment that (ideally) starts with ρ_a, applies the sequence of gates $\mathcal{C}$, and measures the observable P_b.

B.3.1 Basic definitions and vanilla algorithm

Let us first define the task of learning a Pauli noise model. Given a reduced noise model described by an embedding map $\mathcal{Q}$. Recall that L_R describes the linear space for all learnable functions about the noise model (and that Theorem 2.1 says *linear* learnable functions suffice). Thus, we just need to do the following:

1. Specify a basis for L_R, denoted as $\{\boldsymbol{f}_j\}_{j=1}^{|L_R|}$,

2. Although ρ_a is not uniquely defined when P_a contains identity, choosing any ρ_a that satisfies this condition will make no observable difference.

2. Specify the experiments to learn each $\boldsymbol{f}_j$.

By conducting those experiments, we will have effectively learned every function in L_R, which are all we can learn about the noise model.

However, for an arbitrary basis vector $\boldsymbol{f}_j \in L_R$, it is not straightforward to find the experiments that learn it, despite their existence. We would like to find a reduced basis that is more closely related to experiments. Recall that for the complete model, the learnable space L corresponds to the cycle space Z of PTG, and each rooted cycle can be extracted from one experiment (Theorem 2.1(b)). This motivates the following definition.

Definition 2.20 (Reduced cycle)**.** *For any $\boldsymbol{f} \in L_R$, if there exists a $\boldsymbol{g} \in L$ that is a cycle and that $\boldsymbol{f} = \mathcal{Q}^{\mathsf{T}}(\boldsymbol{g})$, we call $\boldsymbol{f}$ a **reduced cycle** corresponding to $\boldsymbol{g}$. If $\boldsymbol{g}$ is also a rooted cycle, we call $\boldsymbol{f}$ a **rooted reduced cycle**.*

We remind the reader that not all $\boldsymbol{g} \in L$ represents a cycle, as some might only be linear combinations of cycles. Since Corollary 2.19 states that $L_R = \mathcal{Q}^{\mathsf{T}}(L)$, given a rooted cycle basis $\{\boldsymbol{g}_i\}_{i=1}^{|L|}$ for L, we know that $\{\mathcal{Q}^{\mathsf{T}}(\boldsymbol{g}_i)\}_{i=1}^{|L|}$ spans L_R, thus we can pick a subset $\{\boldsymbol{f}_j\}_{j=1}^{|L_R|} \subseteq \{\mathcal{Q}^{\mathsf{T}}(\boldsymbol{g}_i)\}_i$ as a basis for L_R consisting solely of rooted reduced cycles. Theorem 2.1(b) thus immediately gives an experiment to learn $\boldsymbol{f}_j$.

As a concrete example, consider the rooted cycle basis introduced in Lemma 2.15,

$$\mathsf{B} \coloneqq \{\boldsymbol{e}_u^S + \boldsymbol{e}_u^M : u \neq \boldsymbol{0}\} \cup \{\boldsymbol{e}_{\mathrm{pt}(a)}^S + \boldsymbol{e}_a^{\mathcal{G}} + \boldsymbol{e}_{\mathrm{pt}(\mathcal{G}(a))}^M : \mathcal{G} \in \mathfrak{G}, a \neq I\}. \tag{B.17}$$

We have the following procedure for noise learning.

Algorithm 1 [A simple algorithm for learning gate set Pauli noise].

1. Select a subset $\mathsf{B}_R \subseteq \mathsf{B}$ such that $\mathcal{Q}^{\mathsf{T}}(\mathsf{B}_R)$ forms a basis for L_R.

2. Learn all $\boldsymbol{e}_u^S + \boldsymbol{e}_u^M \in \mathsf{B}_R$ by preparing $\tilde{\rho}_0$ and measuring $\tilde{Z}^u$.

3. Learn all $\boldsymbol{e}^S_{\mathrm{pt}(a)} + \boldsymbol{e}^{\mathcal{G}}_a + \boldsymbol{e}^M_{\mathrm{pt}(\mathcal{G}(a))} \in \mathsf{B}_R$ by preparing $\tilde{\rho}_a$, applying $\tilde{\mathcal{G}}$, and measuring $\tilde{P}_{\mathcal{G}(a)}$.

As an obvious challenge to this approach, the cardinality of B is $2^n - 1 + |\mathfrak{G}|(4^n - 1)$, scaling exponentially with n. Even if the dimension of L_R is polynomial in n, Step 1 of picking a basis for L_R still requires exponential computational complexity in general. However, for physically-motivated noise ansatzs such as spatially local noise, it is often easy to efficiently construct such a basis. This will be discussed in details when we turn to concrete noise models. Conditioned on that B_R can be efficiently constructed, step 2 and 3 involves at most $|L_R|$ experiments to run, which is polynomial in n for any feasible reduced noise model. One might notice that this algorithm resembles standard quantum process tomography [MRL08]. Specifically, each gate from the gate set is learned one by one. No concatenation or composition of different gates is used.

B.3.2 On learning gate noise to relative precision

In practice, there are often considerations beyond merely constructing a complete set of experiments. One important factor to consider is the precision-efficiency trade-off in estimating the noise parameters. Many existing RB-type noise characterization protocols can learn gate noise parameter to relative precision (with respect to the infidelity) using a small number of initialization and measurements [FW20, HHF$^+$19, EWP$^+$19]. More precisely, suppose we are interested in learning certain gate fidelity parameter λ, RB-type protocol can generally construct a family of experiments $\{(\rho_0, \mathcal{C}^d, P)\}_d$ such that the circuit fidelities satisfy

$$f_d \coloneqq \mathrm{Tr}[\tilde{P}\tilde{\mathcal{C}}^d(\tilde{\rho}_0)] = \alpha\lambda^d, \tag{B.18}$$

where α represents some SPAM fidelity, d is the number of concatenation of the relevant gates. We assume both α and λ are sufficiently close to 1. Now choose $d \in \{0, m\}$, estimate

each $\widehat{f}_d \coloneqq f_d + \hat{\varepsilon}_m$, and construct the following ratio estimator,

$$\hat{\lambda} \coloneqq \frac{\widehat{f}_m}{\widehat{f}_0} = \left(\frac{\alpha\lambda^m + \hat{\varepsilon}_m}{\alpha + \hat{\varepsilon}_0}\right)^{\frac{1}{m}} \approx \lambda\left(1 + \frac{\hat{\varepsilon}_m\lambda^{-m} - \hat{\varepsilon}_0}{m\alpha}\right). \tag{B.19}$$

The last approximation holds given that $|\hat{\varepsilon}_0/\alpha|, |(\hat{\varepsilon}_m\lambda^{-m} - \hat{\varepsilon}_0)/m\alpha| \ll 1$. If we choose $1/m \approx \log\lambda^{-1} = 1 - \lambda + o(1-\lambda)$, we would obtain that

$$\hat{\lambda} \approx \lambda\left(1 + (1-\lambda)\frac{\hat{\varepsilon}_m e - \hat{\varepsilon}_0}{\alpha}\right). \tag{B.20}$$

This means as long as we estimate f_0, f_m to $\pm\varepsilon$ additive precision, we would obtain $O(\varepsilon)$ relative precision estimation for λ with respect to the infidelity $1-\lambda$. As a side note, one can efficiently find the required m using a search subroutine described in [HHF$^+$19, Sec. V]. The basic idea is to estimate f_d for d's chosen from an exponentially increasing list, and output $m = d$ when $\widehat{f}_d/\widehat{f}_0$ is below certain threshold.

The above analysis shows that, as long as one can construct a family of experiments as in Eq. (B.18), one can amplify λ by concatenation and learn it to relative precision. We now explain how to fit this into the framework of gate set Pauli noise learning.

Definition 2.21. *For any reduced Pauli noise model $(X_R, \mathcal{Q})$, the reduced learnable space for gate parameters is defined as $L_R^G \coloneqq \mathcal{Q}^{\mathsf{T}}(L \cap X^G)$.*

Theorem 2.4 (Noise parameter amplification)**.** *Given a reduced model $(X_R, \mathcal{Q})$, for any reduced cycle $\boldsymbol{f} \in L_R^G$, there exists a family of experiments $\{(P, \mathcal{C}^m, \rho_P)\}_m$ and a rooted cycle $\boldsymbol{\alpha} \in L$ such that*

$$\mathrm{Tr}[\tilde{P}\tilde{\mathcal{C}}^m(\tilde{\rho}_P)] = \exp(-\boldsymbol{\alpha}(\mathcal{Q}(\boldsymbol{r})))\exp(-m\boldsymbol{f}(\boldsymbol{r})), \quad \forall m \in \mathbb{N}. \tag{B.21}$$

In other words, any functions on the reduced model that corresponds to learnable functions consisting solely of gate noise parameters in the complete model can be amplified via con-

catenation, and can thus be learned to relative precision. Here, the gate sequence $\mathcal{C}$ can be a composition of multiple gates from the gate set interleaved by noiseless single-qubit gates, which is similar to the concept of "germs" in gate set tomography [NGR+21], and also appears in other Pauli noise learning protocols [CLO+23, VDBMKT23, CDDH+23].

Proof. By definition, there exists a cycle $\boldsymbol{g} \in L \cap X^G$ such that $\boldsymbol{f} = \mathcal{Q}^{\mathsf{T}}(\boldsymbol{g})$. Let u be any vertex of the PTG that is contained in $\boldsymbol{g}$. Then $\boldsymbol{\alpha} \coloneqq \boldsymbol{e}_u^S + \boldsymbol{e}_u^M$ is a rooted cycle also containing u, and it is not hard to see that $\boldsymbol{\alpha} + m\boldsymbol{g}$ for any $m \in \mathbb{N}$ forms a rooted cycle. Theorem 2.1(b) then implies that there exists an experiment $(P, \mathcal{C}_m, \rho_P)$ such that

$$(\boldsymbol{\alpha} + m\boldsymbol{g}) \cdot \mathcal{Q}(\boldsymbol{r}) = -\log(\mathrm{Tr}[\tilde{P}\tilde{\mathcal{C}}_m(\tilde{\rho}_P)]). \tag{B.22}$$

Finally, it is clear from the proof of Theorem 2.1(b) that each $\mathcal{C}_m$ can actually be chosen as $\mathcal{C}^m$ (i.e., concatenating $\mathcal{C}$ m times) for an appropriate gate sequence $\mathcal{C}$. This completes the proof. □

In light of Theorem 2.4, one can design an algorithm that not only learns everything in L_R, but also learn a basis of L_R^G to relative precision.

Algorithm 2 [Relative precision learning of gate set Pauli noise].

1. Select a reduced cycle basis of L_R^G, denoted as B_R^G.
2. Augment B_R^G into a basis for L_R using vectors from $Q^{\mathsf{T}}(\mathsf{B})$.
3. Learn all vectors from B_R^G by running the set of experiments $\{(P, \mathcal{C}^m, \rho_P)\}_m$ described by Theorem 2.4.
4. Learn all the other basis vectors using Step 2 and 3 from Algorithm 1.

Again, the major challenge is to efficiently construct the reduced cycle basis, which would generally require doing linear algebra on an exponential-dimensional linear space. Our paper

[CZJF24] present examples that for physically-motivated noise ansatz and realistic gate set, such basis can be constructed efficiently, but I will not include those in this thesis.

We close this section with one final remark: Although specifying a reduced cycle basis of L_R^G is sufficient to specify any functions therein, learning each basis function to relative precision does not guarantee relative precision estimation for arbitrary functions. For example, relative-precision estimators for both $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ does not guarantee a relative-precision estimator for $\boldsymbol{f}_1 - \boldsymbol{f}_2$, especially when the true value of $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ are close to each other. In such cases, one might need to conduct more experiments than merely learning a basis in order to retrieve relative precision estimation for those functions. We will leave this issue for future study.

B.4 Applications to fully local noise model

B.4.1 Basic definitions

We have established general results on the learnability of any reduced Pauli noise models. Now we apply our theory to concrete physically-motivated noise ansatz. In this section, we will focus on the fully local noise model, or the so-called crosstalk-free noise model. As we will soon see, the reduced gauge space in this case is generated by certain subsystem depolarizing channel. We will also give efficient construction of the learning algorithms.

Before defining into the spatially local Pauli noise model, let us first introduce the *layer-uncorrelated noise model*, which is mostly for mathematical convenience but also practically relevant:

Definition 2.22. *For a reduced noise model $(X_R, \mathcal{Q})$, if X_R and $\mathcal{Q}$ can be decomposed as*

$$\begin{aligned} X_R &= X_R^S \oplus X_R^M \bigoplus_{\mathcal{G}\in\mathfrak{G}} X_R^{\mathcal{G}}, \\ \mathcal{Q} &= \mathcal{Q}^S \oplus \mathcal{Q}^M \bigoplus_{\mathcal{G}\in\mathfrak{G}} \mathcal{Q}^{\mathcal{G}}, \end{aligned} \tag{B.23}$$

where $\mathcal{Q}^\square : X_R^\square \mapsto X^\square$ satisfies $\mathcal{Q}^\square(\boldsymbol{r}^\square) = \boldsymbol{x}^\square$ for all $\square \in \{S, M\} \cup \mathfrak{G}$, then we call it a ***layer-uncorrelated noise model****.*

In words, a reduced noise model being layer-uncorrelated means that, there is an independent set of reduced noise parameters for state preparation, measurement, and each layer of gates, respectively. Many noise models studied in the literature of quantum benchmarking satisfies this assumption (e.g., [Fla22, CDDH$^+$23]). One prominent exception is the so-called "gate-independent noise model" studied in e.g. standard randomized benchmarking [EAŻ05, KLR$^+$08, DCEL09, MGE11, MGJ$^+$12] where the noise channel associated with any Clifford gate is assumed to be the same. Though we will not consider such a noise model in this section, we remark that the general framework established in the previous section is still applicable.

Lemma 2.23. *For a layer-uncorrelated model $(X_R, \mathcal{Q})$, the reduced gauge space T_R satisfies*

$$\mathcal{Q}(T_R) = \bigcap_{\square\in\{S,M\}\cup\mathfrak{G}} \left\{ \boldsymbol{z} \in T : \boldsymbol{z}^\square \in \mathrm{Im}\mathcal{Q}^\square \right\}. \tag{B.24}$$

Proof. Corollary 2.19 says $\mathcal{Q}(T_R) = T \cap \mathrm{Im}\mathcal{Q}$, which satisfies

$$\begin{aligned} T \cap \mathrm{Im}\mathcal{Q} &= \{\boldsymbol{z} \in T : \boldsymbol{z} \in \mathrm{Im}\mathcal{Q}^S \oplus \mathrm{Im}\mathcal{Q}^M \bigoplus_{\mathcal{G}\in\mathfrak{G}} \mathrm{Im}\mathcal{Q}^{\mathcal{G}}\} \\ &= \{\boldsymbol{z} \in T : \boldsymbol{z}^S \in \mathrm{Im}\mathcal{Q}^S,\ \boldsymbol{z}^M \in \mathrm{Im}\mathcal{Q}^M,\ \boldsymbol{z}^{\mathcal{G}} \in \mathrm{Im}\mathcal{Q}^{\mathcal{G}},\ \forall \mathcal{G} \in \mathfrak{G}\} \\ &= \bigcap_{\square\in\{S,M\}\cup\mathfrak{G}} \left\{ \boldsymbol{z} \in T : \boldsymbol{z}^\square \in \mathrm{Im}\mathcal{Q}^\square \right\}. \end{aligned} \tag{B.25}$$

□

As a result, we can separately analyze the gauge vectors allowed by state preparation, measurements, and each layer of gates. The intersection of all of them will give the correct reduced gauge space. In what follows, we will use this approach to analyze the gauge space for the fully local noise model, and for the quasi-local noise model in the next section.

B.4.2 Learnability of fully local noise model

Let us first define the fully local noise model. For any $\mathcal{G} \in \mathfrak{G}$, we use $\mathrm{supp}(\mathcal{G})$ to denote the support of $\mathcal{G}$, i.e., the subsystem of qubits that $\mathcal{G}$ non-trivially acts on. The restriction of $\mathcal{G}$ to its support is denoted by $\hat{\mathcal{G}}$. The number of qubits in the support is denoted by $|\mathcal{G}|$. Recall that we use $\{\boldsymbol{\vartheta}_i\}$ and $\{\boldsymbol{e}_i\}$ to denote the standard basis for X_R and X, respectively. That is, the reduced and complete noise parameters are encoded into vectors according to $\boldsymbol{r} = \sum_i r_i \boldsymbol{\vartheta}_i$ and $\boldsymbol{x} = \sum_i x_i \boldsymbol{e}_i$, respectively.

We will first give the formal definition of the fully local noise model, and explain the intuition in a moment.

Definition 2.24 (fully local noise)**.** *Consider a layer-uncorrelated noise model $(X_R, \mathcal{Q})$:*

- *If the state preparation noise parameters are given by $\{r_j^S : j \in [n]\}$ and that*

$$\mathcal{Q}(\boldsymbol{\vartheta}_j^S) = \sum_{u \neq 0_n} \mathbb{1}[u_j = 1] \boldsymbol{e}_u^S, \quad \forall j \in [n], \tag{B.26}$$

  *we say the model has **fully local state preparation** noise. Similar for the measurement.*

- *If the noise parameters for some $\mathcal{G} \in \mathfrak{G}$ are given by $\{r_a^{\mathcal{G}} : a \in \mathsf{P}^{|\mathcal{G}|}, a \neq I_{|\mathcal{G}|}\}$ and that*

$$\mathcal{Q}(\boldsymbol{\vartheta}_a^{\mathcal{G}}) = \sum_{b \neq I_n} \mathbb{1}[b_{\mathrm{supp}(\mathcal{G})} = a] \boldsymbol{e}_b^{\mathcal{G}}, \quad \forall a \in \mathsf{P}^{|\mathcal{G}|}, a \neq I_{|\mathcal{G}|}, \tag{B.27}$$

we say the model satisfies fully local noise for $\mathcal{G}$. If this holds for all $\mathcal{G} \in \mathfrak{G}$ we say the model has ***fully local gate noise****.*

The model is ***fully local*** *if it has fully local state preparation, measurement, and gate noise.*

The following lemma describes the transformation between the reduced parameters and the complete parameters within a fully local model, as an immediate consequence of the above definitions.

Lemma 2.25. *If $(X_R, \mathcal{Q})$ satisfies fully local state preparation noise, then*

$$\begin{cases} x_u^S = \sum\limits_{j\in[n]} \mathbb{1}[u_j = 1] r_j^S, \quad \forall u \in \{0,1\}^n, u \neq 0_n. \\ r_j^S = x_{1_j}^S, \quad \forall j \in [n]. \end{cases} \tag{B.28}$$

Similar for the measurement side; If $(X_R, \mathcal{Q})$ satisfies fully local noise for some gate $\mathcal{G}$, then

$$\begin{cases} x_b^{\mathcal{G}} = r^{\mathcal{G}}_{b_{\mathrm{supp}(\mathcal{G})}}, \quad \forall b \in \mathsf{P}^n, b \neq I_n. \\ r_a^{\mathcal{G}} = x_b^{\mathcal{G}} \ s.t.\ b_{\mathrm{supp}(\mathcal{G})} = a, \quad \forall a \in \mathsf{P}^{|\mathcal{G}|}, a \neq I_{|\mathcal{G}|}. \end{cases} \tag{B.29}$$

Proof. By definition of the fully local noise model, one can easily verify that

$$\boldsymbol{x}^S = \mathcal{Q}^S(\boldsymbol{r}^S) = \sum_{u\neq 0_n} \sum_{j\in[n]} \mathbb{1}[u_j = 1] r_j^S \boldsymbol{e}_u^S. \tag{B.30}$$

$$\boldsymbol{x}^{\mathcal{G}} = \mathcal{Q}^{\mathcal{G}}(\boldsymbol{r}^{\mathcal{G}}) = \sum_{b\neq I_n} r^{\mathcal{G}}_{b_{\mathrm{supp}(\mathcal{G})}} \boldsymbol{e}_b^{\mathcal{G}}. \tag{B.31}$$

Comparing these with the definitions of $\boldsymbol{x}$ and $\boldsymbol{r}$ immediately gives us the transformation from $\{r_i\}$ to $\{x_i\}$. The correctness of the inverse transformation can be verified by substitution. □

In words, the fully local state preparation or measurement noise is a qubit-wise independent channel, and the fully local gate noise is a channel acting only on the gate's support

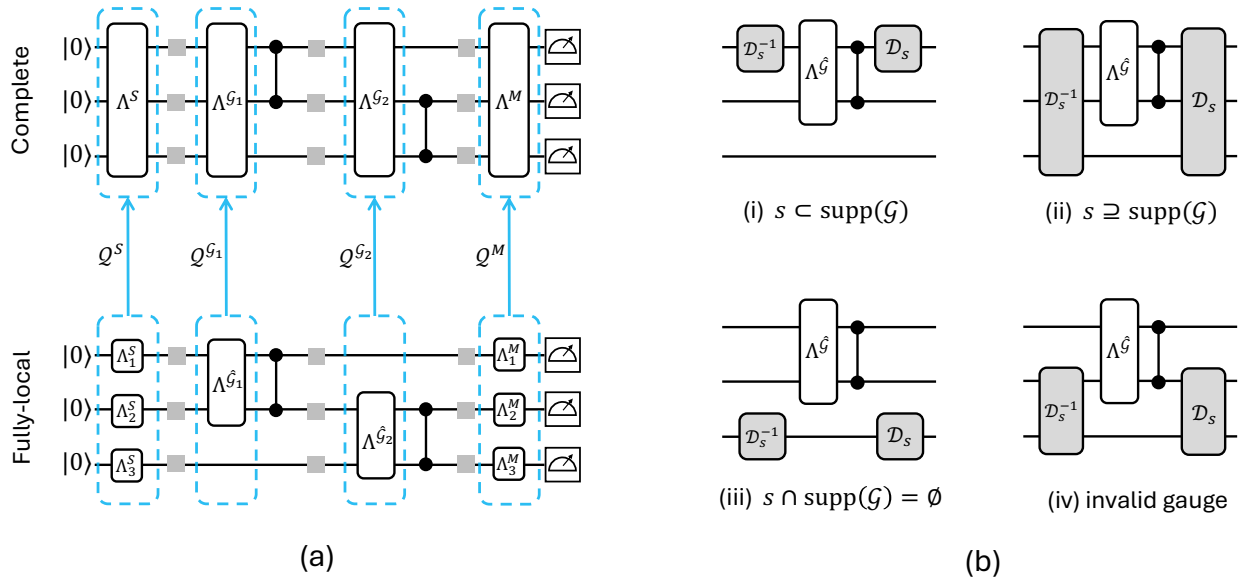


Figure B.2: (a) Illustration of a fully local noise model (bottom) and how it is embedded into a complete noise model (above). Note that the state prep., meas., and each layer of gates have an independent embedding map, which means this is a layer-uncorrelated noise model. (b) Illustration of the subsystem depolarizing gauges (denoted by $\mathcal{D}_s$) acting on a gate with fully local noise. Here, (i)-(iii) shows the three types of valid SDG's as given by Theorem 2.5(b), out of which only case (i) can non-trivially changes the noise channel. (iv) shows a SDG that is not an allowed gauge transformation (when the gate has a fully-connected pattern transfer subgraph). Intuitively, it propagates the 2-qubit noise channel into a 3-qubit channel, violating the fully local assumptions.

and inducing no crosstalk on idling qubits. See Fig. B.2(a) for an illustration. We study these cases separately, as there could be situations where the SPAM noise is local while gate noise is not, or the other way around.

We also note that whether fully local gate noise describes contextual noise depends on the definition of $\mathfrak{G}$: If one define $\mathfrak{G}$ in a way that each $\mathcal{G} \in \mathfrak{G}$ acts non-trivially at, say, at most 2-qubit. Let $\mathcal{G}_1$, $\mathcal{G}_2$ be two Clifford layers from $\mathfrak{G}$ that act non-trivially on disjoint sets of qubits. The noise channel for applying both gates in parallel can then be viewed as the tensor product of two associated noise channels. However, if one believe the parallel application should suffer from a different noise channel, one can add it as an independent layer into $\mathfrak{G}$. Although one needs to be careful not to makes the size $\mathfrak{G}$ exponentially large.

The advantages of a fully local noise model is a small number of model parameters (as small as $O(n)$ for a system with a linear connectivity), which makes working with such model potentially computationally efficient. Previous work, including ACES [Fla22], has focused on characterizing such a noise model. It is thus of great interest to understand the learnable/gauge degrees of freedom within this model. For this purpose, let us introduce the following set of gauge vectors.

Definition 2.26 (Subsystem depolarizing gauge, SDG)**.** *For any $s \in \{0,1\}^n \backslash 0_n$, let $\mathfrak{d}_s$ be the cut vector specified by the vertex partition*

$$\mathsf{V}_0 = \{u : u_s = 0_{|s|}\} \cup \{v_R\}, \quad \mathsf{V}_0^c = \{u : u_s \neq 0_{|s|}\}.$$

*We call such $\mathfrak{d}_s$ a **subsystem depolarizing gauge** (SDG).*

To see why $\mathfrak{d}_s$ is called a subsystem depolarizing gauge, let us first write down its expansion

on the standard basis:

$$\begin{aligned}
\boldsymbol{e}_u^S \cdot \mathfrak{d}_s &= \mathbb{1}[u_s \neq 0_{|s|}], && \forall u \in \{0,1\}^n \backslash 0_n. \\
\boldsymbol{e}_u^M \cdot \mathfrak{d}_s &= -\mathbb{1}[u_s \neq 0_{|s|}], && \forall u \in \{0,1\}^n \backslash 0_n. \\
\boldsymbol{e}_a^{\mathcal{G}} \cdot \mathfrak{d}_s &= \mathbb{1}[\mathrm{pt}(\mathcal{G}(a))_s \neq 0_{|s|}] - \mathbb{1}[\mathrm{pt}(a)_s \neq 0_{|s|}], && \forall \mathcal{G} \in \mathfrak{G},\ \forall a \in \mathsf{P}^n \backslash I_n.
\end{aligned} \tag{B.32}$$

which can be obtained using the definition of cut vectors and the PTG. Now, consider a transformation of noise parameters given by $\boldsymbol{x} \mapsto \boldsymbol{x} + \eta \mathfrak{d}_s$. It is not hard to verify this corresponds to the following gauge transformation,

$$\tilde{\rho}_0 \mapsto \mathcal{D}_{s,\eta}(\tilde{\rho}_0), \quad \tilde{E}_j \mapsto \mathcal{D}_{s,\eta}^{-1}(\tilde{E}_j), \quad \tilde{\mathcal{G}}_s \mapsto \mathcal{D}_{s,\eta} \circ \tilde{\mathcal{G}} \circ \mathcal{D}_{s,\eta}^{-1}, \tag{B.33}$$

where $\mathcal{D}_{s,\eta} \coloneqq \sum_a e^{-\eta \mathbb{1}[a_s \neq I_{|s|}]} |\sigma_a\rangle\!\rangle\langle\!\langle \sigma_a|$ is a partially depolarizing channel (with strength parameter η) acting on the subsystem specified by the index set s. For notational convenience, we define

$$\mathcal{D}_s \coloneqq \mathcal{D}_{s,1} = \sum_a e^{-\mathbb{1}[a_s \neq I_{|s|}]} |\sigma_a\rangle\!\rangle\langle\!\langle \sigma_a|, \tag{B.34}$$

and call $\mathcal{D}_s$ simply as the subsystem depolarizing channel on s.

Lemma 2.27. *The set of SDGs $\{\mathfrak{d}_s\}_{s \neq 0_n}$ forms a basis for the gauge space T.*

The proof is given in Sec. B.7.3. In fact, the gauge space of the fully local noise model is spanned by a subset of the SDG's. We have the following results:

Theorem 2.5. *Given a layer-uncorrelated noise model $(X_R, \mathcal{Q})$*

(a) If the model satisfies fully local state preparation noise, then

$$\{\boldsymbol{z} \in T : \boldsymbol{z}^S \in \mathrm{Im}\mathcal{Q}^S\} = \mathrm{span}\{\mathfrak{d}_s : |s| = 1\}. \tag{B.35}$$

Similarly for the measurement noise.

(b) If the model satisfies fully local noise for some gate $\mathcal{G} \in \mathfrak{G}$, then

$$\{\boldsymbol{z} \in T : \boldsymbol{z}^{\mathcal{G}} \in \mathrm{Im}\mathcal{Q}^{\mathcal{G}}\} \supseteq \mathrm{span}\{\mathfrak{d}_s : s \subset \mathrm{supp}(\mathcal{G}) \text{ or } s \supseteq \mathrm{supp}(\mathcal{G}) \text{ or } s \cap \mathrm{supp}(\mathcal{G}) = \varnothing\}. \tag{B.36}$$

Furthermore, the two sides are equal if $\hat{\mathcal{G}}$ induces a connected pattern transfer subgraph.

(c) If the model is fully local, its embedded gauge space is given by

$$\mathcal{Q}(T_R) = \mathrm{span}\{\mathfrak{d}_s : |s| = 1\}. \tag{B.37}$$

And thus $\dim T_R = \dim \mathcal{Q}(T_R) = n$.

Here, the pattern transfer subgraph of $\hat{\mathcal{G}}$ is a $|\mathcal{G}|$-qubit PTG with the only gate being $\hat{\mathcal{G}}$, and with the SPAM (root) node removed. Examples of 2-qubit gates that have connected pattern transfer subgraph include CZ, iSWAP, but not SWAP [CLO$^+$23], see Fig. B.3. As another example, the pattern transfer subgraph for CZ$\otimes$CZ is not connected. The allowed/forbidden gauge transformation for the fully local gate noise are depicted in Fig. B.2(b).

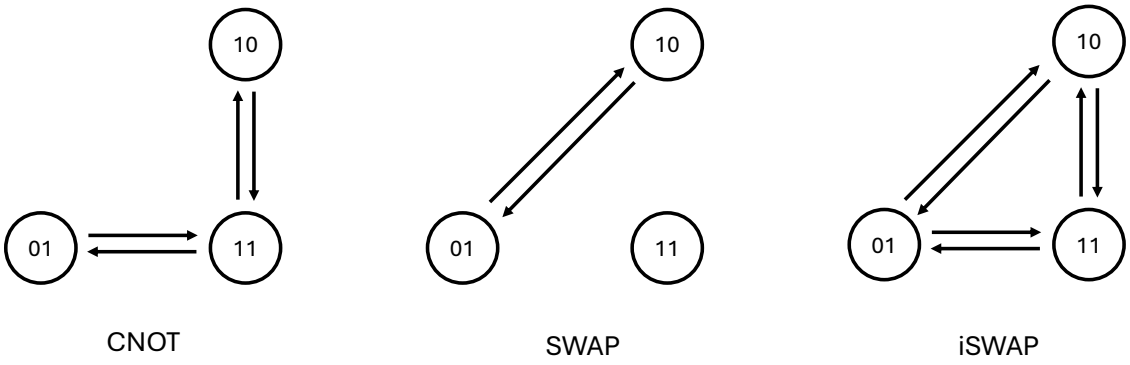


Figure B.3: The possible pattern transfer subgraphs of non-trivial 2-qubit Clifford gates. Labels of edges are omitted. Self loops are omitted. Multi-edges are represented by a single edge.

Theorem 2.5(c) characterizes the learnability of fully local Pauli noise models. In words, the embedded gauge space $\mathcal{Q}(T_R)$ is spanned exactly by those n single-qubit depolarizing

gauges. One can also write down the reduced gauge space as $T_R = \text{span}\{\mathcal{Q}^{-1}(\mathfrak{d}_s) : |s| = 1\}$, where each $\mathcal{Q}^{-1}(\mathfrak{d}_s)$ is well-defined[3] and can be efficiently calculated using the inverse transformation given in Lemma 2.25. The learnable space L_R can be obtained by, e.g., computing the orthogonal complement of T_R within X_R using standard linear algebra procedures. We will come back to the construction of L_R later.

Now we present the proof for Theorem 2.5(a), the direct part of 2.5(b), and 2.5(c). The converse part of 2.5(b) is slightly more involved and is postponed to App. B.7.3 for clarity.

Proof of Theorem 2.5(a). Let $\pi^S : X \mapsto X^S$ be the projection map $\pi^S(\boldsymbol{x}) = \boldsymbol{x}^S$, i.e., restriction to the state preparation parameters. We first show *L.H.S.* $\supseteq$ *R.H.S.*, for which we only need to prove that $\pi^S(\mathfrak{d}_{1_j}) \in \text{Im}\mathcal{Q}^S$ for all $j \in [n]$. Indeed, we have that,

$$\pi^S(\mathfrak{d}_{1_j}) = \sum_u \mathbb{1}[u_j \neq 0]\boldsymbol{e}_u^S = \mathcal{Q}^S(\boldsymbol{\vartheta}_j^S). \tag{B.38}$$

The two equations are by Definition 2.26 and Definition 2.24, respectively. Thus we have *L.H.S.* $\supseteq$ *R.H.S.*.

Conversely, notice that π^S is an invertible map when restricted to T. To see this, consider the canonical cut basis $\{\mathfrak{y}_u : u \neq 0_n\}$ for T such that each $\mathfrak{y}_u$ is induced by the vertex partition $\mathsf{V} = \mathsf{V}_0 \cup \mathsf{V}_0^c$ where $\mathsf{V}_0 = \{u\}$. We have $\pi^S(\mathfrak{y}_u) = -\boldsymbol{e}_u^S$. This means π^S is a basis transform between T and X^S, thus invertible. Therefore,

$$\dim(\{\boldsymbol{z} \in T : \pi^S(\boldsymbol{z}) \in \text{Im}\mathcal{Q}^S\}) = \dim(\text{Im}\mathcal{Q}^S) \leq \dim(X_R^S) = n, \tag{B.39}$$

where the first equation uses the invertibility of π^S restricted to T. This means the dimension of L.H.S. is no larger than R.H.S. Consequently, they must be equal. The proof works similarly on the measurement side. □

3. Though $\mathcal{Q}$ is not invertible, it is injective and we know that $\forall |s| = 1$, $\mathfrak{d}_s \in \text{Im}\mathcal{Q}$, thus $\mathcal{Q}^{-1}(\mathfrak{d}_s)$ is uniquely defined.

Proof of Theorem 2.5(b). Let $\pi^{\mathcal{G}} : X \mapsto X^{\mathcal{G}}$ be the projection map $\pi^{\mathcal{G}}(\boldsymbol{x}) = \boldsymbol{x}^{\mathcal{G}}$. Recall that

$$\boldsymbol{e}_a^{\mathcal{G}} \cdot \mathfrak{d}_s = \mathbb{1}[\mathrm{pt}(\mathcal{G}(a))_s \neq 0_{|s|}] - \mathbb{1}[\mathrm{pt}(a)_s \neq 0_{|s|}], \quad \forall a \neq I_n.$$

- If $s \cap \mathrm{supp}(\mathcal{G}) = \varnothing$: we have $\mathrm{pt}(\mathcal{G}(a))_s = \mathrm{pt}(a)_s$ and thus $\boldsymbol{e}_a^{\mathcal{G}} \cdot \mathfrak{d}_s = 0$ for all a. Therefore, we trivially have $\pi^{\mathcal{G}}(\mathfrak{d}_s) \in \mathrm{Im}\mathcal{Q}^{\mathcal{G}}$.

- If $s \supseteq \mathrm{supp}(\mathcal{G})$: we have $\mathrm{pt}(\mathcal{G}(a))_s = 0_{|s|} \Leftrightarrow \mathrm{pt}(a)_s = 0_{|s|}$ because $\hat{\mathcal{G}}(a_{\mathrm{supp}(\mathcal{G})}) = I_{|\mathcal{G}|}$ if and only if $a_{\mathrm{supp}(\mathcal{G})} = I_{|\mathcal{G}|}$. Thus $\boldsymbol{e}_a^{\mathcal{G}} \cdot \mathfrak{d}_s = 0$ and $\pi^{\mathcal{G}}(\mathfrak{d}_s) \in \mathrm{Im}\mathcal{Q}^{\mathcal{G}}$ trivially holds.

- If $s \subset \mathrm{supp}(\mathcal{G})$: choose an $\mathfrak{r}^{\mathcal{G}} \in X_R^{\mathcal{G}}$ such that

$$\boldsymbol{\vartheta}_b^{\mathcal{G}} \cdot \mathfrak{r}^{\mathcal{G}} = \mathbb{1}[\mathrm{pt}(\hat{\mathcal{G}}(b))_s \neq 0_{|s|}] - \mathbb{1}[\mathrm{pt}(b)_s \neq 0_{|s|}]. \tag{B.40}$$

  Then we can see that

$$\boldsymbol{e}_a^{\mathcal{G}} \cdot \mathcal{Q}^{\mathcal{G}}(\mathfrak{r}^{\mathcal{G}}) = \mathbb{1}[\mathrm{pt}(\hat{\mathcal{G}}(a_{\mathrm{supp}(\mathcal{G})}))_s \neq 0_{|s|}] - \mathbb{1}[\mathrm{pt}(a_{\mathrm{supp}(\mathcal{G})})_s \neq 0_{|s|}] = \boldsymbol{e}_a^{\mathcal{G}} \cdot \mathfrak{d}_s. \tag{B.41}$$

  This means $\pi^{\mathcal{G}}(\mathfrak{d}_s) = \mathcal{Q}^{\mathcal{G}}(\mathfrak{r}^{\mathcal{G}})$. Thus $\pi^{\mathcal{G}}(\mathfrak{d}_s) \in \mathrm{Im}\mathcal{Q}^{\mathcal{G}}$.

Combining all the three cases and using linearity, we conclude that

$$\{\boldsymbol{z} \in T : \boldsymbol{z}^{\mathcal{G}} \in \mathrm{Im}\mathcal{Q}^{\mathcal{G}}\} \supseteq \mathrm{span}\{\mathfrak{d}_s : s \subset \mathrm{supp}(\mathcal{G}) \text{ or } s \supseteq \mathrm{supp}(\mathcal{G}) \text{ or } s \cap \mathrm{supp}(\mathcal{G}) = \varnothing\}. \tag{B.42}$$

Conversely, the two sides are equal if $\hat{\mathcal{G}}$ induces a connected pattern transfer subgraph, which is proven in App. B.7.3. The proof works by establishing an upper bound on the dimension of the L.H.S., and then show by counting that this bound equals the dimension of the R.H.S. □

We make a few remarks regarding the proof of Theorem 2.5(b). First, it is clear from the proof that $\mathrm{span}\{\mathfrak{d}_s : s \subset \mathrm{supp}(\mathcal{G})\}$ are gauges that non-trivially change the noise parameters

of $\mathcal{G}$, while $\mathrm{span}\{\mathfrak{d}_s : s \supseteq \mathrm{supp}(\mathcal{G}) \text{ or } s \cap \mathrm{supp}(\mathcal{G}) = \varnothing\}$ are gauges that "pass through" $\mathcal{G}$ leaving gate noise parameters unchanged; Second, The proof for Eq. (B.42) taking equality crucially relies on the connectivity of the pattern transfer subgraph of $\widehat{\mathcal{G}}$. If this is not satisfied, there could be more gauges that can trivially pass through $\mathcal{G}$, which needs to be analyzed *ad hoc* using Theorem 2.3.

Proof of Theorem 2.5(c). Thanks to Lemma 2.23, $\mathcal{Q}(T_R)$ is given by the intersection of Eq. (B.35) for SPAM and Eq. (B.36) for all $\mathcal{G} \in \mathfrak{G}$. It's easy to see that Eq. (B.35) is always a subset of Eq. (B.36) for any $\mathcal{G} \in \mathfrak{G}$, thus their intersection is simply given by Eq. (B.35), the single-qubit SDG's. □

B.4.3 Efficient learning of fully local noise model

Having characterized the reduced gauge space of the fully local noise model, we turn to the task of designing efficient learning experiments. Following the discussion in Sec. B.3, the key is to find an appropriate reduced cycle basis for L_R. Let us start with the first type of construction which gives the simplest experimental design (but does not care about relative precision estimation). Recall the rooted cycle basis defined in Eq. (B.12),

$$\mathsf{B} \coloneqq \{\boldsymbol{e}^S_u + \boldsymbol{e}^M_u : u \neq 0_n\} \cup \{\boldsymbol{e}^S_{\mathrm{pt}(a)} + \boldsymbol{e}^{\mathcal{G}}_a + \boldsymbol{e}^M_{\mathrm{pt}(\mathcal{G}(a))} : \mathcal{G} \in \mathfrak{G}, a \neq I_n\}. \tag{B.43}$$

Consider the following subset of B, denoted as B',

$$\mathsf{B}' \coloneqq \{\boldsymbol{e}^S_{1_j} + \boldsymbol{e}^M_{1_j} : j \in [n]\} \cup \{\boldsymbol{e}^S_{\mathrm{pt}(a)} + \boldsymbol{e}^{\mathcal{G}}_a + \boldsymbol{e}^M_{\mathrm{pt}(\mathcal{G}(a))} : \mathcal{G} \in \mathfrak{G}, \mathrm{supp}(a) \subseteq \mathrm{supp}(\mathcal{G}), a \neq I\}. \tag{B.44}$$

Proposition 2.28. *Let* $\mathsf{B}_R \coloneqq \mathcal{Q}^{\mathsf{T}}(\mathsf{B}')$. *Then,* B_R *forms a reduced cycle basis for* L_R.

Proof. We have

$$\begin{aligned}
&\mathsf{B}_R \\
&:= \mathcal{Q}^{\mathsf{T}} \left(\{ \boldsymbol{e}_u^S + \boldsymbol{e}_u^M : |u| = 1 \} \cup \{ \boldsymbol{e}_{\mathrm{pt}(a)}^S + \boldsymbol{e}_a^{\mathcal{G}} + \boldsymbol{e}_{\mathrm{pt}(\mathcal{G}(a))}^M : \mathcal{G} \in \mathfrak{G}, \mathrm{supp}(a) \subseteq \mathrm{supp}(\mathcal{G}) \} \right) \\
&= \left\{ \boldsymbol{\vartheta}_j^S + \boldsymbol{\vartheta}_j^M : j \in [n] \right\} \cup \{ \sum_{j \in \mathrm{supp}(\mathcal{G})} (\mathbb{1}[\mathrm{pt}(b)_j \neq 0] \boldsymbol{\vartheta}_j^S + \mathbb{1}[\mathrm{pt}(\widehat{\mathcal{G}}(b))_j \neq 0] \boldsymbol{\vartheta}_j^M) + \boldsymbol{\vartheta}_b^{\mathcal{G}} : \mathcal{G} \in \mathfrak{G} \}.
\end{aligned} \tag{B.45}$$

Recall that $\widehat{\mathcal{G}}$ is $\mathcal{G}$ restricted to its support. It is not hard to see that B_R forms a linearly-independent set. Meanwhile, $|\mathsf{B}_R| = n + \sum_{\mathcal{G} \in \mathfrak{G}} (4^{|\mathcal{G}|} - 1) = \dim(X_R) - \dim(T_R) = \dim(L_R)$. This means B_R is indeed a basis for L_R. □

Following Algorithm 1, the following set of experiments is sufficient to learn the noise model:

1. For $j \in [n]$, prepare $\tilde{\rho}_0$ and measure $\tilde{Z}$ on the jth qubit.

2. For $\mathcal{G} \in \mathfrak{G}$, $b \in \mathsf{P}^{|\mathcal{G}|} \backslash I_{|\mathcal{G}|}$, prepare $\tilde{\rho}_b$, apply $\tilde{\mathcal{G}}$, and measure $\tilde{P}_{\widehat{\mathcal{G}}(b)}$ on the support of $\mathcal{G}$.

The total number of experiments is $n + \sum_{\mathcal{G} \in \mathfrak{G}} (4^{|\mathcal{G}|} - 1)$. If each $\mathcal{G}$ acts non-trivially only on a constant number of qubits, the number of experiments needed is $O(n)$, and is efficient to conduct. We also remark that, here we have not taken advantages of the fact that multiple observables can be extracted from the same experiments. In practice, the sample complexity can thus be further reduced.

B.5 Applications to quasi-local noise model

We have presented a comprehensive treatment on learning the fully local Pauli noise model. In practice, the effect of spatially correlated noise can sometimes be non-negligible, both for SPAM [BSK+21] and for gate [FM14, NB19]. It is thus desirable to study a quasi-local noise model that takes a finite spatial correlation (e.g., nearest-neighbor) into consideration while keeping the total number of parameters manageable. Such models have been studied

in the literature of Pauli noise learning [FW20, VDBMKT23, EWP$^+$19, WKBK23]. In this section, we will introduce the quasi-local model in the framework of gate set Pauli noise learning, provide results on its learnability, and discuss efficient learning algorithms. As a specific example, we will apply our theory to a noise model that has been intensively studied in recent experiments [VDBMKT23, KEA$^+$23] and obtain a self-consistent algorithm for learning the noise model, which has not been previously achieved.

B.5.1 Basic definitions

We first introduce some notations that are needed for defining the quasi-local noise model. For $a, b \in \mathsf{P}^n$, we write $b \triangleleft a$ if $b_i \neq I \Rightarrow a_i = b_i$ for all $i \in [n]$, calling b majorized by a. For any $S \subseteq [n]$, we write $a \sim S$ if $a_i \neq I \Rightarrow i \in S$. For any $\mathbb{S} \subseteq 2^{[n]}$, we write $a \sim \mathbb{S}$ if $a \sim S$ for at least one $S \in \mathbb{S}$.

Let Ω be a subset of $2^{[n]}$ that satisfies $\forall \nu \in \Omega$, $\forall \varnothing \subsetneq \mu \subseteq \nu$, $\mu \in \Omega$. We call such Ω a *factor set*. We also use Ω_* to denote the set of maximal subsets within Ω, i.e., those that do not belong to any other subsets. As an example, $\Omega = \{\{1\}, \{2\}, \{3\}, \{1,2\}, \{2,3\}\}$ is a valid factor set, with the corresponding $\Omega_* = \{\{1,2\}, \{2,3\}\}$. Obviously, Ω, Ω_* uniquely determines each other.

Definition 2.29. *An n-qubit Pauli channel Λ is Ω-local if its Pauli eigenvalues satisfies*

$$\lambda_a = \prod_{b \triangleleft a, b \sim \Omega} \exp(-r_b), \quad \forall a \neq I_n, \tag{B.46}$$

for some $\{r_b \in \mathbb{R} : b \sim \Omega, b \neq I_n\}$ and a factor set Ω.

For notational simplicity, we sometimes omit the requirement that $b \neq I_n$ and set $r_{I_n} = 0$.

A few remarks regarding this definition: First, all n-qubit Pauli channels (with all Pauli eigenvalues being positive) are $2^{[n]} \backslash \varnothing$-local. Note that we do not require r_b to be non-negative. Second, this definition basically says that the eigenvalues $\{\lambda_a\}_a$ factorize ac-

cording to a Markov Random Field (MRF) [Cli90] specified by the factor decomposition Ω. This is opposed to [FW20] where Pauli channels with error rates $\{p_a\}_a$ factorized according to an MRF are considered. It is unclear whether these two models are equivalent, even approximately. Alternatively, our definition of Ω-local is equivalent to requiring that Λ can be expressed as a concatenation of Pauli diagonal maps[4] acting on each subsystem from Ω, as proven in [WKBK23] (In fact, this is the picture we will use in most figures shown in this section). One can also show that the sparse Pauli-Lindblad model used in [VDBMKT23, KEA+23, HGM+24] is equivalent to our definition. Finally, throughout this paper we assume Ω is known a priori, from, e.g., knowledge about the device topology, leaving the topic of structure learning (see e.g. [RF23]) for future study. We provide more details about the quasi-local Pauli noise models in Appendix B.7.4.

B.5.2 Learnability of quasi-local noise model

Note that, in terms of x_a, Definition 2.29 gives that $x_a = \sum_{b \lhd a, b \sim \Omega} r_b$. This motivates the following definition of a quasi-local Pauli noise model. Note that, we have taken into account that the SPAM noise channels are symmetric Pauli channels (i.e., eigenvalues only depend on the Pauli pattern). An example of quasi-local noise model is given in Fig. B.4

Definition 2.30 (Quasi-local noise)**.** *Consider a layer-uncorrelated noise model $(X_R, \mathcal{Q})$:*

- *If the state preparation noise parameters are given by $\{r^S_\nu : \nu \in \Omega^S\}$, where Ω^S is a factor set, such that*

$$\mathcal{Q}(\boldsymbol{\vartheta}^S_\nu) = \sum_{\mu \neq \varnothing} \mathbb{1}[\nu \subseteq \mu] \boldsymbol{e}^S_\mu, \quad \forall \nu \in \Omega^S, \nu \neq \varnothing, \tag{B.47}$$

4. Note that, we do not require those Pauli diagonal maps to be completely positive. Similarly, when we say our quasi-local model is equivalent to the sparse Pauli-Lindblad model, we have not required the positivity of the model parameters.

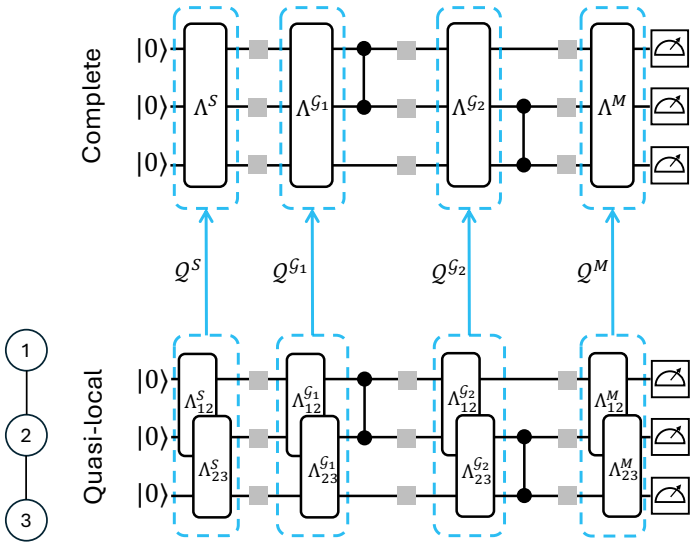


Figure B.4: Example of a 3-qubit quasi-local noise model and the embedding into a complete model. Here, the noise channels for state prep., meas., and all gates are assumed to be Ω-local where $\Omega^* = \{\{1,2\},\{2,3\}\}$, with the factor graph shown in the bottom-left side. In other words, $\Lambda^S = \Lambda^S_{12} \circ \Lambda^S_{23}$ where Λ^S_{12}, Λ^S_{23} are Pauli diagonal maps supporting on $\{1,2\}$, $\{2,3\}$, respectively. Similar for other Pauli noise channels.

we say the model has Ω^S***-local state preparation*** *noise. Similar for the measurement.*

- *If the noise parameters for some* $\mathcal{G} \in \mathfrak{G}$ *are given by* $\{r_b^{\mathcal{G}} : b \sim \Omega^{\mathcal{G}}, b \neq I_n\}$*, where* $\Omega^{\mathcal{G}}$ *is a factor set, such that*

$$\mathcal{Q}(\boldsymbol{\vartheta}_b^{\mathcal{G}}) = \sum_{a \neq I_n} \mathbb{1}[b \triangleleft a] \boldsymbol{e}_a^{\mathcal{G}}, \quad \forall b \sim \Omega^{\mathcal{G}}, b \neq I_n, \tag{B.48}$$

We say the model has $\Omega^{\mathcal{G}}$***-local noise for*** $\mathcal{G}$*.*

- *If the state preparation, measurements, and all layers have quasi-local noise with locality parameters* $\{\Omega^S, \Omega^M, \Omega^{\mathcal{G}} : \mathcal{G} \in \mathfrak{G}\}$*, we say the model is quasi-local with the corresponding parameters.*

The following lemma describes the transformation between the reduced and the complete parameters within a quasi-local model.

Lemma 2.31. *If $(X_R, \mathcal{Q})$ satisfies Ω^S-local state preparation noise, then*

$$\begin{cases} x^S_\mu = \sum\limits_{\nu\subseteq\mu,\ \nu\in\Omega^S} r^S_\nu, \quad \forall \mu \in 2^{[n]}\backslash\varnothing. \\ r^S_\nu = \sum\limits_{\mu\subseteq\nu,\ \mu\in\Omega^S} (-1)^{|\nu|-|\mu|} x^S_\mu, \quad \forall \nu \in \Omega^S. \end{cases} \tag{B.49}$$

Similar for the measurement side; If $(X_R, \mathcal{Q})$ satisfies $\Omega^{\mathcal{G}}$-local noise for a gate $\mathcal{G}$, then

$$\begin{cases} x^{\mathcal{G}}_a = \sum\limits_{b\lhd a,\ b\sim\Omega^{\mathcal{G}}} r^{\mathcal{G}}_b, \quad \forall a \in \mathsf{P}^n\backslash I_n. \\ r^{\mathcal{G}}_b = \sum\limits_{a\lhd b,\ a\sim\Omega^{\mathcal{G}}} (-1)^{w(b)-w(a)} x^{\mathcal{G}}_a, \quad \forall b \sim \Omega^{\mathcal{G}},\ b \neq I_n. \end{cases} \tag{B.50}$$

Proof. The transformation from r's to x's can be obtained by expanding $\boldsymbol{x} = \mathcal{Q}(\boldsymbol{r})$ in the standard basis and compare coefficients. The inverse transformation can be verified by substitution. As a side comment, these pairs of transformation are known as the Möbius transformation. They are also used in [WKBK23] to study Pauli noise learning in the logical level. □

As a sanity check, one can retrieve the fully local model in Definition 2.24 by setting

$$\Omega^S = \Omega^M = \{\{j\} : j \in [n]\}, \quad \Omega^{\mathcal{G}} = \{\nu : \varnothing \neq \nu \subseteq \mathrm{supp}(\mathcal{G})\},\ \forall \mathcal{G} \in \mathfrak{G}. \tag{B.51}$$

Now, we make two important remarks about the gate noise models. First, we will take a more general approach to deal with parallel gates than the fully local case. Instead of letting each $\mathcal{G}$ to have a constantly small support, we allow it to be a tensor product of many few-qubit gates, i.e., $\mathcal{G} = \hat{\mathcal{G}}_1 \otimes \cdots \otimes \hat{\mathcal{G}}_M \otimes \mathrm{id}$, where the identity acts on $[n]\backslash \bigcup_i \mathrm{supp}(\mathcal{G}_i)$. Second, note that a Clifford gate affected by Pauli noise can be modeled in two equivalent

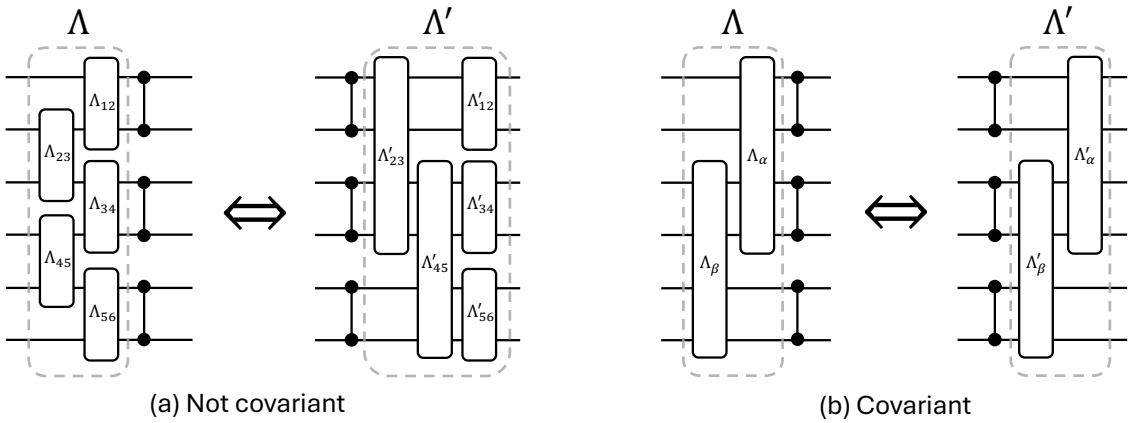


Figure B.5: Examples of quasi-local gate noise models that are incovariant and covariant, respectively. Here we consider a 6-qubit system and let $\mathcal{G} = \mathrm{CZ}^{\otimes 3}$. (a) The nearest-neighbor 2-local noise model ($\Omega^* = \{\{1,2\},\{2,3\},\cdots,\{5,6\}\}$) which is not $\mathcal{G}$-covariant. Indeed, if we commute the noise channel from before-$\mathcal{G}$ (left) to after-$\mathcal{G}$ (right), certain 2-body noise terms (e.g., Λ_{23}) will generically be propagated to be 4-body, making the whole noise channel no longer Ω-local; (b) A 4-local noise model ($\Omega^* = \{\alpha = \{1,2,3,4\}, \beta = \{3,4,5,6\}\}$) which is $\mathcal{G}$-covariant. Indeed, commuting the noise channel from before-$\mathcal{G}$ to after-$\mathcal{G}$ preserves the Ω-local structure. Note that all Pauli diagonal maps commute with each other.

ways: whether the noise channel happens before or after the gate, i.e.,

$$\tilde{\mathcal{G}} = \mathcal{G} \circ \Lambda = \Lambda' \circ \mathcal{G}, \quad \text{where } \Lambda' = \mathcal{G} \circ \Lambda \circ \mathcal{G}^\dagger. \tag{B.52}$$

Throughout this chapter, we have been adopting the noise-before-gate convention, but now it is worth looking at the other convention. Specifically, when Λ is an Ω-local Pauli channel, Λ' might not be Ω-local in general. This raised the following question: is there a physical reason to believe the gate noise has certain locality constraint in the noise-before-gate convention, but not in the noise-after-gate convention? To the best of our knowledge, no physical justification of this kind has been established in the literature. To circumvent the dilemma, we propose to study a specific type of locality assumption:

Definition 2.32. *An Ω-local noise model is called $\mathcal{G}$-**covariant**, if for any Ω-local Pauli channel Λ, $\mathcal{G} \circ \Lambda \circ \mathcal{G}^\dagger$ and $\mathcal{G}^\dagger \circ \Lambda \circ \mathcal{G}$ are also Ω-local Pauli channels.*

Since a $\mathcal{G}$-covariant noise model has the same locality constraints in either convention, it is much easier to justify their physical relevance. An example of covariant vs. incovariant noise model is given in Fig. B.5. When $\mathcal{G}$ represents a parallel application of many Clifford gates with disjoint supports, we have the following sufficient condition for a noise model to be $\mathcal{G}$-covariant:

Lemma 2.33. *Given $\mathcal{G} = \left(\bigotimes_{i=1}^{M} \hat{\mathcal{G}}_i\right) \otimes \mathrm{id}$, define the extended support map $\Xi_{\mathcal{G}} : [n] \mapsto [n]$ as*

$$\Xi_{\mathcal{G}}(s) \coloneqq \left(\bigcup_{\mathrm{supp}(\hat{\mathcal{G}}_i) \cap s \neq \varnothing} \mathrm{supp}(\hat{\mathcal{G}}_i) \right) \cup s. \tag{B.53}$$

For an Ω-local noise model, if $\forall s \in \Omega$, $\Xi_{\mathcal{G}}(s) \in \Omega$, then Ω is $\mathcal{G}$-covariant.

Clearly, for a Pauli P with support s, the support of $\mathcal{G}(P)$ cannot exceed $\Xi_{\mathcal{G}}(s)$. Lemma 2.33 roughly says that if for all Pauli $P \sim \Omega$, one has $\mathcal{G}(P) \sim \Omega$, then Ω must be $\mathcal{G}$-covariant. The proof is given in Appendix. B.7.4.

We are ready to present our main results on the learnability of quasi-local Pauli noise model. Perhaps not surprisingly, the gauge degrees of freedom can still be described using the SDG's in Definition 2.26, as follows:

Theorem 2.6. *Given a layer-uncorrelated noise model $(X_R, \mathcal{Q})$,*

(a) If the model has Ω^S-local state preparation noise, then

$$\{\boldsymbol{z} \in T : \boldsymbol{z}^S \in \mathrm{Im}\mathcal{Q}^S\} = \mathrm{span}\{\mathfrak{d}_\nu : \nu \in \Omega^S\}. \tag{B.54}$$

Similarly for the measurement noise.

(b) If the model has $\Omega^{\mathcal{G}}$-local noise for some $\mathcal{G} \in \mathfrak{G}$ and is $\mathcal{G}$-covariant, then

$$\{\boldsymbol{z} \in T : \boldsymbol{z}^{\mathcal{G}} \in \mathrm{Im}\mathcal{Q}^{\mathcal{G}}\} \supseteq \mathrm{span}\{\mathfrak{d}_\nu : \nu \in \Omega^{\mathcal{G}} \text{ or } \Xi_{\mathcal{G}}(\nu) = \nu\}. \tag{B.55}$$

(c) If $\Omega^S \subseteq \Omega^M$, $\Omega^{\mathcal{G}}$ is $\mathcal{G}$ covariant for all $\mathcal{G} \in \mathfrak{G}$, and that $\forall \nu \in \Omega^S, \forall \mathcal{G} \in \mathfrak{G}$, either $\nu \in \Omega^{\mathcal{G}}$ or $\Xi_{\mathcal{G}}(\nu) = \nu$, then the embedded gauge space for the whole model is given by

$$\mathcal{Q}(T_R) = \text{span}\{\mathfrak{d}_\nu : \nu \in \Omega^S\}. \tag{B.56}$$

Theorem 2.6(c) confirms that, given certain "goodness" assumptions of the quasi-local noise model, the embedded gauge space is indeed spanned by a subset of SDG's consistent with the factor sets. Similar to the fully local case, one can efficiently compute the reduced gauge space $T_R = \text{span}\{\mathcal{Q}^{-1}(\mathfrak{d}_\nu) : \nu \in \Omega^S\}$ by applying the Möbius inverse transformation given in Lemma 2.31.

Proof of Theorem 2.6(a). Let $\pi^S : X \mapsto X^S$ be the projection map $\pi^S(\boldsymbol{x}) = \boldsymbol{x}^S$. Recall from the proof of Theorem 2.5(a) that π^S is an isomorphism between T and X^S. We claim the following two linear spaces are equal.

$$\text{span}\{\mathcal{Q}^S(\boldsymbol{\vartheta}_\nu^S) : \nu \in \Omega^S\} = \text{span}\{\pi^S(\mathfrak{d}_\nu) : \nu \in \Omega^S\}. \tag{B.57}$$

Now we prove this claim. By definition, for any $\nu \in \Omega^S$ we have

$$\mathcal{Q}^S(\boldsymbol{\vartheta}_\nu^S) = \sum_{\mu \neq \varnothing} \mathbb{1}[\nu \subseteq \mu] \boldsymbol{e}_\mu^S =: \boldsymbol{\alpha}_\nu, \tag{B.58}$$

$$\pi^S(\mathfrak{d}_\nu) = \sum_{\mu \neq \varnothing} \mathbb{1}[\nu \cap \mu \neq \varnothing] \boldsymbol{e}_\mu^S =: \boldsymbol{\beta}_\nu. \tag{B.59}$$

Since both $\{\boldsymbol{\alpha}_\nu\}, \{\boldsymbol{\beta}_\nu\}$ have the same cardinality, and we know $\{\boldsymbol{\beta}_\nu\}$ are linearly-independent thanks to the linear independence of the SDG's and π^S being an isomorphism from T to X^S, we just need to show $\{\boldsymbol{\alpha}_\nu\}$ can linearly represent $\{\boldsymbol{\beta}_\nu\}$. We claim the following holds,

$$\boldsymbol{\beta}_\nu = -\sum_{\nu' \in \Omega^S} \mathbb{1}[\nu' \subseteq \nu](-1)^{|\nu'|} \boldsymbol{\alpha}_{\nu'}, \quad \forall \nu \in \Omega^S. \tag{B.60}$$

Indeed, expanding the RHS yields,

$$\begin{aligned} R.H.S. &= -\sum_{\mu\neq\varnothing} \boldsymbol{e}_\mu^S \sum_{\nu'\in\Omega^S} \mathbb{1}[\nu'\subseteq\nu]\mathbb{1}[\nu'\subseteq\mu](-1)^{|\nu'|} \\ &= -\sum_{\mu\neq\varnothing} \boldsymbol{e}_\mu^S \sum_{\nu'\in\Omega^S} \mathbb{1}[\nu'\subseteq\nu\cap\mu](-1)^{|\nu'|} \\ &= -\sum_{\mu\neq\varnothing} \boldsymbol{e}_\mu^S \sum_{\substack{\nu'\subseteq\nu\cap\mu,\\ \nu'\neq\varnothing}} (-1)^{|\nu'|} \\ &= \sum_{\mu\neq\varnothing} \boldsymbol{e}_\mu^S \mathbb{1}[\nu\cap\mu\neq\varnothing] = \boldsymbol{\beta}_\nu, \end{aligned} \tag{B.61}$$

where the third line uses the defining property of Ω^S that for any $\nu\in\Omega^S$, any nonempty $\nu'\subseteq\nu$ also satisfies $\nu'\in\Omega^S$; The last line uses the binomial theorem. This completes the proof of Eq. (B.57). Consequently, we have

$$\begin{aligned} \{\boldsymbol{z}\in T : \boldsymbol{z}^S\in \mathrm{Im}\mathcal{Q}^S\} &= \{\boldsymbol{z}\in T : \pi^S(\boldsymbol{z})\in \mathrm{span}\{\mathcal{Q}^S(\boldsymbol{\vartheta}_\nu^S) : \nu\in\Omega^S\}\} \\ &= \{\boldsymbol{z}\in T : \pi^S(\boldsymbol{z})\in \mathrm{span}\{\pi^S(\mathfrak{d}_\nu) : \nu\in\Omega^S\}\} \\ &= \mathrm{span}\{\mathfrak{d}_\nu : \nu\in\Omega^S\}. \end{aligned} \tag{B.62}$$

The second line uses Eq. (B.57). The last line uses π^S being an isomorphism between T and X^S. The proof works similarly for Ω^M-local measurement noise. □

Proof of Theorem 2.6(b). Let $\pi^{\mathcal{G}} : X\mapsto X^{\mathcal{G}}$ be the projection map $\pi^{\mathcal{G}}(\boldsymbol{x}) = \boldsymbol{x}^{\mathcal{G}}$. Recall that for all $a\neq I_n$, $\nu\neq\varnothing$,

$$\begin{aligned} \boldsymbol{e}_a^{\mathcal{G}}\cdot\mathfrak{d}_\nu &= \mathbb{1}[\mathrm{pt}(\mathcal{G}(a))_\nu\neq 0_\nu] - \mathbb{1}[\mathrm{pt}(a)_\nu\neq 0_\nu], \\ &= -\log\left(\langle\!\langle\sigma_a|\mathcal{G}^\dagger\mathcal{D}_\nu\mathcal{G}\mathcal{D}_\nu^{-1}|\sigma_a\rangle\!\rangle\right). \end{aligned} \tag{B.63}$$

Here $\mathcal{D}_\nu$ is the subsystem depolarizing channel on ν defined in Eq. (B.34).

- If $\Xi_{\mathcal{G}}(\nu)=\nu$, then $\mathcal{G}$ is a tensor product of gates on ν and on $[n]\backslash\nu$. Consequently, $\mathcal{G}$

commutes with $\mathcal{D}_\nu$. Thus, $\boldsymbol{e}_a^{\mathcal{G}} \cdot \mathfrak{d}_\nu = 0$ for all a, which means $\pi^{\mathcal{G}}(\mathfrak{d}_\nu) \in \mathrm{Im}\mathcal{Q}^{\mathcal{G}}$ trivially holds. See Fig. B.6(ii) for an example.

- If $\nu \in \Omega^{\mathcal{G}}$, one can see that both $\mathcal{D}_\nu$ and $\mathcal{D}_\nu^{-1}$ are $\Omega^{\mathcal{G}}$-local Pauli channels. Since the noise model is $\mathcal{G}$-covariant, $\mathcal{G}^\dagger \mathcal{D}_\nu \mathcal{G}$ is $\Omega^{\mathcal{G}}$-local. Since the concatenation of two $\Omega^{\mathcal{G}}$-local Pauli channels is clearly also $\Omega^{\mathcal{G}}$-local, $\mathcal{G}^\dagger \mathcal{D}_\nu \mathcal{G} \mathcal{D}_\nu^{-1}$ is Ω-local. This means,

$$\boldsymbol{e}_a^{\mathcal{G}} \cdot \mathfrak{d}_\nu = \sum_{b \lhd a, b \sim \Omega^{\mathcal{G}}} r'_b, \tag{B.64}$$

  for some coefficients $\{r'_b \in \mathbb{R} : b \sim \Omega^{\mathcal{G}}, b \neq I_n\}$. Comparing this to the definition of quasi-local gate noise, we see that $\pi^{\mathcal{G}}(\mathfrak{d}_\nu) \in \mathrm{Im}\mathcal{Q}^{\mathcal{G}}$. See Fig. B.6(i) for an example.

Combining the above two cases and using linearity, we complete the proof for Theorem 2.6(b). □

We make a few remarks about Theorem 2.6(b). First, it is clear from the proof that $\mathrm{span}\{\mathfrak{d}_\nu : \Xi_{\mathcal{G}}(\nu) = \nu\}$ are gauges that pass trivially through $\mathcal{G}$ leaving its noise parameters unchanged, while $\mathrm{span}\{\mathfrak{d}_\nu : \nu \in \Omega^{\mathcal{G}}\}$ are gauges that non-trivially transform the noise parameters of $\mathcal{G}$; Second, here we do not have a converse side as in Theorem 2.5(b), meaning there could be more allowed gauges than what we have characterized. See Fig. B.6 for an example of different types of gauges. We do not pursue such a tight characterization in the quasi-local case as it looks considerably more complicated. For our purpose of proving Theorem 2.6(c), the current results are sufficient; Finally, the requirement of noise being $\mathcal{G}$-covariant is indispensable for the theorem to hold. In Fig. B.7 we show a counterexample where $\mathfrak{d}_\nu$ is not an allowed gauge even if $\nu \in \Omega^{\mathcal{G}}$, when the gate noise model is not $\mathcal{G}$-covariant. This will be further discussed in Sec. B.6 when we encounter some realistic examples.

Proof of Theorem 2.6(c). Thanks to Lemma 2.23, $\mathcal{Q}(T_R)$ is given by the intersection of Eq. (B.54) for SPAM and Eq. (B.55) for all $\mathcal{G} \in \mathfrak{G}$. Under the given conditions, it is easy to

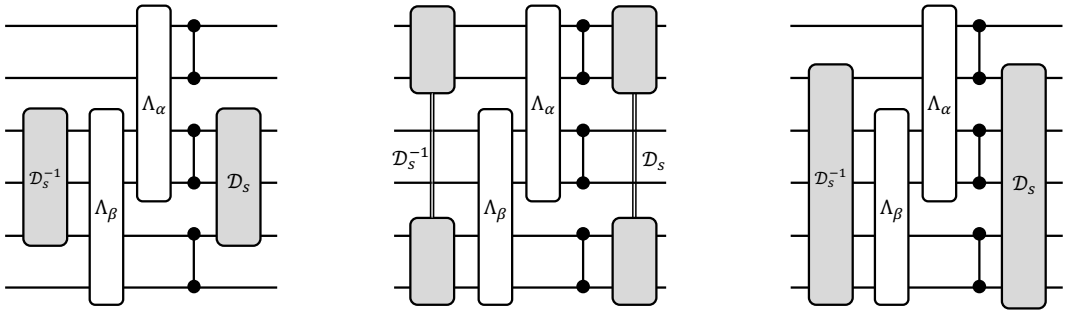


Figure B.6: Examples of three different cases of SDG's ($\mathcal{D}_s$) acting on $\mathcal{G} = \mathrm{CZ}^{\otimes 3}$ with covariant Ω-local noise model, $\Omega^* = \{\{1,2,3,4\},\{3,4,5,6\}\}$. Case (i) and (ii) are allowed gauges as shown in Theorem 2.6(b). Specifically, In Case (i) we have $s = \{3,4,5\} \in \Omega$, thus $\mathcal{D}_s$ is a valid gauge that non-trivially transform the noise channel; In Case (ii) we have $s = \{1,2,5,6\} = \Xi_{\mathcal{G}}(s)$, or in words, any Pauli supports within s still supports within s after the action of $\mathcal{G}$ (see Lemma 2.33). Here, even though $s \notin \Omega$, $\mathcal{D}_s$ commutes through $\mathcal{G}$ leaving the noise channel unchanged, and is thus a trivially valid gauge; Case (iii) shows an invalid gauge ($s = \{2,3,4,5,6\}$), which would generate a 6-body noise term, violating the locality assumption.

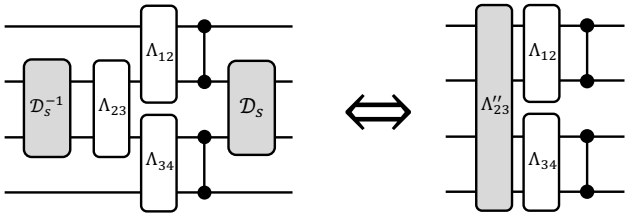


Figure B.7: Example of SDG acting on $\mathcal{G} = \mathrm{CZ}^{\otimes 2}$ with nearest-neighbor 2-local noise model (i.e., $\Omega^* = \{\{1,2\},\{2,3\},\{3,4\}\}$) that is *not* $\mathcal{G}$-covariant. In this case, though $s = \{2,3\} \in \Omega$, The SDG $\mathcal{D}_s$ would create a 4-body noise term (shown in the right) violating the Ω-local assumption, and is thus not an allowed gauge. This highlights the necessity of $\mathcal{G}$-covariance in obtaining Theorem 2.6(b).

see that Eq. (B.54) is always a subset of Eq. (B.55) for any $\mathcal{G} \in \mathfrak{G}$, thus their intersection is simply given by Eq. (B.54). □

B.5.3 Efficient learning of quasi-local noise model

We now discuss algorithms for learning a quasi-local noise model. In this section, we will assume all conditions of Theorem 2.6(c) are satisfied. That is, $\Omega^S \subseteq \Omega^M$; $\Omega^{\mathcal{G}}$ is $\mathcal{G}$-covariant for all $\mathcal{G} \in \mathfrak{G}$; and that $\forall \nu \in \Omega^S, \forall \mathcal{G} \in \mathfrak{G}$ either $\nu \in \Omega^{\mathcal{G}}$ or $\Xi_{\mathcal{G}}(\nu) = \nu$. For such a well-conditioned quasi-local noise model, we can easily find a reduced cycle basis for L_R as follows:

Consider the following subset of the rooted cycle basis B defined in Eq. (B.12), denoted by B'',

$$\mathsf{B}'' \coloneqq \{\boldsymbol{e}^S_\mu + \boldsymbol{e}^M_\mu : \mu \in \Omega^M\} \cup \{\boldsymbol{e}^S_{\mathrm{supp}(a)} + \boldsymbol{e}^{\mathcal{G}}_a + \boldsymbol{e}^M_{\mathrm{supp}(\mathcal{G}(a))} : \mathcal{G} \in \mathfrak{G},\ a \sim \Omega^{\mathcal{G}}\}. \tag{B.65}$$

Lemma 2.34. *Let* $\mathsf{B}_R \coloneqq \mathcal{Q}^{\mathsf{T}}(\mathsf{B}'')$. *Then* B_R *forms a reduced cycle basis for* L_R.

Proof. First, acting $\mathcal{Q}^{\mathsf{T}}$ on the rooted cycle basis B defined in Lemma 2.15 yields,

$$\begin{aligned}
\mathcal{Q}^{\mathsf{T}}(\boldsymbol{e}^S_\mu + \boldsymbol{e}^M_\mu) &= \sum_{\substack{\nu \subseteq \mu,\\ \nu \sim \Omega^S}} \boldsymbol{\vartheta}^S_\nu + \sum_{\substack{\nu \subseteq \mu,\\ \nu \sim \Omega^M}} \boldsymbol{\vartheta}^M_\nu, && \forall \mu \neq \varnothing, \\
\mathcal{Q}^{\mathsf{T}}(\boldsymbol{e}^S_{\mathrm{supp}(a)} + \boldsymbol{e}^{\mathcal{G}}_a + \boldsymbol{e}^M_{\mathrm{supp}(\mathcal{G}(a))}) &= \sum_{\substack{\nu \subseteq \mathrm{supp}(a),\\ \nu \sim \Omega^S}} \boldsymbol{\vartheta}^S_\nu + \sum_{\substack{b \lhd a,\\ b \sim \Omega^{\mathcal{G}}}} \boldsymbol{\vartheta}^{\mathcal{G}}_b + \sum_{\substack{\nu \subseteq \mathrm{supp}(\mathcal{G}(a)),\\ \nu \sim \Omega^M}} \boldsymbol{\vartheta}^M_\nu, && \forall a \neq I_n,\ \forall \mathcal{G} \in \mathfrak{G}.
\end{aligned} \tag{B.66}$$

Thus, $\mathsf{B}_R \coloneqq \mathcal{Q}^{\mathsf{T}}(\mathsf{B}'')$ takes the following form,

$$\mathsf{B}_R = \left\{ \sum_{\substack{\nu\subseteq\mu,\\ \nu\sim\Omega^S}} \boldsymbol{\vartheta}_\nu^S + \sum_{\nu\subseteq\mu} \boldsymbol{\vartheta}_\nu^M : \mu\in\Omega^M \right\} \cup \left\{ \sum_{\substack{\nu\subseteq\mathrm{supp}(a),\\ \nu\sim\Omega^S}} \boldsymbol{\vartheta}_\nu^S + \sum_{b\lhd a} \boldsymbol{\vartheta}_b^{\mathcal{G}} + \sum_{\substack{\nu\subseteq\mathrm{supp}(\mathcal{G}(a)),\\ \nu\sim\Omega^M}} \boldsymbol{\vartheta}_\nu^M : \mathcal{G}\in\mathfrak{G},\ a\sim\Omega^{\mathcal{G}} \right\}, \tag{B.67}$$

where we have used the defining property of Ω^M and $\Omega^{\mathcal{G}}$. let us first prove the linear independence of B_R. Consider a linear combination of all vectors from B_R that yields $\mathbf{0}$. Since only the second part of B_R contains gate noise parameters, and one can show $\{\sum_{b\lhd a} \boldsymbol{\vartheta}_b^{\mathcal{G}} : \mathcal{G}\in\mathfrak{G},\ a\sim\Omega^{\mathcal{G}}\}$ are linearly-independent (as they can represent $\{\boldsymbol{\vartheta}_a^{\mathcal{G}} : \mathcal{G}\in\mathfrak{G}, a\sim\Omega^{\mathcal{G}}\}$ using the inclusion-exclusion principle, see Appendix B.7.4 for similar expressions), the coefficients for all vectors from the second part must be zero; On the other hand, because $\{\sum_{\nu\subseteq\mu} \boldsymbol{\vartheta}_\nu^M : \nu\in\Omega^M\}$ are also linearly-independent (also thanks to the inclusion-exclusion principle), the coefficients for vectors from the first part must also be zero. This means B_R must be linearly independent. On the other hand, Theorem 2.6(c) implies $\dim(T_R) = |\Omega^S|$. We can see from the expression of B_R that $|\mathsf{B}_R| = \dim(X_R) - |\Omega^S| = \dim(X_R) - \dim(T_R) = \dim(L_R)$. We can thus conclude that B_R forms a basis for L_R. □

The construction of B_R implies the following set of experiments can fully learn the quasi-local reduced noise model,

1. For all $\mu\in\Omega^M$, prepare $\tilde{\rho}_0$ and measure $\tilde{Z}^\mu$.
2. For all $\mathcal{G}\in\mathfrak{G}$, $b\sim\Omega^{\mathcal{G}}$, prepare $\tilde{\rho}_b$, apply $\tilde{\mathcal{G}}$, and measure $\tilde{P}_{\mathcal{G}(b)}$.

If Ω^M and $\Omega^{\mathcal{G}}$ have polynomial size, the number of experiments needed is also polynomial, and is thus efficient. Besides, if each $\mathcal{G}$ is a parallel application of many Clifford gates each

acting on a constant number of qubits, many of the above experiments can be conducted in parallel. We leave such an optimization of experiment design as future research.

Of course, one might also want to learn gate noise parameter to relative precision in the quasi-local noise model. We will not pursue a general efficient algorithm for finding reduced cycle basis for $L_R^{\mathcal{G}}$ in the quasi-local case due to its complexity. Instead, we will study one practically-interesting examples in the following section.

B.6 Case study: CZ gates with 1D topology

In this section, we will apply our theory to a concrete gate set where the Control-Z (CZ) gates are the only multi-qubit entangling gates. Such a gate set, augmented with single-qubit gates, are able to conduct universal quantum computation, and is adopted by state-of-the-art experimental platforms (e.g., IBM Quantum [ibm22][5]). We will discuss the noise learnability and learning algorithms with several different noise ansatz. We hope the results presented here can not only illustrate our theoretical results, but also serve as a new noise characterization protocol for practical use. A quick summary of results is given in Table B.2.

Section	n	$\mathfrak{G}$	Noise model	$\dim(X_R)$	$\dim(L_R)$	$\dim(X_R^G)$	$\dim(L_R^G)$
B.6.1	2	$\{\mathrm{CZ}\}$	complete	21	18	15	13
B.6.2	$n \geq 4$	$\{\mathrm{CZ}_{i,i+1}\}$	fully local	$17n$	$16n$	$15n$	$14n$
B.6.3	$n \geq 6$	$\{\mathrm{CZ}_e, \mathrm{CZ}_o\}$	nearest-neighbor	$28n$	$27n$	$24n$	$23n$
B.6.4	$n \geq 6$	$\{\mathrm{CZ}_e, \mathrm{CZ}_o\}$	covariant 4-local	$244n$	$242n$	$240n$	–

Table B.2: Summary of results of this section. Each row corresponds to a gate set and noise model studied in one subsection.

5. Note that the Control-NOT (CNOT) and the Echoed Cross-Resonance (ECR) gates are all equivalent to CZ up to single-qubit rotations. Our theory can similarly be applied to them.

B.6.1 Single CZ, complete noise

We begin with a minimal example: $n = 2$, $\mathfrak{G} = \{\mathrm{CZ}\}$. The goal is to learn the complete Pauli noise model. Recall that the associated PTG for this model is given in Fig. B.1. Thanks to Theorem 2.2, the parameter space X, gauge space T, and learnable space L are given by (let $\mathcal{G} = \mathrm{CZ}$ for notational simplicity),

- X: $\dim(X) = 21$. Basis: $\{\boldsymbol{e}_u^S, \boldsymbol{e}_u^M, \boldsymbol{e}_a^{\mathcal{G}} : u \neq 00, a \neq II\}$.
- T: $\dim(T) = 3$. Basis: $\{\mathfrak{d}_{10}, \mathfrak{d}_{01}, \mathfrak{d}_{11}\}$.
- L: $\dim(L) = 18$. Cycle basis: $\left\{\boldsymbol{e}_u^S + \boldsymbol{e}_u^M : u \in \{01, 10, 11\}\right\} \cup \{\boldsymbol{e}^S_{\mathrm{pt}(a)} + \boldsymbol{e}_a^{\mathcal{G}} + \boldsymbol{e}^M_{\mathrm{pt}(\mathcal{G}(a))} : a \in \mathsf{P}^2 \backslash II\}$.

The learnable space for gate noise $L^G \coloneqq L \cap X^G$ (which is the cycle space for the subgraph of PTG with edges only for gate noise parameters) is spanned by the following cycle basis

$$\begin{aligned} &\left\{\boldsymbol{e}_{IZ}^{\mathcal{G}}, \boldsymbol{e}_{ZI}^{\mathcal{G}}, \boldsymbol{e}_{ZZ}^{\mathcal{G}}, \boldsymbol{e}_{XX}^{\mathcal{G}}, \boldsymbol{e}_{YY}^{\mathcal{G}}, \boldsymbol{e}_{XY}^{\mathcal{G}}, \boldsymbol{e}_{YX}^{\mathcal{G}},\right. \\ &\left.\boldsymbol{e}_{XI}^{\mathcal{G}} + \boldsymbol{e}_{XZ}^{\mathcal{G}}, \boldsymbol{e}_{YI}^{\mathcal{G}} + \boldsymbol{e}_{YZ}^{\mathcal{G}}, \boldsymbol{e}_{XI}^{\mathcal{G}} + \boldsymbol{e}_{YZ}^{\mathcal{G}}, \boldsymbol{e}_{IX}^{\mathcal{G}} + \boldsymbol{e}_{ZX}^{\mathcal{G}}, \boldsymbol{e}_{IY}^{\mathcal{G}} + \boldsymbol{e}_{ZY}^{\mathcal{G}}, \boldsymbol{e}_{IX}^{\mathcal{G}} + \boldsymbol{e}_{ZY}^{\mathcal{G}}\right\}. \end{aligned} \tag{B.68}$$

We have $\dim(L^G) = 13$. Note that $\dim(X^G) - \dim(L^G) = 2$. This is because $\pi^{\mathcal{G}}(\mathfrak{d}_{11}) = \mathbf{0}$, which means the global depolarizing gauge does not change gate noise parameters. Only the two single-qubit depolarizing gauges $\{\mathfrak{d}_{01}, \mathfrak{d}_{10}\}$ changes gate parameters nontrivially. The above basis for L^G can be augmented to a basis of L by adding the following 5 rooted cycles,

$$\left\{\boldsymbol{e}_u^S + \boldsymbol{e}_u^M : u \in \{01, 10, 11\}\right\} \cup \left\{\boldsymbol{e}_{10}^S + \boldsymbol{e}_{XI}^{\mathcal{G}} + \boldsymbol{e}_{11}^M, \boldsymbol{e}_{01}^S + \boldsymbol{e}_{IX}^{\mathcal{G}} + \boldsymbol{e}_{11}^M\right\} \tag{B.69}$$

We remark that Eq. (B.68) was previously given in [CLO+23]. The new insight from our theory is to include SPAM parameters as the target of learning, and to explicitly parameterize the gauge degrees of freedom. We also note that [CLO+23] has presented experimental designs to learn all of these gate parameters to relative precision. Concretely, any of the

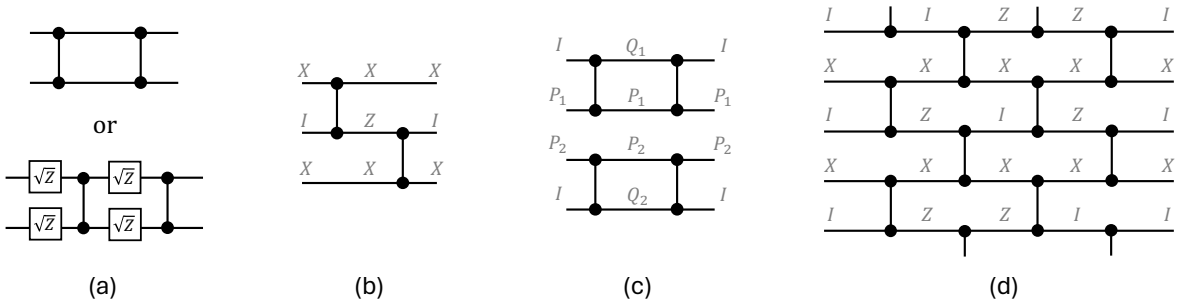


Figure B.8: Circuits for learning different types of cycles. (a) Circuits for learning all cycles of a single CZ (and for certain cycles of the other models). Any cycles from Eq. (B.68) can be learned by inputting an appropriate Pauli eigenstate to the upper or lower circuit and measure the same Pauli observable. (b) Circuits for learning the 2-gate cycle from Eq. (B.71). The evolution of the corresponding Pauli operators is shown in gray. (c) Circuits for learning cycles from Eq. (B.75). (d) Circuits for learning for cycles from Eq. (B.76).

above parameters can be learned by concatenating CZ or $\mathrm{CZ} \circ (\sqrt{Z} \otimes \sqrt{Z})$ an even number of time, input the correct Pauli eigenstate, and measurement in the correct Pauli eigenbasis. See Fig. B.8 (a).

B.6.2 CZ's on 1D ring, fully local noise

Let $n > 2$, consider the case where the n qubits are placed in an 1D ring, and one can apply an CZ between any neighboring pair of qubits, i.e., $\mathfrak{G} = \{\mathrm{CZ}_{i,i+1} : i \in [n]\}$, where the $(i+1)$-th qubit is the same as the first qubit. Suppose the noise model is fully local (c.f. Definition 2.24), using Theorem 2.5, we obtain the following characterization for the learnability of this model (let $\mathcal{G}_i = \mathrm{CZ}_{i,i+1}$),

- X_R: $\dim(X_R) = 17n$. Basis: $\{\boldsymbol{\vartheta}^S_j, \boldsymbol{\vartheta}^M_j, \boldsymbol{\vartheta}^{\mathcal{G}_i}_a : j \in [n], i \in [n], a \in \mathsf{P}^2 \backslash II\}$.
- T_R: $\dim(T_R) = n$. Basis: $\{\mathcal{Q}^{-1}(\mathfrak{d}_{\{i\}}) : i \in [n]\}$.
- L_R: $\dim(L_R) = 16n$. Corresponding cycle basis: $\{\boldsymbol{e}^S_{\{i\}} + \boldsymbol{e}^M_{\{i\}} : i \in [n]\} \cup \{\boldsymbol{e}^S_{\mathrm{pt}(a)} + \boldsymbol{e}^{\mathcal{G}}_a + \boldsymbol{e}^M_{\mathrm{pt}(\mathcal{G}_i(a))} : i \in [n], a \in \mathsf{P}^{\{i,i+1\}} \backslash I_n\}$.

Note that $\mathcal{Q}^{-1}$ is well-defined when acting within the image of $\mathcal{Q}$. For L_R, instead of

writing down a reduced cycle basis, we write down the corresponding cycle basis in L whose connection to experiment design is more direct. Acting $\mathcal{Q}^{\mathsf{T}}$ on that basis will give a reduced cycle basis for L_R.

Using the algorithms introduced in Sec. B.4.3, we can construct a reduced cycle basis for the learnable spaces of gate parameters L_R^G. In fact, there exists a basis where each basis vector contains parameters of only one gate or two neighboring gates, as follows

- 1-gate (same as Eq. (B.68)):

$$\Big\{\boldsymbol{\vartheta}_{IZ}^{\mathcal{G}_i}, \boldsymbol{\vartheta}_{ZI}^{\mathcal{G}_i}, \boldsymbol{\vartheta}_{ZZ}^{\mathcal{G}_i}, \boldsymbol{\vartheta}_{XX}^{\mathcal{G}_i}, \boldsymbol{\vartheta}_{YY}^{\mathcal{G}_i}, \boldsymbol{\vartheta}_{XY}^{\mathcal{G}_i}, \boldsymbol{\vartheta}_{YX}^{\mathcal{G}_i}, \\ \boldsymbol{\vartheta}_{XI}^{\mathcal{G}_i} + \boldsymbol{\vartheta}_{XZ}^{\mathcal{G}_i}, \boldsymbol{\vartheta}_{YI}^{\mathcal{G}_i} + \boldsymbol{\vartheta}_{YZ}^{\mathcal{G}_i}, \boldsymbol{\vartheta}_{XI}^{\mathcal{G}_i} + \boldsymbol{\vartheta}_{YZ}^{\mathcal{G}_i}, \boldsymbol{\vartheta}_{IX}^{\mathcal{G}_i} + \boldsymbol{\vartheta}_{ZX}^{\mathcal{G}_i}, \boldsymbol{\vartheta}_{IY}^{\mathcal{G}_i} + \boldsymbol{\vartheta}_{ZY}^{\mathcal{G}_i}, \boldsymbol{\vartheta}_{IX}^{\mathcal{G}_i} + \boldsymbol{\vartheta}_{ZY}^{\mathcal{G}_i} : i \in [n]\Big\}. \tag{B.70}$$

- 2-gate:

$$\left\{\boldsymbol{\vartheta}_{XI}^{\mathcal{G}_i} + \boldsymbol{\vartheta}_{ZX}^{\mathcal{G}_{i+1}} : i \in [n]\right\}. \tag{B.71}$$

We have $\dim(L^G) = 14n$. The 1-gate basis vector can be learned by running the same learning protocol for a single CZ in each consecutive 2-qubit subsystem, shown in Fig. B.8 (a); The 2-gate basis can be learned using the circuits shown in Fig. B.8 (b) on each consecutive 3-qubit subsystem. By learning disjoint regions in parallel, a constant number of experiments is sufficient to learn all the gate noise parameters. We also note that the above basis for L_R^G can be augmented to a basis for L_R by adding the following $2n$ reduced cycles,

$$\left\{\boldsymbol{\vartheta}_j^S + \boldsymbol{\vartheta}_j^M : j \in [n]\right\} \cup \left\{\boldsymbol{\vartheta}_j^S + \boldsymbol{\vartheta}_{XI}^{\mathcal{G}_j} + \boldsymbol{\vartheta}_j^M + \boldsymbol{\vartheta}_{j+1}^M : j \in [n]\right\}. \tag{B.72}$$

B.6.3 CZ's on ring, nearest-neighbor noise

Let $n \geq 6$ be an even number. We again consider n qubits placed in an 1D ring, but this time let the gate set be $\mathfrak{G} = \{\mathcal{G}_e, \mathcal{G}_o\}$ where

$$\mathcal{G}_e = \mathrm{CZ}_{1,2} \otimes \mathrm{CZ}_{3,4} \otimes \cdots \otimes \mathrm{CZ}_{n-1,n}, \quad \mathcal{G}_o = \mathrm{CZ}_{2,3} \otimes \mathrm{CZ}_{4,5} \otimes \cdots \otimes \mathrm{CZ}_{n,1}, \tag{B.73}$$

and we assume the state preparation, measurement, and gate noise for both $\mathcal{G}_e, \mathcal{G}_o$ are all Ω-local, with Ω associated with $\Omega_* = \{\{i, i+1\} : i \in [n]\}$. The setup is illustrated in Fig. B.9. This setup is essentially the one used in [KEA+23] for conducting an error-mitigated quantum simulation task. It is thus of great interest to understand the learnability of this noise model.

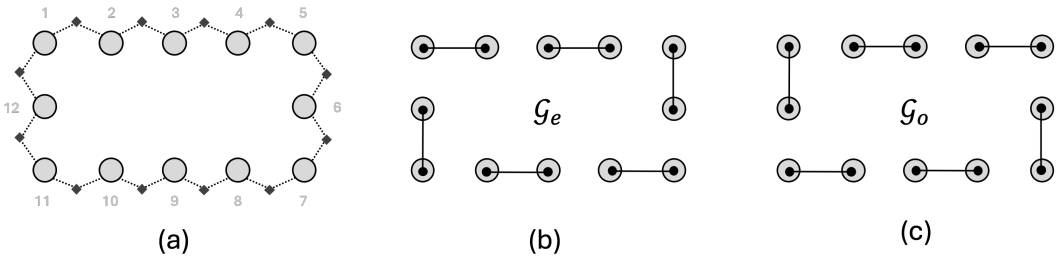


Figure B.9: Illustration of the setup in Sec. B.6.3 with $n = 12$. (a) Qubit topology and noise model. Each circle represents a qubit with their label annotated. Each diamond connects to a maximal factor from Ω_*, specifying the nearest-neighbor noise model for both SPAM and gate noises. (b), (c) Illustration of $\mathcal{G}_e, \mathcal{G}_o$, respectively.

To begin with, this noise model is not gate-covariant in the sense of Definition 2.32. To see this, consider the Pauli channel $\Lambda'(\rho) = \eta X_2 X_3 \rho X_2 X_3 + (1-\eta)\rho$ with some $\eta < 1/2$, which is Ω-local thanks to Lemma 2.36. We have $\mathcal{G}_e \Lambda' \mathcal{G}_e^\dagger = \eta Z_1 X_2 X_3 Z_4 \rho Z_1 X_2 X_3 Z_4 + (1-\eta)\rho$, which is not Ω-local, again by Lemma 2.36. Similarly for $\mathcal{G}_o$. This means we cannot directly apply Theorem 2.6(c). In fact, we can show that the reduced gauge space is given by $T_R = \mathrm{span}\{\mathcal{Q}^{-1}(\mathfrak{d}_{\{j\}}) : j \in [n]\}$, i.e., only the single-qubit depolarizing gauges. Intuitively, any (linear combination of) two-qubit depolarizing gauge transformation will violate the Ω-local noise assumption of either $\mathcal{G}_e$ or $\mathcal{G}_o$. A formal proof is given in Appendix B.7.5.

Therefore, we obtain the following characterization of the model learnability,

- X_R: $\dim(X_R) = 28n$. Basis: $\{\boldsymbol{\vartheta}_\nu^S, \boldsymbol{\vartheta}_\nu^M, \boldsymbol{\vartheta}_a^{\mathcal{G}_e}, \boldsymbol{\vartheta}_a^{\mathcal{G}_o} : \nu \in \Omega, a \sim \Omega\}$.
- T_R: $\dim(T_R) = n$. Basis: $\{\mathcal{Q}^{-1}(\mathfrak{d}_{\{i\}}) : i \in [n]\}$.
- L_R: $\dim(L_R) = 27n$.

Note that the cardinality of $\{a \in \mathsf{P}^n \backslash I_n : a \sim \Omega\}$ can be computed using the inclusion-exclusion principle: $(4^2 - 1)n - (4 - 1)n = 12n$.

We have also found a reduced cycle basis for L_R^G. To make the connection to experiments more straightforward, we will write down a cycle basis that generates the reduced cycle basis by applying $\mathcal{Q}^{\mathsf{T}}$ on them (c.f. Definition 2.20). The cycle bases are divided into two parts, depending on whether the basis vector involves one or two gates, as follows,

- 1-gate:
  - For $k = 1, ..., \frac{n}{2}$, similar as Eq. (B.68),

$$\begin{aligned} &\{\boldsymbol{e}^{\mathcal{G}_e}_{I_{2k-1}Z_{2k}}, \cdots, \boldsymbol{e}^{\mathcal{G}_e}_{X_{2k-1}I_{2k}} + \boldsymbol{e}^{\mathcal{G}_e}_{X_{2k-1}Z_{2k}}, \cdots\} \\ \cup &\{\boldsymbol{e}^{\mathcal{G}_o}_{I_{2k}Z_{2k+1}}, \cdots, \boldsymbol{e}^{\mathcal{G}_o}_{X_{2k}I_{2k+1}} + \boldsymbol{e}^{\mathcal{G}_o}_{X_{2k}Z_{2k+1}}, \cdots\} \end{aligned} \tag{B.74}$$

  - For $k = 1, ..., \frac{n}{2}$,

$$\begin{aligned} &\{\boldsymbol{e}^{\mathcal{G}_e}_a + \boldsymbol{e}^{\mathcal{G}_e}_{a'} : a \in \mathsf{P}^{\{2k,2k+1\}}, w(a) = 2, a' = \mathcal{G}_e(a)\} \\ \cup &\{\boldsymbol{e}^{\mathcal{G}_o}_a + \boldsymbol{e}^{\mathcal{G}_o}_{a'} : a \in \mathsf{P}^{\{2k+1,2k+2\}}, w(a) = 2, a' = \mathcal{G}_o(a)\} \end{aligned} \tag{B.75}$$

- 2-gate: For $k = 1, \cdots, \frac{n}{2}$,

$$\begin{aligned} &\left\{\boldsymbol{e}^{\mathcal{G}_o}_{IXIXI_{\{2k-1,\cdots,2k+3\}}} + \boldsymbol{e}^{\mathcal{G}_e}_{IXZXZ_{\{2k-1,\cdots,2k+3\}}} + \boldsymbol{e}^{\mathcal{G}_o}_{ZXIXZ_{\{2k-1,\cdots,2k+3\}}} + \boldsymbol{e}^{\mathcal{G}_e}_{ZXZXI_{\{2k-1,\cdots,2k+3\}}}\right\} \\ \cup &\left\{\boldsymbol{e}^{\mathcal{G}_o}_{IXIXI_{\{2k,\cdots,2k+4\}}} + \boldsymbol{e}^{\mathcal{G}_e}_{ZXZXI_{\{2k,\cdots,2k+4\}}} + \boldsymbol{e}^{\mathcal{G}_o}_{ZXIXZ_{\{2k,\cdots,2k+4\}}} + \boldsymbol{e}^{\mathcal{G}_e}_{IXZXZ_{\{2k,\cdots,2k+4\}}}\right\} \end{aligned} \tag{B.76}$$

Eqs. (B.74), (B.75), (B.76) contributes $13n$, $9n$, n cycles, respectively, and is consistent with the fact that $\dim(L_R^G) = 23n$. Elusive as they might look, each type of cycles corresponds to an explicit family of experiment designs, as shown in Fig. B.8 (a), (c), (d), respectively.

The proof that the above cycles constitute a cycle basis for L_R^G and a way to supplement it into a complete basis for L_R can be found in the Appendix of [CZJF24], and is omitted in this thesis.

B.6.4 CZ's on ring, covariant quasi-local noise

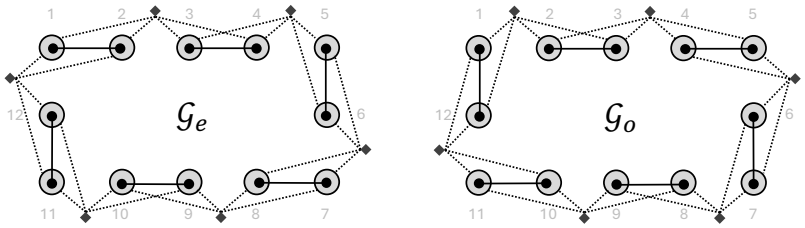


Figure B.10: The covariant quasi-local gate noise model from Sec. B.6.4. The left and right shows Ω^e, Ω^o, respectively, with each diamond connecting to a maximal factor from Ω_*^e (resp. Ω_*^o).

Finally, we present a minimal example of quasi-local noise model that is gate-covariant. Let $n \geq 6$ be even. Consider the same gate set $\mathfrak{G} = \{\mathcal{G}_e, \mathcal{G}_o\}$ as in the last section. Assume the SPAM noise is still Ω-local for $\Omega_* = \{\{i, i+1\} : i \in [n]\}$. Assume the noise for $\mathcal{G}_e$, $\mathcal{G}_o$ is $\Omega^{\mathcal{G}_e}$, $\Omega^{\mathcal{G}_o}$-local, respectively, where

$$\begin{aligned} \Omega_*^{\mathcal{G}_e} &= \{\{2k-1, 2k, 2k+1, 2k+2\} : k \in [n/2]\}, \\ \Omega_*^{\mathcal{G}_o} &= \{\{2k, 2k+1, 2k+2, 2k+3\} : k \in [n/2]\}. \end{aligned} \tag{B.77}$$

The factor graphs for each noise model are given in Fig. B.10. The fact that this model is gate-covariant follows from Lemma 2.33. Indeed, each maximal factor is weight-4, containing exactly the support of two neighboring CZ. Consequently, Theorem 2.6 gives us that

- X_R: $\dim(X_R) = 244n$. Basis: $\{\boldsymbol{\vartheta}_\nu^S, \boldsymbol{\vartheta}_\nu^M, \boldsymbol{\vartheta}_a^{\mathcal{G}_e}, \boldsymbol{\vartheta}_b^{\mathcal{G}_o} : \nu \in \Omega, a \sim \Omega^{\mathcal{G}_e}, b \sim \Omega^{\mathcal{G}_o}\}$.

- T_R: $\dim(T_R) = 2n$. Basis: $\{\mathcal{Q}^{-1}(\mathfrak{d}_\nu) : \nu \in \Omega\}$.
- L_R: $\dim(L_R) = 242n$. Corresponding cycle basis: $\{\boldsymbol{e}^S_\nu + \boldsymbol{e}^M_\nu : \nu \in \Omega\} \cup \{\boldsymbol{e}^S_{\mathrm{pt}(a)} + \boldsymbol{e}^{\mathcal{G}}_a + \boldsymbol{e}^M_{\mathrm{pt}(\mathcal{G}(a))} : \mathcal{G} \in \{\mathcal{G}_e, \mathcal{G}_o\}, a \sim \Omega^{\mathcal{G}}\}$.

Note that the cardinality of $\{a \in \mathsf{P}^n \backslash I_n : a \sim \Omega^{\mathcal{G}_e}\}$ can again be computed using the inclusion-exclusion principle: $\frac{n}{2} \times (4^4 - 4^2) = 120n$. One can see that a linear number of experiments that involves measuring at most weight-4 Pauli observables is sufficient to learn the whole noise model. It is also not hard to see that, by paralleling many of the experiments, a constant number of experiments would also suffice.

We will not pursue a reduced cycle basis for L_R^G in this example due to its complexity, and leave this for future investigation.

B.7 Technical Details

B.7.1 Proof of Theorem 2.1

The proof follows [CLO$^+$23, App. D] with some modification to fit in our framework.

Proof of Theorem 2.1(a). For any experiment F, we need to find a finite set of rooted cycles (*i.e.*, linear functions corresponding to a rooted cycle) $\{f_j\}_{j=1}^M$ and a function $\widehat{F}$ such that

$$F(\boldsymbol{x}) = \widehat{F}(f_1(\boldsymbol{x}), \cdots, f_M(\boldsymbol{x})), \quad \forall \boldsymbol{x} \in X. \tag{B.78}$$

By definition, $F(\boldsymbol{x})$ is a 2^n-dimensional real vector such that $F_k(\boldsymbol{x}) = \mathrm{Tr}[\widetilde{E}_k \widetilde{\mathcal{C}}(\widetilde{\rho}_0)]$ where the noisy initial state, circuit, and POVM element are determined by the noise parameter $\boldsymbol{x}$. Expanding the SPAM noise parameters yields

$$F_k(\boldsymbol{x}) = \frac{1}{4^n} \sum_{u,u' \in \{0,1\}^n} (-1)^{k \cdot u'} \lambda^M_{u'} \lambda^S_u \,\mathrm{Tr}\left[Z^{u'} \widetilde{\mathcal{C}}(Z^u)\right]. \tag{B.79}$$

On the other hand, any noisy circuit $\mathcal{C}$ satisfying our assumptions can be written as

$$\widetilde{\mathcal{C}} = \mathcal{U}^{(m)} \circ \widetilde{\mathcal{G}}_m \circ \cdots \circ \mathcal{U}^{(1)} \circ \widetilde{\mathcal{G}}_1 \circ \mathcal{U}^{(0)}, \tag{B.80}$$

where each $\mathcal{U}^{(i)} = \bigotimes_{l=1}^{n} \mathcal{U}_l^{(i)}$ is a layer of parallel single-qubit gate, and each $\widetilde{\mathcal{G}}_i$ is the noisy realization of a Clifford gate $\mathcal{G}_i$ from our gate set $\mathfrak{G}$. Note that consecutive layers of single-qubit gates can always be merged into a single layer, and we have assumed single-qubit gate to be noiseless, so the above form of $\widetilde{\mathcal{C}}$ is indeed general.

An important property for single-qubit gate layers is that, even if they are non-Clifford, they always preserve the pattern of an input Pauli operator. That is,

$$\mathcal{U}^{(i)}(P_a) = \sum_{b:\mathrm{pt}(b)=\mathrm{pt}(a)} c_{b,a}^{(i)} P_b, \quad \forall P_a \in \mathsf{P}^n, \tag{B.81}$$

for some real coefficients $c_{b,a}^{(i)}$. This comes from the fact that any single-qubit unitary channel is trace-preserving and unital. Thanks to this property, the action of $\widetilde{\mathcal{C}}$ on any input Pauli

operator can be expressed as

$$
\begin{aligned}
\widetilde{\mathcal{C}}(P_a) &= (\mathcal{U}^{(m)} \circ \widetilde{\mathcal{G}}_m \circ \cdots \circ \mathcal{U}^{(1)} \circ \widetilde{\mathcal{G}}_1 \circ \mathcal{U}^{(0)})(P_a) \\
&= (\mathcal{U}^{(m)} \circ \widetilde{\mathcal{G}}_m \circ \cdots \circ \mathcal{U}^{(1)} \circ \widetilde{\mathcal{G}}_1) \left(\sum_{b_0:\mathrm{pt}(b_0)=\mathrm{pt}(a)} c^{(0)}_{b_0,a} P_{b_0} \right) \\
&= (\mathcal{U}^{(m)} \circ \widetilde{\mathcal{G}}_m \circ \cdots \circ \mathcal{U}^{(1)}) \left(\sum_{b_0:\mathrm{pt}(b_0)=\mathrm{pt}(a)} c^{(0)}_{b_0,a} \lambda^{\mathcal{G}_1}_{b_0} P_{\mathcal{G}_1(b_0)} \right) \\
&= (\mathcal{U}^{(m)} \circ \widetilde{\mathcal{G}}_m \circ \cdots \circ \mathcal{U}^{(2)}) \left(\sum_{\substack{b_0:\mathrm{pt}(b_0)=\mathrm{pt}(a),\\ b_1:\mathrm{pt}(b_1)=\mathrm{pt}(\mathcal{G}_1(b_0))}} c^{(1)}_{b_1,\mathcal{G}_1(b_0)} c^{(0)}_{b_0,a} \lambda^{\mathcal{G}_2}_{b_1} \lambda^{\mathcal{G}_1}_{b_0} P_{\mathcal{G}_2(b_1)} \right) \\
&= \cdots \\
&= \sum_{\substack{b_0:\mathrm{pt}(b_0)=\mathrm{pt}(a),\\ b_1:\mathrm{pt}(b_1)=\mathrm{pt}(\mathcal{G}_1(b_0)),\\ \cdots \\ b_m:\mathrm{pt}(b_m)=\mathrm{pt}(\mathcal{G}_m(b_{m-1}))}} c^{(m)}_{b_m,\mathcal{G}_m(b_{m-1})} \cdots c^{(1)}_{b_1,\mathcal{G}_1(b_0)} c^{(0)}_{b_0,a} \lambda^{\mathcal{G}_m}_{b_{m-1}} \cdots \lambda^{\mathcal{G}_2}_{b_1} \lambda^{\mathcal{G}_1}_{b_0} P_{b_m}.
\end{aligned}
\tag{B.82}
$$

Now, substitute this back into the expression of $F_k(\boldsymbol{x})$,

$$
\begin{aligned}
& F_k(\boldsymbol{x}) \\
=& \sum_{\substack{u,u' \in \{0,1\}^n,\\ b_0:\mathrm{pt}(b_0)=\mathrm{pt}(u),\\ b_1:\mathrm{pt}(b_1)=\mathrm{pt}(\mathcal{G}_1(b_0)),\\ \cdots \\ b_m:\mathrm{pt}(b_m)=\mathrm{pt}(\mathcal{G}_m(b_{m-1}))}} \frac{(-1)^{k\cdot u'}}{4^n} \, c^{(m)}_{b_m,\mathcal{G}_m(b_{m-1})} \cdots c^{(1)}_{b_1,\mathcal{G}_1(b_0)} c^{(0)}_{b_0,a} \lambda^{M}_{u'} \lambda^{\mathcal{G}_m}_{b_{m-1}} \cdots \lambda^{\mathcal{G}_2}_{b_1} \lambda^{\mathcal{G}_1}_{b_0} \lambda^{S}_{u} \operatorname{Tr}\left[Z^{u'} P_{b_m} \right] \\
=& \sum_{\substack{u \in \{0,1\}^n,\\ b_0:\mathrm{pt}(b_0)=\mathrm{pt}(u),\\ b_1:\mathrm{pt}(b_1)=\mathrm{pt}(\mathcal{G}_1(b_0)),\\ \cdots \\ b_m:\mathrm{pt}(b_m)=\mathrm{pt}(\mathcal{G}_m(b_{m-1}))}} \frac{(-1)^{k\cdot \mathrm{pt}(b_m)}}{2^n} \, c^{(m)}_{b_m,\mathcal{G}_m(b_{m-1})} \cdots c^{(0)}_{b_0,a} \lambda^{M}_{\mathrm{pt}(b_m)} \lambda^{\mathcal{G}_m}_{b_{m-1}} \cdots \lambda^{\mathcal{G}_1}_{b_0} \lambda^{S}_{u} \, \mathbb{1}[Z^{\mathrm{pt}(b_m)} = P_{b_m}].
\end{aligned}
\tag{B.83}
$$

Let us look at the indexes of the summation. For $u = \mathbf{0}$, we must have $b_m = \cdots = b_0 = I$, and the corresponding term in the sum becomes $1/2^n$ that is a constant; For $u \neq \mathbf{0}$, we must have $b_i \neq I$ for all i, and the corresponding term in the sum depends on the noise parameters

via

$$\begin{aligned}\lambda^{M}_{\mathrm{pt}(b_m)}\lambda^{\mathcal{G}_m}_{b_{m-1}}\cdots\lambda^{\mathcal{G}_2}_{b_1}\lambda^{\mathcal{G}_1}_{b_0}\lambda^{S}_{u} &= \exp\Big(-\Big(x^{M}_{\mathrm{pt}(b_m)}+x^{\mathcal{G}_m}_{b_{m-1}}+\cdots+x^{\mathcal{G}_1}_{b_0}+x^{S}_{u}\Big)\Big)\\ &=: \exp\Big(-f_{\boldsymbol{b},u}(\boldsymbol{x})\Big),\end{aligned} \tag{B.84}$$

where $f_{\boldsymbol{b},u}$ is a linear function of $\boldsymbol{x}$. What's more, thanks to the restriction of $(b_0,\cdots,b_m)$, one can see that $f_{\boldsymbol{b},u}$ is actually a rooted cycle in the pattern transfer graph. Therefore, $F_k(\boldsymbol{x})$ (and thus $F(\boldsymbol{x})$) is indeed determined by $\{f_{\boldsymbol{b},u}(\boldsymbol{x})\}_{\boldsymbol{b},u}$ consisting of finitely many rooted cycles. This completes the proof of Theorem 2.1(a). □

The following proof again comes from [CLO$^+$23] with minor modification, and is presented here for completeness.

Proof of Theorem 2.1(b). For any rooted cycles f, we need to find an experiment F and a function $\widehat{f}$ such that

$$f(\boldsymbol{x})=\widehat{f}(F(\boldsymbol{x})),\quad \forall \boldsymbol{x}\in X. \tag{B.85}$$

A generic rooted cycle f of depth m takes the following form (here, we write down the vertices for clarity, though they are not needed to define a cycle),

$$\begin{aligned}&(v_R,x^{S}_{b_0},v_{b_0},x^{\mathcal{G}_1}_{P_1},v_{b_1},x^{\mathcal{G}_2}_{P_2},\cdots,x^{\mathcal{G}_m}_{P_m},v_{b_m},x^{M}_{b_m},v_R),\\ &\text{s.t.}\quad b_{i-1}=\mathrm{pt}(P_i),\ \mathrm{pt}(\mathcal{G}_i(P_i))=b_i,\ \forall i=1,\cdots,m.\end{aligned} \tag{B.86}$$

Consider the following noisy circuit

$$\widetilde{\mathcal{C}}=\mathcal{U}^{(m)}\circ\widetilde{\mathcal{G}}_m\circ\cdots\circ\mathcal{U}^{(1)}\circ\widetilde{\mathcal{G}}_1\circ\mathcal{U}^{(0)},$$

$$\text{where } \mathcal{U}^{(i)}\equiv\bigotimes_{j=1}^{n}\mathcal{U}^{(i)}_j \text{ is Clifford, and satisfies } \begin{cases}\mathcal{U}^{(0)}(Z^{b_0})=P_1;\\ \mathcal{U}^{(i)}(\mathcal{G}_i(P_i))=P_{i+1},\ \forall i=1,\cdots,m-1;\\ \mathcal{U}^{(m)}(\mathcal{G}_m(P_m))=Z^{b_m}.\end{cases} \tag{B.87}$$

Note that, thanks to the constraints from Eq. (B.86), the above requirements for $\mathcal{U}^{(i)}$ are literally mapping one Pauli operator to another *with the same pattern*, thus can indeed be achieved by a layer of single-qubit Clifford gates.

This circuit yields the following experiment (*i.e.*, experimental outcome distribution)

$$F_k(\boldsymbol{x}) = \mathrm{Tr}[\widetilde{E}_k \widetilde{\mathcal{C}}(\widetilde{\rho}_0)] = \frac{1}{4^n} \sum_{u,u' \in \{0,1\}^n} (-1)^{k \cdot u'} \lambda_{u'}^M \lambda_u^S \,\mathrm{Tr}\left[Z^{u'} \widetilde{\mathcal{C}}(Z^u) \right], \quad \forall k \in \{0,1\}^n. \tag{B.88}$$

Compute the following function of experimental outcome (which is, intuitively, the expectation value of Z^{b_m} on the equivalent noisy state right before a perfect measurement)

$$\begin{aligned}
\sum_{k \in \{0,1\}^n} (-1)^{k \cdot b_m} F_k(\boldsymbol{x}) &= \frac{1}{4^n} \sum_{k,u,u'} (-1)^{k \cdot (b_m + u')} \lambda_{u'}^M \lambda_u^S \,\mathrm{Tr}\left[Z^{u'} \widetilde{\mathcal{C}}(Z^u) \right] \\
&= \frac{1}{2^n} \sum_u \lambda_{b_m}^M \lambda_u^S \,\mathrm{Tr}\left[Z^{b_m} \widetilde{\mathcal{C}}(Z^u) \right] \\
&= \frac{1}{2^n} \sum_u \lambda_{b_m}^M \lambda_{P_m}^{\mathcal{G}_m} \cdots \lambda_{P_1}^{\mathcal{G}_1} \lambda_u^S \,\mathrm{Tr}\left[Z^{b_0} Z^u \right] \\
&= \lambda_{b_m}^M \lambda_{P_m}^{\mathcal{G}_m} \cdots \lambda_{P_1}^{\mathcal{G}_1} \lambda_{b_0}^S \\
&= \exp\left(-\left(x_{b_0}^S + x_{P_1}^{\mathcal{G}_1} + \cdots + x_{P_m}^{\mathcal{G}_m} + x_{b_m}^M \right) \right) \equiv \exp\left(-f(\boldsymbol{x}) \right).
\end{aligned} \tag{B.89}$$

The third line is obtained by acting $\widetilde{\mathcal{C}}^\dagger$ on Z^{b_m} inside the trace, and then evolve Z^{b_m} backwards according to the definition of $\widetilde{\mathcal{C}}$ from Eq. (B.87). As a result, we have

$$f(\boldsymbol{x}) = -\log\left(\sum_{k \in \{0,1\}^n} (-1)^{k \cdot b_m} F_k(\boldsymbol{x}) \right). \tag{B.90}$$

This completes the proof of Theorem 2.1(b). Recall that this means any rooted cycle f can be learned from a single experiment F. □

B.7.2 Proof of Theorem 2.3

Proof. We first show that $T_R = \mathrm{Ker}(\mathcal{P} \circ \mathcal{Q})$. For any experiment F, Theorem 2.1(b) implies that there exists some function $\widehat{F}$ such that $F(\mathcal{Q}(\boldsymbol{r})) = \widehat{F}(\mathcal{P} \circ \mathcal{Q}(\boldsymbol{r}))$ for all $\boldsymbol{r} \in X_R$. Thus, for any $\boldsymbol{t} \in \mathrm{Ker}(\mathcal{P} \circ \mathcal{Q})$,

$$F(\mathcal{Q}(\boldsymbol{r} + \eta\boldsymbol{t})) = \widehat{F}(\mathcal{P} \circ \mathcal{Q}(\boldsymbol{r} + \eta\boldsymbol{t})) = \widehat{F}(\mathcal{P} \circ \mathcal{Q}(\boldsymbol{r})) = F(\mathcal{Q}(\boldsymbol{r})), \quad \forall \boldsymbol{r} \in X_R, \ \forall \eta \in \mathbb{R}. \tag{B.91}$$

The second equality uses linearity of $\mathcal{P} \circ \mathcal{Q}$. The above means $\boldsymbol{t} \in T_R$ by definition. Therefore, $\mathrm{Ker}(\mathcal{P} \circ \mathcal{Q}) \subseteq T_R$; On the other hand, for any $\boldsymbol{t} \notin \mathrm{Ker}(\mathcal{P} \circ \mathcal{Q})$, $\mathcal{Q}(\boldsymbol{t}) \notin \mathrm{Ker}(\mathcal{P})$, which means there is at least one f_i such that $f_i(\mathcal{Q}(\boldsymbol{t})) \neq 0$. Now, Theorem 2.1(b) says that there exists an experiment F_i and a function $\widehat{f_i}$ such that $f_i(\mathcal{Q}(\boldsymbol{t})) = \widehat{f_i}(F_i(\mathcal{Q}(\boldsymbol{t})))$ for all $\boldsymbol{r}$. Thus

$$\begin{aligned} f_i(\mathcal{Q}(\boldsymbol{t})) \neq 0 &\Rightarrow f_i(\mathcal{Q}(\boldsymbol{r})) \neq f_i(\mathcal{Q}(\boldsymbol{r} + \boldsymbol{t})), \quad \forall \boldsymbol{r} \in X_R, \\ &\Rightarrow \widehat{f_i}(F(\mathcal{Q}(\boldsymbol{r}))) \neq \widehat{f_i}(F(\mathcal{Q}(\boldsymbol{r} + \boldsymbol{t}))), \quad \forall \boldsymbol{r} \in X_R, \\ &\Rightarrow F(\mathcal{Q}(\boldsymbol{r})) \neq F(\mathcal{Q}(\boldsymbol{r} + \boldsymbol{t})), \quad \forall \boldsymbol{r} \in X_R, \end{aligned} \tag{B.92}$$

where the first line uses linearity of f_i and $\mathcal{Q}$. This means $\boldsymbol{t} \notin T_R$. Therefore, $T_R \subseteq \mathrm{Ker}(\mathcal{P} \circ \mathcal{Q})$. This proves that $T_R = \mathrm{Ker}(\mathcal{P} \circ \mathcal{Q})$.

We then show that $L_R = \mathrm{Im}(\mathcal{Q}^\mathsf{T} \circ \mathcal{P}^\mathsf{T})$. For any $f \in \mathrm{Im}(\mathcal{Q}^\mathsf{T} \circ \mathcal{P}^\mathsf{T})$, there exists some linear function $h : \mathbb{R}^{|\mathcal{Z}|} \mapsto \mathbb{R}$ such that $f(\boldsymbol{r}) = \mathcal{Q}^\mathsf{T} \circ \mathcal{P}^\mathsf{T}(h)(\boldsymbol{r}) \equiv h(\mathcal{P} \circ \mathcal{Q}(\boldsymbol{r}))$, $\forall \boldsymbol{r} \in X_R$. Theorem 2.1 implies that each entry of $\mathcal{P}(\mathcal{Q}(\boldsymbol{r}))$ can be learned from some experiments, thus $f \in L_R$. This means $\mathrm{Im}(\mathcal{Q}^\mathsf{T} \circ \mathcal{P}^\mathsf{T}) \subseteq L_R$; On the other hand, $L_R \perp T_R$, according to similar arguments as in Lemma 2.8. We have shown $T_R = \mathrm{Ker}(\mathcal{P} \circ \mathcal{Q})$, thus $L_R \subseteq \mathrm{Im}(\mathcal{Q}^\mathsf{T} \circ \mathcal{P}^\mathsf{T})$ thanks to the orthogonality between kernel and adjoint image. This proves that $L_R = \mathrm{Im}(\mathcal{Q}^\mathsf{T} \circ \mathcal{P}^\mathsf{T})$. □

B.7.3 Additional proofs for Sec. B.4

Proof of Lemma 2.27

Proof. Since $\#\{\mathfrak{d}_s\}_s = \dim(T) = 2^n - 1$ (recall that the PTG is strongly connected thanks to the root vertex), we only need to show the completeness of $\{\mathfrak{d}_s\}_s$. It is sufficient to show that $\{\mathfrak{d}_s\}_s$ can express the canonical cut basis $\mathfrak{y}_z$ generated by cutting off a single vertex $\mathsf{V}_0 = \{v_z\}$ for all non-zero bit string z. (Note that the root node gives a redundant degree of freedom and can thus be ignored). By definition, we have

$$\mathfrak{d}_s = -\sum_z \mathbb{1}[z_s \neq 0]\mathfrak{y}_z. \tag{B.93}$$

For two n-bit string s, t we write $t \preceq s$ if $t_i = 0$ for any index i such that $s_i = 0$. Let $\bar{s}$ denote the bitwise flip of s (e.g., $\overline{1101} = 0010$). We claim the following relation holds,

$$\mathfrak{y}_z = \sum_{s\neq\mathbf{0}:\bar{z}\preceq s} (-1)^{|s|-|\bar{z}|}\mathfrak{d}_s, \quad \forall z \neq \mathbf{0}. \tag{B.94}$$

This is basically a consequence of the inclusion-exclusion principle. To see this, thanks to Eq. (B.93),

$$\begin{aligned}
R.H.S. &= -\sum_{s\neq\mathbf{0}:\bar{z}\preceq s} (-1)^{|s|-|\bar{z}|} \sum_{t:t_s\neq 0} \mathfrak{y}_t \\
&= -\sum_t \sum_{s:\bar{z}\preceq s,\ s\not\preceq\bar{t}} (-1)^{|s|-|\bar{z}|}\mathfrak{y}_t \\
&= \sum_t \sum_{s:\bar{z}\preceq s\preceq\bar{t}} (-1)^{|s|-|\bar{z}|}\mathfrak{y}_t \\
&= \sum_t \mathbb{1}[\bar{z}\preceq\bar{t}] \sum_{k=0}^{|\bar{t}|-|\bar{z}|} (-1)^k \binom{|\bar{t}|-|\bar{z}|}{k} \mathfrak{y}_t \\
&= \sum_t \mathbb{1}[\bar{z}\preceq\bar{t}]\mathbb{1}[|\bar{t}|-|\bar{z}|=0]\mathfrak{y}_t \\
&= \mathfrak{y}_z.
\end{aligned} \tag{B.95}$$

The order change of sum in the second line uses the fact that $t_s \neq 0$ is equivalent to $s \not\preceq \bar{t}$, and that $\mathbf{0} \preceq \bar{t}$ for any t, so we can drop the constraint $s \neq \mathbf{0}$. The third line uses the following

$$\sum_{s:\bar{z}\preceq s\preceq t}(-1)^{|s|-|\bar{z}|}+\sum_{s:\bar{z}\preceq s,\ s\not\preceq t}(-1)^{|s|-|\bar{z}|}=\sum_{s:\bar{z}\preceq s}(-1)^{|s|-|\bar{z}|}=\sum_{k=0}^{n-|\bar{z}|}\binom{n-|\bar{z}|}{k}(-1)^k=0. \tag{B.96}$$

Note that $z \neq \mathbf{0}$ implies $n - |\bar{z}| > 0$. The fifth line uses the Binomial theorem. This proves Eq. (B.94), and thus establishes the fact that the SDG's $\{\mathfrak{d}_s\}_s$ form a basis for T and U. □

Proof of converse part of Theorem 2.5(b)

Proof. Let $(X_R, \mathcal{Q})$ satisfies a fully local noise model for gate $\mathcal{G}$. Define

$$\begin{aligned} T_0 &\coloneqq \{\boldsymbol{z} \in T : \boldsymbol{z}^{\mathcal{G}} \in \mathrm{Im}\mathcal{Q}^{\mathcal{G}}\}, \\ D_0 &\coloneqq \mathrm{span}\{\mathfrak{d}_s : s \subset \mathrm{supp}(\mathcal{G}) \text{ or } s \supseteq \mathrm{supp}(\mathcal{G}) \text{ or } s \cap \mathrm{supp}(\mathcal{G}) = \varnothing\}. \end{aligned} \tag{B.97}$$

It has been proved that $T_0 \supseteq D_0$. Now, assume that $\hat{\mathcal{G}}$ induces a connected pattern transfer subgraph. Let $k = |\mathcal{G}|$ and assume $\mathcal{G}$ supports on the first k qubits, without loss of generality. By definition, any $\boldsymbol{z}^{\mathcal{G}} \in \mathrm{Im}\mathcal{Q}^{\mathcal{G}}$ satisfies

$$z^{\mathcal{G}}_{b\oplus c_1} = z^{\mathcal{G}}_{b\oplus c_2}, \quad \forall b \in \mathsf{P}^k\backslash\{I_k\},\ \forall c_1, c_2 \in \mathsf{P}^{n-k}. \tag{B.98}$$

Here $\oplus$ means concatenation of bit string, i.e., $P_{b\oplus c} = P_b \otimes P_c$. Equivalently, $\boldsymbol{z} \cdot (\boldsymbol{e}^{\mathcal{G}}_{b\oplus c_1} - \boldsymbol{e}^{\mathcal{G}}_{b\oplus c_2}) = 0$. Now, let us pick a set of k-qubit Paulis $\{b_j \in \mathsf{P}^k\backslash\{I\}\}_{j=1}^{2^k-2}$ such that the set of edges $\{\boldsymbol{e}^{\mathcal{G}}_{b_j\oplus\mathbf{0}}\}_{j=1}^{2^k-2}$, viewed as undirected, form a spanning tree for the subgraph of vertices $\{t \oplus 0_{n-k} \in \mathsf{V} : t \in \{0,1\}^k, t \neq 0_k\}$. Such a choice always exists, as we have assumed that the pattern transfer subgraph of $\widehat{\mathcal{G}}$ is strongly connected. Next, consider the following

vectors

$$\boldsymbol{k}_{j,c} \coloneqq \boldsymbol{e}^{\mathcal{G}}_{b_j \oplus I} - \boldsymbol{e}^{\mathcal{G}}_{b_j \oplus c}, \quad \forall j = 1, \cdots, 2^k - 2, \ \forall c \in \mathsf{P}^{n-k} \backslash \{I\}. \tag{B.99}$$

It is not hard to see $\{\boldsymbol{k}_{j,c}\}$ forms a linearly-independent set, as each $\boldsymbol{e}^{\mathcal{G}}_{b_j \oplus c}$ $(c \neq I)$ appears in exactly one $\boldsymbol{k}_{j,c}$. Let $K \coloneqq \mathrm{span}\{\boldsymbol{k}_{j,c}\}$. We thus have

$$\dim(K) = (2^k - 2) \times (2^{n-k} - 1) = 2^n - 2^k - 2^{n-k+1} + 2 \tag{B.100}$$

Since K supports on a union of spanning trees, the intersection between K and the cycle space Z must be trivial, i.e., $K \cap Z = \{\mathbf{0}\}$ (see e.g., [Ber01]). Meanwhile, one has $T_0 \perp K$ by Eq. (B.98) and $T_0 \perp Z$ by the orthogonality between cycle space and cut space. Combining all of these, we have $X \supseteq T_0 \oplus K \oplus Z$, and that

$$\begin{aligned} \dim(T_0) &\leq \dim(X) - \dim(Z) - \dim(K) \\ &= \dim(T) - \dim(K) \\ &= 2^k + 2^{n-k+1} - 3. \end{aligned} \tag{B.101}$$

The second line uses the fact that T is the complement of Z in X. On the other hand, by directly counting the number of SDG's in D_0, we have $\dim(D_0) = (2^k - 2) + 2^{n-k} + (2^{n-k} - 1) \geq \dim(T_0)$. Since we have shown $T_0 \supseteq D_0$, we conclude $D_0 = T_0$. This completes our proof. □

B.7.4 About quasi-local Pauli channels

Throughout this section, when we talk about a Pauli channel, we always assume all of its eigenvalues to be strictly positive, so that $x_a = -\log \lambda_a$ is well-defined.

Lemma 2.35. *Let $\Omega = 2^{[n]} \backslash \varnothing$. Then any n-qubit Pauli channel is Ω-local.*

Proof. For any set of parameters $\{x_a : a \in \mathsf{P}^n\}$, we claim there always exist $\{r_b : b \in \mathsf{P}^n\}$

such that,

$$x_a = \sum_{b \lhd a} r_b, \qquad \forall a \in \mathsf{P}^n. \tag{B.102}$$

$$r_b = \sum_{a \lhd b} (-1)^{w(a)-w(b)} x_a, \qquad \forall b \in \mathsf{P}^n. \tag{B.103}$$

Indeed, substituting the second expression into R.H.S. of the first yields

$$\begin{aligned}
\sum_{b \lhd a} \sum_{c \lhd b} (-1)^{w(c)-w(b)} x_c &= \sum_{c \lhd a} x_c \sum_{b : c \lhd b \lhd a} (-1)^{w(c)-w(b)} \\
&= \sum_{c \lhd a} x_c \sum_{k=0}^{|w(a)-w(c)|} \binom{w(a)-w(c)}{k} (-1)^k \\
&= \sum_{c \lhd a} x_c \mathbb{1}[w(a) = w(c)] \\
&= x_a.
\end{aligned} \tag{B.104}$$

The first line changes the order of summation. The third line uses the binomial theorem. Further note that $x_{I_n} = 0 \Leftrightarrow r_{I_n} = 0$. Since any Pauli channel is trace-preserving (i.e., $x_{I_n} = 0$), it is $2^{[n]} \backslash \varnothing$-local by choosing r_b according to Eq. (B.103). □

The following Lemma explains the equivalence between the quasi-local Pauli channels we consider and the sparse Pauli-Lindblad noise model. We note that basically the same results are given by [VDBMKT23, Theorem SIV.1].

Lemma 2.36. *Given an n-qubit Pauli channel Λ and a factor set Ω, if Λ can be written in the following concatenating form*

$$\Lambda(\cdot) = \bigcirc_{b \sim \Omega} \left(\eta_b P_b(\cdot) P_b + (1 - \eta_b)(\cdot) \right), \tag{B.105}$$

with real parameters $\{\eta_b\}_{b \sim \Omega}$ such that $\eta_b < 1/2$, $\forall b \neq I_n$, then Λ is Ω-local. Conversely, any Ω-local Pauli channel can be written in the above form.

Proof. Suppose Eq. (B.105) holds. By definition of the Pauli eigenvalues,

$$\lambda_a = \mathrm{Tr}[P_a \Lambda(P_a)]/2^n = \prod_{b \sim \Omega} (1 - 2\eta_b)^{\langle a,b \rangle}. \tag{B.106}$$

Taking a log on both sides yields (recall that we require $\eta_b < 1/2$)

$$x_a \coloneqq -\log(\lambda_a) = -\sum_{b \sim \Omega} \langle a, b \rangle \log(1 - 2\eta_b) =: \sum_{b \sim \Omega} \langle a, b \rangle \tau_b, \tag{B.107}$$

where we define $\tau_b \coloneqq -\log(1-2\eta_b)$. On the other hand, recall $\{r_b\}$ as defined in Eq. (B.103). Substituting Eq. (B.107) into the expression of r_b yields,

$$r_b = \sum_{c \sim \Omega} \sum_{a \lhd b} (-1)^{w(a)-w(b)} \langle a, c \rangle \tau_c. \tag{B.108}$$

Now we claim that $r_b = 0$ unless $b \sim \Omega$. If the condition does not hold, for any $c \sim \Omega$, there is at least one $i \in [n]$ such that $c_i = I$ but $b_i \neq I$. For any $a \lhd b$ such that $a_i = I$, let $\tilde{a}$ be obtained by replacing the ith entry of a by b_i. Then we obviously have

$$(-1)^{w(a)-w(b)} \langle a, c \rangle + (-1)^{w(\tilde{a})-w(b)} \langle \tilde{a}, c \rangle = 0.$$

Further note that $\{a \in \mathsf{P}^n : a \lhd b\}$ can be decomposed into disjoint pairs of the above form $(a,\, \tilde{a})$, we thus conclude that r_b must be 0, justifying our claim. We can thus rewrite the expression for $\boldsymbol{x}$ as

$$x_a = \sum_{b \lhd a, b \sim \Omega} r_b, \tag{B.109}$$

depending on a set of real parameters $\{r_b\}_{b \sim \Omega}$. Exponentiating both sides yields,

$$\lambda_a = \prod_{b \lhd a, b \sim \Omega} \exp(-r_b). \tag{B.110}$$

This completes the proof that Λ is Ω-local.

Conversely, let us assume that Λ is Ω-local, i.e., Eq. (B.109) holds. Define τ_b according to

$$\tau_b \coloneqq \mathbb{1}[b \neq I_n] \cdot \frac{-2}{4^n} \sum_a (-1)^{\langle a,b \rangle} x_a, \quad \forall b \in \mathsf{P}^n, \tag{B.111}$$

which gives us $x_a = \sum_b \langle a,b \rangle \tau_b$. Indeed,

$$\begin{aligned}
\sum_b \langle a,b \rangle \tau_b &= -\frac{2}{4^n} \sum_{b \neq I_n} \langle a,b \rangle \sum_c (-1)^{\langle c,b \rangle} x_c \\
&= -\frac{2}{4^n} \sum_{b \neq I_n} \frac{1-(-1)^{\langle a,b \rangle}}{2} \sum_c (-1)^{\langle c,b \rangle} x_c \\
&= \frac{1}{4^n} \sum_c x_c \sum_{b \neq I_n} \left(-(-1)^{\langle c,b \rangle} + (-1)^{\langle a+c,b \rangle} \right) \\
&= \sum_c x_c \left(-\mathbb{1}[c = I_n] + \mathbb{1}[c = a] \right) \\
&= x_a - x_{I_n},
\end{aligned} \tag{B.112}$$

which is just x_a since $x_{I_n} = 0$ by the trace preserving condition. Now substituting Eq. (B.109) into Eq. (B.111), considering only $b \neq I_n$,

$$\begin{aligned}
\tau_b &= -\frac{2}{4^n} \sum_a (-1)^{\langle a,b \rangle} \sum_{c \lhd a, c \sim \Omega} r_c \\
&= -\frac{2}{4^n} \sum_{c \sim \Omega} r_c \sum_{a : c \lhd a} (-1)^{\langle a,b \rangle}.
\end{aligned} \tag{B.113}$$

We claim that $\tau_b = 0$ unless $b \sim \Omega$. Similar to the previous case, if $b \sim \Omega$ does not hold, for any $c \sim \Omega$ there exists an index i such that $c_i = I$ but $b_i \neq I$. Consequently, exactly half of the Pauli operators from $\{a : c \lhd a\}$ commute with b while the other half anti-commute with b, depending on whether a_i commutes with b_i. Thus, the above sum always yields zero, proving our claim. By setting $\eta_b = (1 - \exp(-\tau_b))/2$ for all $b \sim \Omega$, we see Eq. (B.105) is a valid representation for Λ. This completes our proof. $\square$

We are now ready to give a proof for Lemma 2.33.

Proof for Lemma 2.33. Thanks to Lemma 2.36, any Ω-local Pauli channel Λ can be written in the form of Eq. (B.105). Conjugate Λ with $\mathcal{G}$ yields,

$$\begin{aligned}\mathcal{G}\Lambda\mathcal{G}^\dagger &= \bigcirc_{b\sim\Omega}\left(\eta_b\mathcal{G}\left(P_b(\mathcal{G}^\dagger(\cdot))P_b\right)+(1-\eta_b)(\cdot)\right)\\ &= \bigcirc_{b\sim\Omega}\left(\eta_b P_{\mathcal{G}(b)}(\cdot)P_{\mathcal{G}(b)}+(1-\eta_b)(\cdot)\right).\end{aligned} \tag{B.114}$$

If one can show that $\mathcal{G}$ is a permutation acting on $\{b : b\sim\Omega\}$, by simple relabeling one can see $\mathcal{G}\Lambda\mathcal{G}^\dagger$ also has the form of Eq. (B.105) and is thus Ω-local. For any $b\sim\Omega$, there must be a $\nu\in\Omega$ such that $\mathrm{supp}(b)=\nu$. it is not hard to see that

$$\mathrm{supp}(\mathcal{G}(b))\subseteq\Xi_{\mathcal{G}}(\nu). \tag{B.115}$$

By assumption, $\Xi_{\mathcal{G}}(\nu)\in\Omega$, thus $\mathrm{supp}(\mathcal{G}(b))\in\Omega$ by the property of Ω, which means $\mathcal{G}(b)\sim\Omega$. Since $\mathcal{G}$ is invertible, this completes the proof that $\mathcal{G}$ acts as a permutation over $\{b : b\sim\Omega\}$, and thus proving $\mathcal{G}\Lambda\mathcal{G}^\dagger$ is Ω-local. Since the same argument holds by replacing $\mathcal{G}$ with $\mathcal{G}^\dagger$, this completes the proof of Lemma 2.33. □

B.7.5 Additional proofs for Sec. B.6.3

Gauge in the nearest-neighbor CZ model

Proposition 2.37. *For the gate set and noise model described in Sec. B.6.3, the embedded gauge space is given by* $\mathcal{Q}(T_R)=\mathrm{span}\{\mathfrak{d}_{\{i\}} : i\in[n]\}$.

Proof. Since both state preparation and measurement noise are Ω-local, Theorem 2.6(a)

says,

$$\begin{aligned}\{\boldsymbol{z} \in T : \pi^S(\boldsymbol{z}) \in \mathrm{Im}\mathcal{Q}^S,\ \pi^M(\boldsymbol{z}) \in \mathrm{Im}\mathcal{Q}^M\} &= \mathrm{span}\{\mathfrak{d}_\nu : \nu \in \Omega\} \\ &= \mathrm{span}\{\mathfrak{d}_{\{i\}},\ \mathfrak{d}_{\{i,i+1\}} : i \in [n]\}.\end{aligned} \tag{B.116}$$

Then, by Lemma 2.23,

$$\begin{aligned}\mathcal{Q}(T_R) &= \mathrm{span}\{\mathfrak{d}_{\{i\}},\ \mathfrak{d}_{\{i,i+1\}} : i \in [n]\} \cap \{\boldsymbol{z} \in T : \pi^{\mathcal{G}_e}(\boldsymbol{z}) \in \mathrm{Im}\mathcal{Q}^{\mathcal{G}_e},\ \pi^{\mathcal{G}_o}(\boldsymbol{z}) \in \mathrm{Im}\mathcal{Q}^{\mathcal{G}_o}\} \\ &= \left\{\boldsymbol{z} \in \mathrm{span}\{\mathfrak{d}_{\{i\}},\ \mathfrak{d}_{\{i,i+1\}} : i \in [n]\} : \pi^{\mathcal{G}_e}(\boldsymbol{z}) \in \mathrm{Im}\mathcal{Q}^{\mathcal{G}_e},\ \pi^{\mathcal{G}_o}(\boldsymbol{z}) \in \mathrm{Im}\mathcal{Q}^{\mathcal{G}_o}\right\}.\end{aligned} \tag{B.117}$$

We are going to show that

$$\begin{aligned}&\left\{\boldsymbol{z} \in \mathrm{span}\{\mathfrak{d}_{\{i\}},\ \mathfrak{d}_{\{i,i+1\}} : i \in [n]\} : \pi^{\mathcal{G}_e}(\boldsymbol{z}) \in \mathrm{Im}\mathcal{Q}^{\mathcal{G}_e}\right\} \\ =&\,\mathrm{span}\left\{\{\mathfrak{d}_{\{i\}} : i \in [n]\} \cup \{\mathfrak{d}_{\{2k-1,2k\}} : k \in [n/2]\}\right\},\end{aligned} \tag{B.118}$$

$$\begin{aligned}&\left\{\boldsymbol{z} \in \mathrm{span}\{\mathfrak{d}_{\{i\}},\ \mathfrak{d}_{\{i,i+1\}} : i \in [n]\} : \pi^{\mathcal{G}_o}(\boldsymbol{z}) \in \mathrm{Im}\mathcal{Q}^{\mathcal{G}_o}\right\} \\ =&\,\mathrm{span}\left\{\{\mathfrak{d}_{\{i\}} : i \in [n]\} \cup \{\mathfrak{d}_{\{2k,2k+1\}} : k \in [n/2]\}\right\},\end{aligned} \tag{B.119}$$

which would then implies $\mathcal{Q}(T_R) = \mathrm{span}\{\mathfrak{d}_{\{i\}} : i \in [n]\}$. Let us look at Eq. (B.118) first. To see L.H.S. $\supseteq$ R.H.S., note that $\mathcal{G}_e^\dagger \mathcal{D}_{\{i\}} \mathcal{G}_e \mathcal{D}_{\{i\}}^{-1}$ and $\mathcal{G}_e^\dagger \mathcal{D}_{\{2k-1,2k\}} \mathcal{G}_e \mathcal{D}_{\{2k-1,2k\}}^{-1}$ are both clearly Ω-local for all $i \in [n], k \in [n/2]$. Here $\mathcal{D}_\nu$ is the subsystem depolarizing channel on ν defined in Eq. (B.34). Using the same argument as in the proof of Theorem 2.6(b), we can conclude those SDGs are indeed allowed. To see L.H.S. $\subseteq$ R.H.S, it remains to show that $\mathrm{span}\{\mathfrak{d}_{\{2k,2k+1\}} : k \in [n/2]\} \cap \text{L.H.S.} = \{\mathbf{0}\}$. Consider any vector of the form $\boldsymbol{z} = \sum_{k\in[n/2]} \alpha_k \mathfrak{d}_{\{2k,2k+1\}}$. By definition of the SDG,

$$z_a^{\mathcal{G}_e} := \boldsymbol{e}_a^{\mathcal{G}_e} \cdot \boldsymbol{z} = \sum_{k\in[n/2]} \alpha_k \left(\mathbb{1}[\mathrm{pt}(\mathcal{G}(a))_{\{2k,2k+1\}} \neq 00] - \mathbb{1}[\mathrm{pt}(a)_{\{2k,2k+1\}} \neq 00]\right). \tag{B.120}$$

Suppose $\pi^{\mathcal{G}_e}(\boldsymbol{z}) \in \mathrm{Im}\mathcal{Q}^{\mathcal{G}_e}$, then $z_a^{\mathcal{G}_e} = \sum_{b \lhd a, b\sim\Omega} t_b$ for some parameters t_b. Specifically, for

any $k \in [n/2]$, if we choose $a = X_{2k-1}X_{2k+3}$, since $\{2k-1, 2k+3\} \notin \Omega$ given that $n \geq 6$, we have

$$z^{\mathcal{G}_e}_{X_{2k-1}X_{2k+3}} = z^{\mathcal{G}_e}_{X_{2k-1}} + z^{\mathcal{G}_e}_{X_{2k+3}}. \tag{B.121}$$

On the other hand, Eq. (B.120) gives (see Fig. B.11 for how these are computed)

$$\begin{aligned} z^{\mathcal{G}_e}_{X_{2k-1}X_{2k+3}} &= \alpha_{k-1}(1-1) + \alpha_k(1-0) + \alpha_{k+1}(1-1) = \alpha_k. \\ z^{\mathcal{G}_e}_{X_{2k-1}} &= \alpha_{k-1}(1-1) + \alpha_k(1-0) = \alpha_k. \\ z^{\mathcal{G}_e}_{X_{2k+3}} &= \alpha_{k+1}(1-1) + \alpha_k(1-0) = \alpha_k. \end{aligned} \tag{B.122}$$

But this means α_k must be 0, which finishes the proof that $\text{span}\{\mathfrak{d}_{\{2k,2k+1\}} : k \in [n/2]\} \cap$

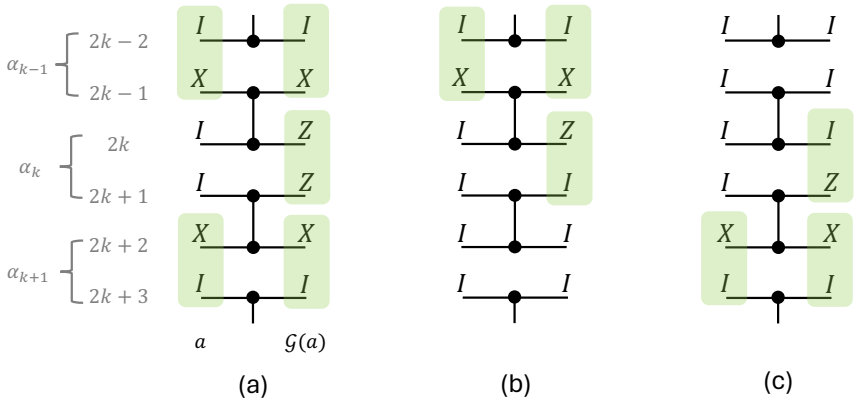


Figure B.11: Illustration for calculations of $z^{\mathcal{G}_e}_a$ where (a) $a = X_{2k-1}X_{2k+3}$, (b) $a = X_{2k-1}$, and (c) $a = X_{2k+3}$, respectively. The labels of qubits and the corresponding coefficient α_k is shown in the leftmost side. The shaded region represents a non-zero pattern in each $\{2k, 2k+1\}$ subsystem, for both a and $\mathcal{G}_e(a)$.

L.H.S. $= \{\mathbf{0}\}$. Eq. (B.119) can be proved similarly by cyclically shifting every index by 1. Combining everything, we complete the proof that $\mathcal{Q}(T_R) = \text{span}\{\mathfrak{d}_{\{i\}} : i \in [n]\}$. □

APPENDIX C

APPENDICES TO CHAPTER 5

C.1 Details of EMEEL protocol

The procedure of EMEEL is described in Algorithm 2. In the input parameters, $\boldsymbol{D} = [d_1, \cdots, d_{|\boldsymbol{D}|}]$ is a list of circuit depths. We require each d_i to be an integer multiple of d_0. See the next section for how to choose $\boldsymbol{D}$. N_C, N_S are number of random circuits per depth and number of measurement shots per circuit, respectively. The unprimed, primed version are parameters for EEL and AuxEFL, respectively; $\boldsymbol{Q} = [a_1, \cdots, a_{|\boldsymbol{Q}|}] \subseteq \mathsf{P}^n$ is set of queries of Pauli fidelities. The protocol will return an estimator $\hat{\lambda}_a$ to the symmetrized Pauli fidelity $\bar{\lambda}_a$ for each $a \in \boldsymbol{Q}$.

Algorithm 2 EMEEL

Input: $N_C, N_S, \boldsymbol{D}, N'_C, N'_S, \boldsymbol{D'}, \boldsymbol{Q}$.

1: $\boldsymbol{\mu}^{\mathsf{EEL}} = \textsc{CollectEEL}(N_C, N_S, \boldsymbol{D})$
2: $\boldsymbol{k}^{\mathsf{AuxEFL}} = \textsc{CollectAuxEFL}(N'_C, N'_S, \boldsymbol{D'})$
3: **for** $a \in \boldsymbol{Q}$ **do**
4: $\quad \hat{\lambda}_{\mathsf{EEL},a} \coloneqq \textsc{FitEEL}(\boldsymbol{\mu}^{\mathsf{EEL}}, a)$
5: $\quad \hat{\lambda}_{\mathsf{AuxEFL},\mathrm{pt}(a)} \coloneqq \textsc{FitAuxEFL}(\boldsymbol{k}^{\mathsf{AuxEFL}}, \mathrm{pt}(a))$
6: $\quad \hat{\lambda}_a \coloneqq \hat{\lambda}_{\mathsf{EEL},a} / \hat{\lambda}_{\mathsf{AuxEFL},\mathrm{pt}(a)}$
7: **return** $\{\hat{\lambda}_a\}_{a \in \boldsymbol{Q}}$.

The subroutines used for EMEEL are described in Algorithm 3-4. SampleGateSequence generates a randomized gate sequence containing d layers of $\mathcal{G}$. Here $\{P_{a_i}\}$ are the Pauli twirling gates while $\{C_i\}$ are the local-Clifford twirling gates. The P_α and P_β are Pauli gates for measurement twirling. The identity gate will be experimentally implemented as a dynamical decoupling sequence that suppresses idling noise and system-ancilla crosstalk [SLT+24]. CollectEEL calls SampleGateSequence as a subroutine to construct EEL experiments. Note that, $\mu_{d,i,j}$ defined in Line 14 is the "twirling-corrected"

Algorithm 3 Subroutines for EMEEL (Data Collection)

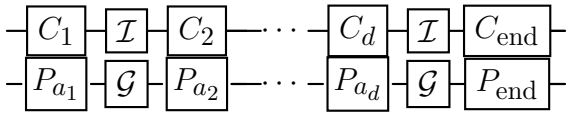
procedure SAMPLEGATESEQUENCE(d)

Uniformly sample d layers of single-qubit Clifford gates $C_1, \cdots, C_d$ and $d+2$ Pauli gates $P_{a_1}, \cdots, P_{a_d}, P_\alpha, P_\beta$.

$C_{\text{end}} \coloneqq P_\alpha C_1^\dagger C_2^\dagger \cdots C_d^\dagger$.

$P_{\text{end}} \coloneqq P_\beta \mathcal{G}^d(P_{a_1}) \mathcal{G}^{d-1}(P_{a_2}) \cdots \mathcal{G}(P_{a_d})$.

Define $\mathcal{T}$ to be the following gate sequence



return $\mathcal{T}, \alpha, \beta$

---

procedure COLLECTEEL($N_C, N_S, \boldsymbol{D}$)

for $d \in \boldsymbol{D}$ **do**

for $i = 1 \cdots N_C$ **do**

$\mathcal{T}, \alpha, \beta \coloneqq$ SAMPLEGATESEQUENCE(d)

for $j = 1 \cdots N_S$ **do**

Prepare $|\Phi\rangle$, apply $\mathcal{T}$, measure in $\{|\Phi_\nu\rangle\}$. Denote outcome as ν.

$\mu_{d,i,j} \coloneqq \nu + \alpha + \beta \in \mathsf{P}^n \cong \mathbb{Z}_2^{2n}$.

return $\boldsymbol{\mu} \coloneqq \{\mu_{d,i,j}\}$

---

procedure COLLECTAUXEFL($N_C, N_S, \boldsymbol{D}$)

for $d \in \boldsymbol{D}$ **do**

for $i = 1 \cdots N_C$ **do**

$\mathcal{T}, \alpha, \beta \coloneqq$ SAMPLEGATESEQUENCE(d)

for $j = 1 \cdots N_S$ **do**

Prepare $|0\rangle^{\otimes 2n}$, apply $\mathcal{T}$, measure ancilla in $\{|l\rangle\}$. Denote outcome as l.

$k_{d,i,j} \coloneqq l + \alpha_x \in \mathbb{Z}_2^n$.

return $\boldsymbol{k} \coloneqq \{k_{d,i,j}\}$

Algorithm 4 Subroutines for EMEEL (Data Processing)

procedure FitEEL($\boldsymbol{\mu}, a$)
    Obtain $N_C, N_S, \boldsymbol{D}$ from $\boldsymbol{\mu}$.
    for $d \in \boldsymbol{D}$ **do**
        $\hat{f}_d \coloneqq \frac{1}{N_C N_S} \sum_{i,j} (-1)^{\langle \mu, a \rangle}$
    Fit $\{\hat{f}_d\}$ to $f_d = \hat{\alpha}_a^{\mathsf{EEL}} (\hat{\lambda}_a^{\mathsf{EEL}})^d$.
    return $\hat{\lambda}_a^{\mathsf{EEL}}$

procedure FitAuxEFL($\boldsymbol{k}, s$)
    Obtain $N_C, N_S, \boldsymbol{D}$ from $\boldsymbol{k}$.
    for $d \in \boldsymbol{D}$ **do**
        $\hat{f}_d \coloneqq \frac{1}{N_C N_S} \sum_{i,j} (-1)^{k \cdot s}$
    Fit $\{\hat{f}_d\}$ to $f_d = \hat{\alpha}_s^{\mathsf{AuxEFL}} (\hat{\lambda}_s^{\mathsf{AuxEFL}})^d$.
    return $\hat{\lambda}_s^{\mathsf{AuxEFL}}$

measurement outcome taking into account P_α, P_β. Similar for CollectAuxEFL. The α_x in Line 23 is the "X part" of the Pauli operator as defined in Eq. (2.2). Finally, FitEEL and FitAuxEFL extract the fidelity per layer via curve fitting to cancel the SPAM noise, as is standard for RB-type methods [KLR+08, MGE11, FW20].

C.2 Proof for the correctness of EMEEL

Proposition 3.1. *For any a, $\hat{\lambda}_a$ given by EMEEL is a consistent (i.e., asymptotically unbiased) estimator for $\bar{\lambda}_a$.*

To prove this claim, note that EMEEL consists of two subroutines, EEL and AuxEFL, which generate an estimator for $\lambda_a^{\mathrm{anc}} \bar{\lambda}_a$ and λ_a^{anc}, respectively. Here we assume $\mathcal{G}$ is fixed and omit the corresponding superscript. The estimator of EMEEL is then given by the ratio of these two estimators. Therefore, we just need to show both EEL and AuxEFL give consistent estimators. This again reduces to showing the fidelity estimator at depth d is an unbiased estimator.

Notations. Let $\{|j\rangle\}_{j=0}^{2^n-1}$ denote the n-qubit computational basis. Define the canonical Bell state $|\Phi\rangle \coloneqq \sum_{j=0}^{2^n-1} |jj\rangle / \sqrt{2^n}$. The Bell basis $\{|\Phi\rangle_\nu\}_{\nu \in \mathsf{P}^n}$ is defined as $|\Phi\rangle_\nu \coloneqq I \otimes P_\nu |\Phi\rangle$, which forms a orthonormal basis for the $2n$-qubit Hilbert space. In the PTM representation, the n-qubit Bell states are given by

$$|\Phi_v\rangle\!\rangle = \mathcal{P}_v \otimes \mathbb{1} |\Phi_+\rangle\!\rangle = \frac{1}{4^n} \sum_{u \in \mathbb{Z}_2^{2n}} (-1)^{\langle u,v \rangle} |P_u \otimes P_u^T\rangle\!\rangle, \quad \forall v \in \mathbb{Z}_2^{2n}, \tag{C.1}$$

which forms a complete basis on $\mathcal{H}_{2^{2n}}$. Also, we use $|j\rangle\!\rangle$ to denote the vectorization of the computational basis state $|j\rangle\langle j|$ for $j = 1 \cdots 2^n - 1$.

Denote the state preparation and measurement noise channels as $\mathcal{E}^{\mathsf{S}}$ and $\mathcal{E}^{\mathsf{M}}$, respectively. Note that both of them acts on $2n$-qubits. Their Pauli eigenvalues is defined by ξ_a^{S} and ξ_a^{M} for any $2n$-qubit Pauli operator a. Use $\mathcal{C}_j(\cdot) \coloneqq C_j(\cdot)C_j^\dagger$, $\mathcal{P}_j(\cdot) \coloneqq P_j(\cdot)P_j^\dagger$ to denote the corresponding unitary channel of the Pauli/Clifford twirling gates.

Consistency of EEL. Let us compute the probability of getting corrected measurement outcome z, averaged over random circuits at depth d,

$$
\begin{aligned}
&\Pr[z] \\
&= \mathbb{E}\langle\!\langle \widetilde{\Phi}_{z+\alpha+\beta} | \mathcal{P}_{\alpha,\beta}^{AS} \left(\mathcal{C}_{\text{end}} \prod_{j=d}^{1} \Lambda_{\text{anc}} \mathcal{C}_j \right)^A \otimes \left(\mathcal{P}_{\text{end}} \prod_{j=d}^{1} \Lambda \mathcal{G} \mathcal{P}_j \right)^S |\widetilde{\Phi}_+\rangle\!\rangle \\
&= \langle\!\langle \Phi_z | \left(\mathbb{E}_{\alpha,\beta} \mathcal{P}_{\alpha,\beta} \mathcal{E}^{\mathsf{M}} \mathcal{P}_{\alpha,\beta} \right)^{AS} \left(\prod_{j=d}^{1} \mathbb{E}_{\mathcal{C}_j'} \mathcal{C}_j'^{\dagger} \Lambda_{\text{anc}} \mathcal{C}_j' \right)^A \otimes \left(\prod_{j=d}^{1} \mathbb{E}_{\mathcal{P}_j'} \mathcal{P}_j' \Lambda \mathcal{P}_j' \mathcal{G} \right)^S |\widetilde{\Phi}_+\rangle\!\rangle \\
&= \langle\!\langle \Phi_z | \sum_{a_1,a_2 \in \mathsf{P}^n} \xi_{a_1,a_2}^{\mathsf{M}} |\sigma_{a_1,a_2}\rangle\!\rangle \langle\!\langle \sigma_{a_1,a_2}|^{AS} \left(\prod_{j=d}^{1} \sum_{a_1 \in \mathsf{P}^n} \lambda_{\text{pt}(a_1)}^{\text{anc}} |\sigma_{a_1}\rangle\!\rangle \langle\!\langle \sigma_{a_1}| \right)^A \otimes \\
&\qquad\qquad \left(\prod_{j=d}^{1} \sum_{a_2 \in \mathsf{P}^n} \lambda_{a_2} |\sigma_{a_2}\rangle\!\rangle \langle\!\langle \sigma_{a_2}| \mathcal{G} \right)^S |\widetilde{\Phi}_+\rangle\!\rangle \\
&= \frac{1}{2^n} \sum_{a \in \mathsf{P}^n} (-1)^{\langle a,z \rangle} \xi_{a,a}^{\mathsf{M}} \langle\!\langle \sigma_{a,a}|^{AS} \left(\sum_{a_1 \in \mathsf{P}^n} (\lambda_{\text{pt}(a_1)}^{\text{anc}})^d |\sigma_{a_1}\rangle\!\rangle \langle\!\langle \sigma_{a_1}| \right)^A \otimes \\
&\qquad\qquad \left(\sum_{a_2 \in \mathsf{P}^n} (\lambda_{a_2} \lambda_{\mathcal{G}^\dagger(a_2)} \cdots \lambda_{\mathcal{G}^{\dagger d-1}(a_2)}) |\sigma_{a_2}\rangle\!\rangle \langle\!\langle \sigma_{a_2}| \right)^S |\widetilde{\Phi}_+\rangle\!\rangle \\
&= \frac{1}{2^n} \sum_{a \in \mathsf{P}^n} (-1)^{\langle a,z \rangle} \xi_{a,a}^{\mathsf{M}} (\lambda_{\text{pt}(a)}^{\text{anc}})^d \bar{\lambda}_a^d \langle\!\langle \sigma_{a,a} | \widetilde{\Phi}_+ \rangle\!\rangle \\
&= \frac{1}{4^n} \sum_{a \in \mathsf{P}^n} (-1)^{\langle a,z \rangle} \xi_{a,a}^{\mathsf{M}} (\lambda_{\text{pt}(a)}^{\text{anc}})^d \bar{\lambda}_a^d \xi_{a,a}^{\mathsf{S}} \\
&\equiv \frac{1}{4^n} \sum_{a \in \mathsf{P}^n} (-1)^{\langle a,z \rangle} \xi_{a,a} (\lambda_{\text{pt}(a)}^{\text{anc}})^d \bar{\lambda}_a^d.
\end{aligned}
\tag{C.2}
$$

Recall that $\sigma_a = P_a/\sqrt{2^n}$ is the normalized Pauli operator, and we write $\sigma_{a_1,a_2} \coloneqq \sigma_{a_1} \otimes \sigma_{a_2}$. In the second line, we use the following change of variables

$$
C_j' \coloneqq C_j C_{j-1} \cdots C_1, \quad P_j' \coloneqq \mathcal{G}(P_j) \mathcal{G}^2(P_{j-1}) \cdots \mathcal{G}^j(P_1), \quad \forall j = 1, \cdots, d, \tag{C.3}
$$

which does not change the average thanks to the unitary invariance of the Haar measure. We also use the following fact of the noisy Bell measurement POVM

$$\begin{aligned}\langle\langle\widetilde{\Phi}_{z+\alpha+\beta}| &= \frac{1}{2^n}\sum_{a\in\mathsf{P}^n}(-1)^{\langle a,z+\alpha+\beta\rangle}\langle\langle\sigma_{a,a}|\mathcal{E}^{\mathsf{M}} \\ &= \frac{1}{2^n}\sum_{a\in\mathsf{P}^n}(-1)^{\langle a,z\rangle}\langle\langle\sigma_{a,a}|\mathcal{P}_{\alpha,\beta}\mathcal{E}^{\mathsf{M}}.\end{aligned} \tag{C.4}$$

The third line uses the property of Pauli and local Clifford twirling (see e.g. [GCM+12b]). Note that Pauli twirling yields general Pauli channels, while local Clifford twirling yields Pauli channels of which the Pauli eigenvalues only depends on the pattern of the Pauli operator.

From Eq. (C.2) and the Walsh-Hadamard transform, we see that $\hat{f}_a(d)[z] \coloneqq (-1)^{\langle z,a\rangle}$ is an unbiased estimator for $\xi_{a,a}(\lambda^{\mathrm{anc}}_{\mathrm{pt}(a)})^d\bar{\lambda}_a^d$, for any $a\in\mathsf{P}^n$. Thus, by picking different d, averaging the estimators over different random circuits and measurement shots, then fitting to the single exponential decay model $\hat{\xi}^{\mathsf{EEL}}_{a,a}(\hat{\lambda}^{\mathsf{EEL}}_a)^d$, we obtain $\hat{\lambda}^{\mathsf{EEL}}_a$ as an consistent estimator for $\lambda^{\mathrm{anc}}_{\mathrm{pt}(a)}\bar{\lambda}_a$.

Consistency of AuxEFL. Let us compute the probability of obtaining corrected outcome z in AuxEFL, again averaged over random circuits at depth d.

$$\begin{aligned}\Pr[z] &= \mathbb{E}\langle\langle\widetilde{z+\alpha_x}|\mathcal{P}_\alpha\mathcal{C}_{\mathrm{end}}\prod_{j=d}^{1}\Lambda_{\mathrm{anc}}\mathcal{C}_j|\widetilde{0}\rangle\rangle \\ &= \langle\langle z|\left(\mathbb{E}_\alpha\mathcal{P}_\alpha\mathcal{E}'^{\mathsf{M}}\mathcal{P}_\alpha\right)\left(\prod_{j=d}^{1}\mathbb{E}_{\mathcal{C}'_j}\mathcal{C}'^{\dagger}_j\Lambda_{\mathrm{anc}}\mathcal{C}'_j\right)|\widetilde{0}\rangle\rangle \\ &= \langle\langle z|\left(\sum_{a\in\mathsf{P}^n}\xi'^{\mathsf{M}}_a|\sigma_a\rangle\rangle\langle\langle\sigma_a|\right)\left(\sum_{a\in\mathsf{P}^n}(\lambda^{\mathrm{anc}}_{\mathrm{pt}(a)})^d|\sigma_a\rangle\rangle\langle\langle\sigma_a|\right)|\widetilde{0}\rangle\rangle \\ &= \frac{1}{2^n}\sum_{\mu\in\{0,1\}^n}(-1)^{z\cdot\mu}\xi'^{\mathsf{M}}_\mu\xi'^{\mathsf{S}}_\mu(\lambda^{\mathrm{anc}}_\mu)^d \\ &\equiv \frac{1}{2^n}\sum_{\mu\in\{0,1\}^n}(-1)^{z\cdot\mu}\xi'_\mu(\lambda^{\mathrm{anc}}_\mu)^d\end{aligned} \tag{C.5}$$

In the second line we do the same change of variables as Eq. (C.3), and use the following fact of noisy computational basis measurement

$$\langle\langle\widetilde{z+\alpha_x}| = \langle\langle z+\alpha_x|\mathcal{E}'^{\mathsf{M}} = \langle\langle z|\mathcal{X}^{\alpha_x}\mathcal{Z}^{\alpha_z}\mathcal{E}'^{\mathsf{M}} = \langle z|\,\mathcal{P}_\alpha\mathcal{E}'^{\mathsf{M}}, \tag{C.6}$$

where we can add $\mathcal{Z}^{\alpha_z}$ as it acts trivially on the computational basis state. (Recall α_x, α_z is the X and Z part of Pauli P_α as defined in Methods Sec A.) We also note that the SPAM noise channel here can be quite different from the EFL experiment, as very different initial state and measurement settings are applied.

Thanks to Eq. (C.5) and the Walsh-Hadamard transform, we see that $\hat{f}_\mu(d)[z] \coloneqq (-1)^{z\cdot\mu}$ is an unbiased estimator for $\xi'_\mu(\lambda^{\mathrm{anc}}_\mu)^d$, for any $\mu \in \{0,1\}^n$. Thus, by picking different d, averaging the estimators over different random circuits and measurement shots, then fitting to the single exponential decay model $\hat{\xi}^{\mathsf{AuxEFL}}_\mu(\hat{\lambda}^{\mathsf{AuxEFL}}_\mu)^d$, we obtain $\hat{\lambda}^{\mathsf{AuxEFL}}_\mu$ as an consistent estimator for $\lambda^{\mathrm{anc}}_\mu$. Choosing $\mu \coloneqq \mathrm{pt}(a)$ gives us the relevant $\lambda^{\mathrm{anc}}_{\mathrm{pt}(a)}$.

C.3 Resource cost analysis

In this section, we analyze the variance overhead of EMEEL in estimating individual symmetrized Pauli eigenvalues, and compare it with the experimental setting reduction. Consider a specific $a \in \mathsf{P}^n$ with weight w. Henceforth, we omit the subscript of a or $\mathrm{pt}(a)$ for notational conciseness. Since $\hat{\lambda}_{\mathsf{EMEEL}} \coloneqq \hat{\lambda}_{\mathsf{EEL}}/\hat{\lambda}_{\mathsf{AuxEFL}}$, assuming both $\hat{\lambda}_{\mathsf{EEL}}$ and $\hat{\lambda}_{\mathsf{AuxEFL}}$ are sufficiently close to the true values, using the first-order Taylor expansion yields,

$$\frac{\mathrm{Var}[\hat{\lambda}_{\mathsf{EMEEL}}]}{\mathrm{Var}[\hat{\lambda}_{\mathsf{EFL}}]} \approx \frac{1}{\lambda^2_{\mathsf{AuxEFL}}}\frac{\mathrm{Var}[\hat{\lambda}_{\mathsf{EEL}}]}{\mathrm{Var}[\hat{\lambda}_{\mathsf{EFL}}]} + \frac{\lambda^2_{\mathsf{EFL}}}{\lambda^2_{\mathsf{AuxEFL}}}\frac{\mathrm{Var}[\hat{\lambda}_{\mathsf{AuxEFL}}]}{\mathrm{Var}[\hat{\lambda}_{\mathsf{EFL}}]}. \tag{C.7}$$

The second term uses the assumption that $\lambda_{\mathsf{EEL}} = \lambda_{\mathsf{EFL}}\lambda_{\mathsf{AuxEFL}}$, i.e., the system and auxiliary noise are independent. Note that $\hat{\lambda}_{\mathsf{EEL}}$, $\hat{\lambda}_{\mathsf{AuxEFL}}$, $\hat{\lambda}_{\mathsf{EFL}}$ are all obtained by fitting an exponential decay in the form of $f_{d,k} = \alpha\lambda^d$. Omitting the label of protocols for now. For simplicity, we assume

each protocol only uses two circuit depths, $\{0, d\}$, in which case the fidelity estimator is given by $\hat{\lambda} \coloneqq (\hat{f}_d/\hat{f}_0)^{1/d}$, called a ratio estimator. Indeed, let $\hat{f}_{0(d)} \coloneqq f_{0(d)} + \hat{\varepsilon}_{0(d)}$,

$$\hat{\lambda} = \left(\frac{\alpha\lambda^d + \hat{\varepsilon}_d}{\alpha + \hat{\varepsilon}_0}\right)^{\frac{1}{d}} \approx \lambda\left(1 + \frac{\hat{\varepsilon}_d - \lambda^d\hat{\varepsilon}_0}{d\alpha\lambda^d}\right), \tag{C.8}$$

up to first-order approximation in $\hat{\varepsilon}_{0(d)}$. Therefore, the variance of $\hat{\lambda}$ is given by

$$\mathrm{Var}[\hat{\lambda}] \approx \frac{\lambda^2}{d^2\alpha^2}\left(\mathrm{Var}[\hat{f}_0] + \mathrm{Var}[\hat{f}_d]/\lambda^{2d}\right) =: \frac{\lambda^2 V}{d^2\alpha^2}, \tag{C.9}$$

where V is a short-hand notation for the term within the bracket. Putting this back into Eq. (C.7) yields,

$$\frac{\mathrm{Var}[\hat{\lambda}_{\mathsf{EMEEL}}]}{\mathrm{Var}[\hat{\lambda}_{\mathsf{EFL}}]} \approx \frac{d_{\mathsf{EFL}}^2\alpha_{\mathsf{EFL}}^2 V_{\mathsf{EEL}}}{d_{\mathsf{EEL}}^2\alpha_{\mathsf{EEL}}^2 V_{\mathsf{EFL}}} + \frac{d_{\mathsf{EFL}}^2\alpha_{\mathsf{EFL}}^2 V_{\mathsf{AuxEFL}}}{d_{\mathsf{AuxEFL}}^2\alpha_{\mathsf{AuxEFL}}^2 V_{\mathsf{EFL}}}. \tag{C.10}$$

To proceed, we need to decide d for each protocols. One sensible choice is to let $\lambda^d \approx 1/e$, which enables a multiplicative precision estimation for λ with respect to the infidelity. There exists a procedure to efficiently search for such d [HHF+19]. Here we assume the search is done *a priori*. On the other hand, if we assume that (1) f_0 is sufficiently close to 1, so the contribution of $\mathrm{Var}[\hat{f}_0]$ to V can be ignored, and that (2) $\hat{f}_d$ has binomial variance $\mathrm{Var}[\hat{f}_d] = (1 - f_d)^2/N$ (which is true if $N_S = 1$ or the noise is sufficiently uniform among different random circuit instances), then the value of V is comparable for all protocols. Finally, if we further make the assumption that $\lambda_{\mathsf{AuxEFL}}$ is at least as large as λ_{EFL} (which means the system gate noise is at least as large as the ancilla idling noise), we have

$$\frac{\mathrm{Var}[\hat{\lambda}_{\mathsf{EMEEL}}]}{\mathrm{Var}[\hat{\lambda}_{\mathsf{EFL}}]} \approx \left(1 + \frac{r_{\mathsf{AuxEFL}}}{r_{\mathsf{EFL}}}\right)^2 \frac{\alpha_{\mathsf{EFL}}^2}{\alpha_{\mathsf{EEL}}^2} + 1. \tag{C.11}$$

Here $r \coloneqq \ln\lambda^{-1} \approx 1 - \lambda$ when λ is close to 1. In conclusion, under all the assumptions we made in above, the variance overhead of EMEEL roughly scales as the SPAM fidelity ratio $(\alpha_{\mathsf{EFL}}/\alpha_{\mathsf{EEL}})^2$. Suppose each Bell pair suffers from identical and independent error, in which case $\alpha_{\mathsf{EEL}} = c^{w/2}$ for some constant $c < 1$. As long as $c > 1/2$ (which is a mild assumption about the Bell pair fidelity), the variance overhead will grow slower than 2^w, and thus EMEEL will have an efficiency enhancement over EFL.

We remark that the above analysis only aims at understanding the rough scaling of the variance overhead rather than giving an exact prediction. In our experiments, not all assumptions made above are satisfied: For example, we choose $d = 0, 16, 32$ and extract the rate using a fit for all protocols instead of searching for individual optimal depths and using the ratio estimator. We also choose $N_S = 500$, therefore $\mathrm{Var}[\hat{f}_d]$ does not satisfy binomial variance. Nevertheless, by directly estimating the sample variance, we obtain a variance overhead scales as 1.33^w in Fig. 5.3, which is considerably smaller than 2^w or 3^w, demonstrating the entanglement enhancement even in presence of noise. We also verify that, although Eq. (C.11) deviates from the empirical variance overhead, Eq. (C.10) gives a precise prediction by inserting the empirical variance for each $\mathrm{Var}[\hat{f}_0]$ and $\mathrm{Var}[\hat{f}_d]$ at $d = 32$. See SM Sec. IV.

REFERENCES

[AA11] Scott Aaronson and Alex Arkhipov. The computational complexity of linear optics. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 333–342, 2011.

[AAA+23] Igor Aleiner, Richard Allen, Trond I Andersen, Markus Ansmann, Frank Arute, Kunal Arya, Abraham Asfaw, Juan Atalaya, Ryan Babbush, Dave Bacon, et al. Suppressing quantum errors by scaling a surface code logical qubit. *Nature*, 614(7949):676–681, 2023.

[AAA+25] Dorit Aharonov, Ori Alberton, Itai Arad, Yosi Atia, Eyal Bairey, Zvika Brakerski, Itsik Cohen, Omri Golan, Ilya Gurwich, Oded Kenneth, et al. On the importance of error mitigation for quantum computation. *arXiv preprint arXiv:2503.17243*, 2025.

[AAAB+25] Rajeev Acharya, Dmitry A Abanin, Laleh Aghababaie-Beni, Igor Aleiner, Trond I Andersen, Markus Ansmann, Frank Arute, Kunal Arya, Abraham Asfaw, Nikita Astrakhantsev, et al. Quantum error correction below the surface code threshold. *Nature*, 638(8052):920–926, 2025.

[AAB+19] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando GSL Brandao, David A Buell, et al. Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510, 2019.

[ABO08] Dorit Aharonov and Michael Ben-Or. Fault-tolerant quantum computation with constant error rate. *SIAM Journal on Computing*, 2008.

[ACQ22] Dorit Aharonov, Jordan Cotler, and Xiao-Liang Qi. Quantum algorithmic measurement. *Nature communications*, 13(1):887, 2022.

[AP08] Panos Aliferis and John Preskill. Fault-tolerant quantum computation against biased noise. *Physical Review A*, 78(5):052331, 2008.

[BATB+21] J Pablo Bonilla Ataides, David K Tuckett, Stephen D Bartlett, Steven T Flammia, and Benjamin J Brown. The xzzx surface code. *Nature communications*, 12(1):1–12, 2021.

[BBC+93] Charles H Bennett, Gilles Brassard, Claude Crépeau, Richard Jozsa, Asher Peres, and William K Wootters. Teleporting an unknown quantum state via dual classical and einstein-podolsky-rosen channels. *Physical review letters*, 70(13):1895, 1993.

[BBP15] Leonardo Banchi, Samuel L Braunstein, and Stefano Pirandola. Quantum fidelity for arbitrary gaussian states. *Physical review letters*, 115(26):260501, 2015.

[BBPS96] Charles H Bennett, Herbert J Bernstein, Sandu Popescu, and Benjamin Schumacher. Concentrating partial entanglement by local operations. *Physical Review A*, 53(4):2046, 1996.

[BC18] Ge Bai and Giulio Chiribella. Test one to test many: a unified approach to quantum benchmarks. *Physical Review Letters*, 120(15):150502, 2018.

[BCG+24] Sergey Bravyi, Andrew W Cross, Jay M Gambetta, Dmitri Maslov, Patrick Rall, and Theodore J Yoder. High-threshold and low-overhead fault-tolerant quantum memory. *Nature*, 627(8005):778–782, 2024.

[BCL20] Sebastien Bubeck, Sitan Chen, and Jerry Li. Entanglement is necessary for optimal quantum property testing. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 692–703. IEEE, 2020.

[BDLR24] Simon Becker, Nilanjana Datta, Ludovico Lami, and Cambyse Rouze. Classical shadow tomography for continuous variables quantum systems. *IEEE Transactions on Information Theory*, 70(5):3427–3452, 2024.

[BDSW96] Charles H Bennett, David P DiVincenzo, John A Smolin, and William K Wootters. Mixed-state entanglement and quantum error correction. *Physical Review A*, 54(5):3824, 1996.

[BEG+24] Dolev Bluvstein, Simon J Evered, Alexandra A Geim, Sophie H Li, Hengyun Zhou, Tom Manovitz, Sepehr Ebadi, Madelyn Cain, Marcin Kalinowski, Dominik Hangleiter, et al. Logical quantum processor based on reconfigurable atom arrays. *Nature*, 626(7997):58–65, 2024.

[Ber01] Claude Berge. *The theory of graphs*. Courier Corporation, 2001.

[BGH+22] Anthony J Brady, Christina Gao, Roni Harnik, Zhen Liu, Zheshen Zhang, and Quntao Zhuang. Entangled sensor-networks for dark-matter searches. *PRX Quantum*, 3(3):030333, 2022.

[BGY+11] Jonas Bylander, Simon Gustavsson, Fei Yan, Fumiki Yoshihara, Khalil Harrabi, George Fitch, David G Cory, Yasunobu Nakamura, Jaw-Shen Tsai, and William D Oliver. Noise spectroscopy through dynamical decoupling with a superconducting flux qubit. *Nature Physics*, 7(7):565–570, 2011.

[BIS+18] Sergio Boixo, Sergei V Isakov, Vadim N Smelyanskiy, Ryan Babbush, Nan Ding, Zhang Jiang, Michael J Bremner, John M Martinis, and Hartmut Neven. Characterizing quantum supremacy in near-term devices. *Nature Physics*, 14(6):595–600, 2018.

[BKdSN+22] Robin Blume-Kohout, Marcus P da Silva, Erik Nielsen, Timothy Proctor, Kenneth Rudinger, Mohan Sarovar, and Kevin Young. A taxonomy of small markovian errors. *PRX Quantum*, 3(2):020335, 2022.

[BKN+13] Robin Blume-Kohout, John King Gamble, Erik Nielsen, Jonathan Mizrahi, Jonathan D. Sterk, and Peter Maunz. Robust, self-consistent, closed-form tomography of quantum logic gates on a trapped ion qubit, 2013.

[Bol98] Béla Bollobás. *Modern graph theory*, volume 184. Springer Science & Business Media, 1998.

[BPK+21] Kelly M Backes, Daniel A Palken, S Al Kenany, Benjamin M Brubaker, SB Cahn, A Droster, Gene C Hilton, Sumita Ghosh, H Jackson, Steve K Lamoreaux, et al. A quantum enhanced search for dark matter axions. *Nature*, 590(7845):238–242, 2021.

[BR02] Stephen Barnett and Paul M Radmore. *Methods in theoretical quantum optics*, volume 15. Oxford University Press, 2002.

[Bro13] Peter Brooks. *Quantum error correction with biased noise*. California Institute of Technology, 2013.

[BSBN02] Stephen D Bartlett, Barry C Sanders, Samuel L Braunstein, and Kae Nemoto. Efficient classical simulation of continuous variable quantum information processes. *Physical Review Letters*, 88(9):097904, 2002.

[BSK+21] Sergey Bravyi, Sarah Sheldon, Abhinav Kandala, David C. Mckay, and Jay M. Gambetta. Mitigating measurement errors in multiqubit experiments. *Phys. Rev. A*, 103:042605, Apr 2021.

[BSST02] Charles H Bennett, Peter W Shor, John A Smolin, and Ashish V Thapliyal. Entanglement-assisted capacity of a quantum channel and the reverse shannon theorem. *IEEE Transactions on Information Theory*, 48(10):2637–2655, 2002.

[BVL05] Samuel L Braunstein and Peter Van Loock. Quantum information with continuous variables. *Reviews of modern physics*, 77(2):513, 2005.

[BW92] Charles H Bennett and Stephen J Wiesner. Communication via one-and two-particle operators on einstein-podolsky-rosen states. *Physical review letters*, 69(20):2881, 1992.

[BW23] Stefanie J Beale and Joel J Wallman. Randomized compiling for subsystem measurements. *arXiv preprint arXiv:2304.06599*, 2023.

[Can20] Clément L Canonne. A short note on learning discrete distributions. *arXiv preprint arXiv:2002.11457*, 2020.

[Car22] Matthias C Caro. Learning quantum processes and hamiltonians via the pauli transfer matrix. *arXiv preprint arXiv:2212.04471*, 2022.

[Cas96] Kenneth R Castleman. *Digital image processing*. Prentice Hall Press, 1996.

[CBB+23] Zhenyu Cai, Ryan Babbush, Simon C Benjamin, Suguru Endo, William J Huggins, Ying Li, Jarrod R McClean, and Thomas E O'Brien. Quantum error mitigation. *Reviews of Modern Physics*, 95(4):045005, 2023.

[CCF+25] Edward H. Chen, Senrui Chen, Laurin Fischer, Andrew Eddins, Luke Govia, Brad Mitchell, Andre He, Youngseok Kim, Liang Jiang, and Alireza Seif. Disambiguating Pauli noise in quantum computers. *under preparation*, 2025.

[CCHL22] Sitan Chen, Jordan Cotler, Hsin-Yuan Huang, and Jerry Li. Exponential separations between learning with and without quantum memory. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 574–585. IEEE, 2022.

[CCHL23] Sitan Chen, Jordan Cotler, Hsin-Yuan Huang, and Jerry Li. The complexity of nisq. *Nature Communications*, 14(1):6001, 2023.

[CDDH+23] Arnaud Carignan-Dugas, Dar Dahlen, Ian Hincks, Egor Ospadov, Stefanie J Beale, Samuele Ferracin, Joshua Skanes-Norman, Joseph Emerson, and Joel J Wallman. The error reconstruction and compiled calibration of quantum computing cycles. *arXiv preprint arXiv:2303.17714*, 2023.

[CDKL01] Juan Ignacio Cirac, Wolfgang Dür, Barabara Kraus, and Maciej Lewenstein. Entangling operations and their implementation using a small amount of entanglement. *Physical Review Letters*, 86(3):544, 2001.

[CG69] Kevin E Cahill and Roy J Glauber. Ordered expansions in boson amplitude operators. *Physical Review*, 177(5):1857, 1969.

[CG19] Eric Chitambar and Gilad Gour. Quantum resource theories. *Reviews of Modern Physics*, 91(2):025001, 2019.

[CG23a] Sitan Chen and Weiyuan Gong. Efficient pauli channel estimation with logarithmic quantum memory. *arXiv preprint arXiv:2309.14326*, 2023.

[CG23b] Sitan Chen and Weiyuan Gong. Futility and utility of a few ancillas for pauli channel learning. *arXiv preprint arXiv:2309.14326*, 2023.

[CGFF17] Joshua Combes, Christopher Granade, Christopher Ferrie, and Steven T Flammia. Logical randomized benchmarking. *arXiv preprint arXiv:1702.03688*, 2017.

[CHL+22] Sitan Chen, Brice Huang, Jerry Li, Allen Liu, and Mark Sellke. Tight bounds for state tomography with incoherent measurements. *arXiv preprint arXiv:2206.05265*, 2022.

[Cho75] Man-Duen Choi. Completely positive linear maps on complex matrices. *Linear algebra and its applications*, 10(3):285–290, 1975.

[Cli90] Peter Clifford. Markov random fields in statistics. *Disorder in physical systems: A volume in honour of John M. Hammersley*, pages 19–32, 1990.

[CLM+14] Eric Chitambar, Debbie Leung, Laura Mančinska, Maris Ozols, and Andreas Winter. Everything you always wanted to know about locc (but were afraid to ask). *Communications in Mathematical Physics*, 328:303–326, 2014.

[CLO21] Sitan Chen, Jerry Li, and Ryan O'Donnell. Toward instance-optimal state certification with incoherent measurements. *arXiv preprint arXiv:2102.13098*, 2021.

[CLO+23] Senrui Chen, Yunchao Liu, Matthew Otten, Alireza Seif, Bill Fefferman, and Liang Jiang. The learnability of pauli noise. *Nature Communications*, 14(1):52, 2023.

[COG+24] Emilio Pelaez Cisneros, Victory Omole, Pranav Gokhale, Rich Rines, Kaitlin N Smith, Michael A Perlin, and Akel Hashim. Average circuit eigenvalue sampling on nisq devices. *arXiv preprint arXiv:2403.12857*, 2024.

[Col13] David Collins. Mixed-state pauli-channel parameter estimation. *Physical Review A*, 87(3):032301, 2013.

[Cov99] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[COZ+24a] Senrui Chen, Changhun Oh, Sisi Zhou, Hsin-Yuan Huang, and Liang Jiang. Tight bounds on pauli channel learning without entanglement. *Physical Review Letters*, 132(18):180805, 2024.

[COZ+24b] Senrui Chen, Changhun Oh, Sisi Zhou, Hsin-Yuan Huang, and Liang Jiang. Tight bounds on pauli channel learning without entanglement. *Physical Review Letters*, 132(18):180805, 2024.

[CRV+11] A Chiuri, V Rosati, G Vallone, S Pádua, H Imai, S Giacomini, C Macchiavello, and P Mataloni. Experimental realization of optimal noise estimation for a general pauli channel. *Physical review letters*, 107(25):253602, 2011.

[CT12] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[CW04] Carlton M Caves and Krzysztof Wódkiewicz. Fidelity of gaussian channels. *Open Systems & Information Dynamics*, 11(4):309–323, 2004.

[CXB20] Zhenyu Cai, Xiaosi Xu, and Simon C Benjamin. Mitigating coherent noise using pauli conjugation. *npj Quantum Information*, 6(1):1–9, 2020.

[CYK+22] Edward H Chen, Theodore J Yoder, Youngseok Kim, Neereja Sundaresan, Srikanth Srinivasan, Muyuan Li, Antonio D Córcoles, Andrew W Cross, and Maika Takita. Calibrated decoders for experimental quantum error correction. *Physical Review Letters*, 128(11):110504, 2022.

[CYZF21] Senrui Chen, Wenjun Yu, Pei Zeng, and Steven T. Flammia. Robust shadow estimation. *PRX Quantum*, 2:030348, Sep 2021.

[CZJF24] Senrui Chen, Zhihan Zhang, Liang Jiang, and Steven T Flammia. Efficient self-consistent learning of gate set pauli noise. *arXiv preprint arXiv:2410.03906*, 2024.

[CZSJ22] Senrui Chen, Sisi Zhou, Alireza Seif, and Liang Jiang. Quantum advantages for Pauli channel estimation. *Physical Review A*, 105(3):032435, 2022.

[dAKBS+20] Rayssa B de Andrade, Hugo Kerdoncuff, Kirstine Berg-Sørensen, Tobias Gehring, Mikael Lassen, and Ulrik L Andersen. Quantum-enhanced continuous-wave stimulated raman scattering spectroscopy. *Optica*, 7(5):470–475, 2020.

[DCEL09] Christoph Dankert, Richard Cleve, Joseph Emerson, and Etera Livine. Exact and approximate unitary 2-designs and their application to fidelity estimation. *Phys. Rev. A*, 80:012304, Jul 2009.

[DGL+23] Yu-Hao Deng, Yi-Chao Gu, Hua-Liang Liu, Si-Qiu Gong, Hao Su, Zhi-Jiong Zhang, Hao-Yang Tang, Meng-Hao Jia, Jia-Min Xu, Ming-Cheng Chen, Jian Qin, Li-Chao Peng, Jiarong Yan, Yi Hu, Jia Huang, Hao Li, Yuxuan Li, Yaojian Chen, Xiao Jiang, Lin Gan, Guangwen Yang, Lixing You, Li Li, Han-Sen Zhong, Hui Wang, Nai-Le Liu, Jelmer J. Renema, Chao-Yang Lu, and Jian-Wei Pan. Gaussian boson sampling with pseudo-photon-number-resolving detectors and quantum computational advantage. *Phys. Rev. Lett.*, 131:150601, Oct 2023.

[DRC17] Christian L Degen, F Reinhard, and Paola Cappellaro. Quantum sensing. *Reviews of modern physics*, 89(3):035002, 2017.

[DTW17] Kasper Duivenvoorden, Barbara M Terhal, and Daniel Weigand. Single-mode displacement sensor. *Physical Review A*, 95(1):012305, 2017.

[ea21] MD SAJID ANIS et al. Qiskit: An open-source framework for quantum computing, 2021.

[EAŻ05] Joseph Emerson, Robert Alicki, and Karol Życzkowski. Scalable noise estimation with random unitary operators. *Journal of Optics B: Quantum and Semiclassical Optics*, 7(10):S347–S352, sep 2005.

[EBL18] Suguru Endo, Simon C Benjamin, and Ying Li. Practical quantum error mitigation for near-future applications. *Physical Review X*, 8(3):031027, 2018.

[EHW+20] Jens Eisert, Dominik Hangleiter, Nathan Walk, Ingo Roth, Damian Markham, Rhea Parekh, Ulysse Chabaud, and Elham Kashefi. Quantum certification and benchmarking. *Nature Reviews Physics*, 2(7):382–390, 2020.

[EWP+19] Alexander Erhard, Joel J. Wallman, Lukas Postler, Michael Meth, Roman Stricker, Esteban A. Martinez, Philipp Schindler, Thomas Monz, Joseph Emerson, and Rainer Blatt. Characterizing large-scale quantum computers via cycle benchmarking. *Nature Communications*, 10(1):5347, Nov 2019.

[FGLE12] Steven T Flammia, David Gross, Yi-Kai Liu, and Jens Eisert. Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators. *New Journal of Physics*, 14(9):095022, 2012.

[FHV+24] Samuele Ferracin, Akel Hashim, Jean-Loup Ville, Ravi Naik, Arnaud Carignan-Dugas, Hammam Qassim, Alexis Morvan, David I Santiago, Irfan Siddiqi, and Joel J Wallman. Efficiently improving the performance of noisy quantum computers. *Quantum*, 8:1410, 2024.

[FI03] Akio Fujiwara and Hiroshi Imai. Quantum parameter estimation of a generalized pauli channel. *Journal of Physics A: Mathematical and General*, 36(29):8093, 2003.

[Fla22] Steven T. Flammia. Averaged Circuit Eigenvalue Sampling. In François Le Gall and Tomoyuki Morimae, editors, *17th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2022)*, volume 232 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 4:1–4:10, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[FM14] Austin G Fowler and John M Martinis. Quantifying the effects of local many-qubit errors and nonlocal two-qubit errors on the surface code. *Physical Review A*, 89(3):032316, 2014.

[FMMD21] Samuele Ferracin, Seth T. Merkel, David McKay, and Animesh Datta. Experimental accreditation of outputs of noisy quantum computers. *Phys. Rev. A*, 104:042603, Oct 2021.

[FO21] Steven T Flammia and Ryan O'Donnell. Pauli error estimation via population recovery. *Quantum*, 5:549, 2021.

[FOF23] Omar Fawzi, Aadil Oufkir, and Daniel Stilck França. Lower bounds on learning pauli channels. *arXiv preprint arXiv:2301.09192*, 2023.

[FOP05] Alessandro Ferraro, Stefano Olivares, and Matteo GA Paris. Gaussian states in continuous variable quantum information. *arXiv preprint quant-ph/0503237*, 2005.

[FW20] Steven T. Flammia and Joel J. Wallman. Efficient estimation of Pauli channels. *ACM Transactions on Quantum Computing*, 1(1), December 2020.

[GAG+24] Srilekha Gandhari, Victor V Albert, Thomas Gerrits, Jacob M Taylor, and Michael J Gullans. Precision bounds on continuous-variable state tomography using classical shadows. *PRX Quantum*, 5(1):010346, 2024.

[GBB+20] Xueshi Guo, Casper R Breum, Johannes Borregaard, Shuro Izumi, Mikkel V Larsen, Tobias Gehring, Matthias Christandl, Jonas S Neergaard-Nielsen, and Ulrik L Andersen. Distributed quantum sensing in a continuous-variable entangled network. *Nature Physics*, 16(3):281–284, 2020.

[GCM+12a] Jay M. Gambetta, A. D. Córcoles, S. T. Merkel, B. R. Johnson, John A. Smolin, Jerry M. Chow, Colm A. Ryan, Chad Rigetti, S. Poletto, Thomas A. Ohki, Mark B. Ketchen, and M. Steffen. Characterization of addressability by simultaneous randomized benchmarking. *Phys. Rev. Lett.*, 109:240504, Dec 2012.

[GCM+12b] Jay M. Gambetta, A. D. Córcoles, S. T. Merkel, B. R. Johnson, John A. Smolin, Jerry M. Chow, Colm A. Ryan, Chad Rigetti, S. Poletto, Thomas A. Ohki, Mark B. Ketchen, and M. Steffen. Characterization of addressability by simultaneous randomized benchmarking. *Phys. Rev. Lett.*, 109:240504, Dec 2012.

[GGH+24] James W Gardner, Tuvia Gefen, Simon A Haine, Joseph J Hope, John Preskill, Yanbei Chen, and Lee McCuller. Stochastic waveform estimation at the fundamental quantum limit. *arXiv preprint arXiv:2404.13867*, 2024.

[GGPCH14] Vittorio Giovannetti, Raul Garcia-Patron, Nicholas J Cerf, and Alexander S Holevo. Ultimate classical communication rates of quantum optical channels. *Nature Photonics*, 8(10):796–800, 2014.

[Gho21] Malay Ghosh. Exponential tail bounds for chisquared random variables. *Journal of Statistical Theory and Practice*, 15(2):1–6, 2021.

[GJN+23] D Ganapathy, W Jia, M Nakano, V Xu, N Aritomi, T Cullen, N Kijbunchoo, SE Dwyer, A Mullavey, L McCuller, et al. Broadband quantum enhancement of the ligo detectors with frequency-dependent squeezing. *Physical Review X*, 13(4):041021, 2023.

[GKP01] Daniel Gottesman, Alexei Kitaev, and John Preskill. Encoding a qubit in an oscillator. *Physical Review A*, 64(1):012310, 2001.

[GLM06] Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. Quantum metrology. *Physical review letters*, 96(1):010401, 2006.

[GLM11] Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. Advances in quantum metrology. *Nature photonics*, 5(4):222–229, 2011.

[GLS03] Petra M. Gleiss, Josef Leydold, and Peter F. Stadler. Circuit bases of strongly connected digraphs. *Discuss. Math. Graph Theory*, 23(2):241–260, 2003.

[Got98] Daniel Gottesman. Theory of fault-tolerant quantum computation. *Physical Review A*, 57(1):127, 1998.

[Gre15] Daniel Greenbaum. Introduction to quantum gate set tomography. *arXiv preprint arXiv:1509.02921*, 2015.

[GS22] Daniel Grier and Luke Schaeffer. The classification of clifford gates over qubits. *Quantum*, 6:734, 2022.

[GT07] Nicolas Gisin and Rob Thew. Quantum communication. *Nature photonics*, 1(3):165–171, 2007.

[GVK12] Emanuel Gavartin, Pierre Verlot, and Tobias J Kippenberg. A hybrid on-chip optomechanical transducer for ultrasensitive force measurements. *Nature nanotechnology*, 7(8):509–514, 2012.

[Hay10] Masahito Hayashi. Quantum channel estimation and asymptotic bound. In *Journal of Physics: Conference Series*, volume 233, page 012016. IOP Publishing, 2010.

[HBC+22] Hsin-Yuan Huang, Michael Broughton, Jordan Cotler, Sitan Chen, Jerry Li, Masoud Mohseni, Hartmut Neven, Ryan Babbush, Richard Kueng, John Preskill, et al. Quantum advantage in learning from experiments. *Science*, 376(6598):1182–1186, 2022.

[HDH25] Evan T Hockings, Andrew C Doherty, and Robin Harper. Scalable noise characterization of syndrome-extraction circuits with averaged circuit eigenvalue sampling. *PRX Quantum*, 6(1):010334, 2025.

[HFP22] Hsin-Yuan Huang, Steven T. Flammia, and John Preskill. Foundations for learning from noisy quantum experiments, 2022.

[HFW20] Robin Harper, Steven T Flammia, and Joel J Wallman. Efficient learning of quantum noise. *Nature Physics*, 16(12):1184–1188, 2020.

[HGM+24] Hong-Ye Hu, Andi Gu, Swarnadeep Majumder, Hang Ren, Yipei Zhang, Derek S Wang, Yi-Zhuang You, Zlatko Minev, Susanne F Yelin, and Alireza Seif. Demonstration of robust and efficient quantum property learning with shallow shadows. *arXiv preprint arXiv:2402.17911*, 2024.

[HHF+19] Robin Harper, Ian Hincks, Chris Ferrie, Steven T Flammia, and Joel J Wallman. Statistical analysis of randomized benchmarking. *Physical Review A*, 99(5):052350, 2019.

[HHJ+17] Jeongwan Haah, Aram W Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. *IEEE Transactions on Information Theory*, 63(9):5628–5641, 2017.

[HIR+21] Jonas Helsen, Marios Ioannou, Ingo Roth, Jonas Kitzinger, Emilio Onorati, Albert H. Werner, and Jens Eisert. Estimating gate-set properties from random sequences, 2021.

[HKP20] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, 2020.

[HKP21] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Information-theoretic bounds on quantum advantage in machine learning. *Phys. Rev. Lett.*, 126:190505, May 2021.

[HM17] Aram W Harrow and Ashley Montanaro. Quantum computational supremacy. *Nature*, 549(7671):203–209, 2017.

[HNG+24] Akel Hashim, Long B Nguyen, Noah Goss, Brian Marinelli, Ravi K Naik, Trevor Chistolini, Jordan Hines, JP Marceaux, Yosep Kim, Pranav Gokhale, et al. A practical introduction to benchmarking and characterization of quantum computers. *arXiv preprint arXiv:2408.12064*, 2024.

[HNM+21] Akel Hashim, Ravi K. Naik, Alexis Morvan, Jean-Loup Ville, Bradley Mitchell, John Mark Kreikebaum, Marc Davis, Ethan Smith, Costin Iancu, Kevin P. O'Brien, Ian Hincks, Joel J. Wallman, Joseph Emerson, and Irfan Siddiqi. Randomized compiling for scalable quantum computing on a noisy superconducting quantum processor. *Phys. Rev. X*, 11:041039, Nov 2021.

[Hoe94] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.

[Hol73] Alexander Semenovich Holevo. Bounds for the quantity of information transmitted by a quantum communication channel. *Problemy Peredachi Informatsii*, 9(3):3–11, 1973.

[Hol11] Alexander S Holevo. *Probabilistic and statistical aspects of quantum theory*, volume 1. Springer Science & Business Media, 2011.

[HP25] Jordan Hines and Timothy Proctor. Pauli noise learning for mid-circuit measurements. *Physical Review Letters*, 134(2):020602, 2025.

[HRO+20] Jonas Helsen, Ingo Roth, Emilio Onorati, Albert H Werner, and Jens Eisert. A general framework for randomized benchmarking. *arXiv preprint arXiv:2010.07974*, 2020.

[Hsi83] Cheng Hsiao. Identification. *Handbook of econometrics*, 1:223–283, 1983.

[HXVW19] Jonas Helsen, Xiao Xue, Lieven MK Vandersypen, and Stephanie Wehner. A new class of efficient randomized benchmarking protocols. *npj Quantum Information*, 5(1):1–9, 2019.

[HYF21a] Robin Harper, Wenjun Yu, and Steven T Flammia. Fast estimation of sparse quantum noise. *PRX Quantum*, 2(1):010322, 2021.

[HYF21b] Robin Harper, Wenjun Yu, and Steven T. Flammia. Fast estimation of sparse quantum noise. *PRX Quantum*, 2:010322, Feb 2021.

[ibm22] IBM Quantum. https://quantum-computing.ibm.com/services?services=systems, 2022.

[IJBE21] Pavithran Iyer, Aditya Jain, Stephen D Bartlett, and Joseph Emerson. Efficient diagnostics for quantum error correction. *arXiv preprint arXiv:2108.10830*, 2021.

[IKY+23] Asuka Inoue, Takahiro Kashiwazaki, Taichi Yamashima, Naoto Takanashi, Takushi Kazama, Koji Enbutsu, Kei Watanabe, Takeshi Umeki, Mamoru Endo, and Akira Furusawa. Toward a multi-core ultra-fast optical quantum processor: 43-ghz bandwidth real-time amplitude measurement of 5-db squeezed light using modularized optical parametric amplifier with 5g technology. *Applied Physics Letters*, 122(10):104001, 2023.

[Ing10] Tadeusz Inglot. Inequalities for quantiles of the chi-square distribution. *Probability and Mathematical Statistics*, 30(2):339–351, 2010.

[ITG+23] Mohsin Iqbal, Nathanan Tantivasadakarn, Thomas M Gatterman, Justin A Gerber, Kevin Gilmore, Dan Gresh, Aaron Hankin, Nathan Hewitt, Chandler V Horst, Mitchell Matheny, et al. Topological order from measurements and feed-forward on a trapped ion quantum computer. *arXiv preprint arXiv:2302.01917*, 2023.

[Jam72] Andrzej Jamiołkowski. Linear transformations which preserve trace and positive semidefiniteness of operators. *Reports on Mathematical Physics*, 3(4):275–278, 1972.

[JNG21] Atharv Joshi, Kyungjoo Noh, and Yvonne Y Gao. Quantum information processing with bosonic qubits in circuit qed. *Quantum Science and Technology*, 6(3):033001, 2021.

[KdSR+14] Shelby Kimmel, Marcus P. da Silva, Colm A. Ryan, Blake R. Johnson, and Thomas Ohki. Robust extraction of tomographic information via randomized benchmarking. *Phys. Rev. X*, 4:011050, Mar 2014.

[KEA+23] Youngseok Kim, Andrew Eddins, Sajant Anand, Ken Xuan Wei, Ewout Van Den Berg, Sami Rosenblatt, Hasan Nayfeh, Yantao Wu, Michael Zaletel, Kristan Temme, et al. Evidence for the utility of quantum computing before fault tolerance. *Nature*, 618(7965):500–505, 2023.

[Kim08] H Jeff Kimble. The quantum internet. *Nature*, 453(7198):1023–1030, 2008.

[KKEG19] Martin Kliesch, Richard Kueng, Jens Eisert, and David Gross. Guaranteed recovery of quantum processes from few measurements. *Quantum*, 3:171, 2019.

[KL04] Keri Kornelson and David Larson. Rank-one decomposition of operators and construction of frames. *Contemporary Mathematics*, 345:203–214, 2004.

[KLJ+22] Hyukgun Kwon, Youngrong Lim, Liang Jiang, Hyunseok Jeong, and Changhun Oh. Quantum metrological power of continuous-variable quantum networks. *Physical Review Letters*, 128(18):180503, 2022.

[KLR+08] Emanuel Knill, Dietrich Leibfried, Rolf Reichle, Joe Britton, R Brad Blakestad, John D Jost, Chris Langer, Roee Ozeri, Signe Seidelin, and David J Wineland. Randomized benchmarking of quantum gates. *Physical Review A*, 77(1):012307, 2008.

[Kni05] Emanuel Knill. Quantum computing with realistically noisy devices. *Nature*, 434(7029):39–44, 2005.

[KOPS15] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *Conference on Learning Theory*, pages 1066–1100. PMLR, 2015.

[LBØ+25] Zheng-Hao Liu, Romain Brunel, Emil EB Østergaard, Oscar Cordero, Senrui Chen, Yat Wong, Jens AH Nielsen, Axel B Bregnsbo, Sisi Zhou, Hsin-Yuan Huang, et al. Quantum learning advantage on a scalable photonic platform. *arXiv preprint arXiv:2502.07770*, 2025.

[LBZ02] Jay Lawrence, Časlav Brukner, and Anton Zeilinger. Mutually unbiased binary observable sets on n qubits. *Physical Review A*, 65(3):032320, 2002.

[LeC73] Lucien LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

[LLM22] Raymond Laflamme, Junan Lin, and Tal Mor. Algorithmic cooling for resolving state preparation and measurement errors in quantum computing. *arXiv preprint arXiv:2203.08114*, 2022.

[LOB+21] Yunchao Liu, Matthew Otten, Roozbeh Bassirianjahromi, Liang Jiang, and Bill Fefferman. Benchmarking near-term quantum computers via random circuit sampling, 2021.

[Mau86] AA Maudsley. Modified carr-purcell-meiboom-gill sequence for nmr fourier imaging applications. *Journal of Magnetic Resonance (1969)*, 69(3):488–491, 1986.

[MG58] Saul Meiboom and David Gill. Modified spin-echo method for measuring nuclear relaxation times. *Review of scientific instruments*, 29(8):688–691, 1958.

[MGE11] Easwar Magesan, J. M. Gambetta, and Joseph Emerson. Scalable and robust randomized benchmarking of quantum processes. *Phys. Rev. Lett.*, 106:180504, May 2011.

[MGE12] Easwar Magesan, Jay M. Gambetta, and Joseph Emerson. Characterizing quantum gates via randomized benchmarking. *Phys. Rev. A*, 85:042311, Apr 2012.

[MGJ+12] Easwar Magesan, Jay M Gambetta, Blake R Johnson, Colm A Ryan, Jerry M Chow, Seth T Merkel, Marcus P Da Silva, George A Keefe, Mary B Rothwell, Thomas A Ohki, et al. Efficient measurement of quantum gate error by interleaved randomized benchmarking. *Physical review letters*, 109(8):080505, 2012.

[MGS+13a] Seth T Merkel, Jay M Gambetta, John A Smolin, Stefano Poletto, Antonio D Córcoles, Blake R Johnson, Colm A Ryan, and Matthias Steffen. Self-consistent quantum process tomography. *Physical Review A*, 87(6):062119, 2013.

[MGS+13b] Seth T. Merkel, Jay M. Gambetta, John A. Smolin, Stefano Poletto, Antonio D. Córcoles, Blake R. Johnson, Colm A. Ryan, and Matthias Steffen. Self-consistent quantum process tomography. *Phys. Rev. A*, 87:062119, Jun 2013.

[MK05] Yongyi Mao and Frank R Kschischang. On factor graphs and the fourier transform. *IEEE Transactions on Information Theory*, 51(5):1635–1649, 2005.

[MLA+22] Lars S Madsen, Fabian Laudenbach, Mohsen Falamarzi Askarani, Fabien Rortais, Trevor Vincent, Jacob FF Bulmer, Filippo M Miatto, Leonhard Neuhaus, Lukas G Helt, Matthew J Collins, et al. Quantum computational advantage with a programmable photonic processor. *Nature*, 606(7912):75–81, 2022.

[MRL08] Masoud Mohseni, Ali T Rezakhani, and Daniel A Lidar. Quantum-process tomography: Resource analysis of different strategies. *Physical Review A*, 77(3):032322, 2008.

[MSB+16] Marios H Michael, Matti Silveri, RT Brierley, Victor V Albert, Juha Salmilehto, Liang Jiang, and Steven M Girvin. New class of quantum error-correcting codes for a bosonic mode. *Physical Review X*, 6(3):031006, 2016.

[MVM+23] A Morvan, B Villalonga, X Mi, S Mandra, A Bengtsson, PV Klimov, Z Chen, S Hong, C Erickson, IK Drozdov, et al. Phase transition in random circuit sampling. *arXiv preprint arXiv:2304.11119*, 2023.

[MZO20] Filip B. Maciejewski, Zoltán Zimborás, and Michał Oszmaniec. Mitigation of readout noise in near-term quantum devices by classical post-processing based on detector tomography. *Quantum*, 4:257, April 2020.

[NB19] Naomi H Nickerson and Benjamin J Brown. Analysing correlated noise on the surface code using adaptive decoding algorithms. *Quantum*, 3:131, 2019.

[NC11] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, USA, 10th edition, 2011.

[NGR+21] Erik Nielsen, John King Gamble, Kenneth Rudinger, Travis Scholten, Kevin Young, and Robin Blume-Kohout. Gate Set Tomography. *Quantum*, 5:557, October 2021.

[NOD19] Gergő Nemes and Adri Olde Daalhuis. Asymptotic expansions for the incomplete gamma function in the transition regions. *Mathematics of Computation*, 88(318):1805–1827, 2019.

[NYBK22] Erik Nielsen, Kevin Young, and Robin Blume-Kohout. First-order gauge-invariant error rates in quantum processors. *Bulletin of the American Physical Society*, 2022.

[OCW+24] Changhun Oh, Senrui Chen, Yat Wong, Sisi Zhou, Hsin-Yuan Huang, Jens AH Nielsen, Zheng-Hao Liu, Jonas S Neergaard-Nielsen, Ulrik L Andersen, Liang Jiang, et al. Entanglement-enabled advantage for learning a bosonic random displacement channel. *Physical Review Letters*, 133(23):230604, 2024.

[OJL22] Changhun Oh, Liang Jiang, and Changhyoup Lee. Distributed quantum phase sensing for arbitrary positive and negative weights. *Physical Review Research*, 4(2):023164, 2022.

[OLLJ20] Changhun Oh, Changhyoup Lee, Seok Hyung Lie, and Hyunseok Jeong. Optimal distributed quantum sensing using gaussian states. *Physical Review Research*, 2(2):023030, 2020.

[ORS+23] Corey Ostrove, Kenneth Rudinger, Stefan Seritan, Kevin Young, and Robin Blume-Kohout. Near-minimal gate set tomography experiment designs. In *2023 IEEE International Conference on Quantum Computing and Engineering (QCE)*, volume 1, pages 1422–1432. IEEE, 2023.

[PBP+24] Lukas Postler, Friederike Butt, Ivan Pogorelov, Christian D Marciniak, Sascha Heußen, Rainer Blatt, Philipp Schindler, Manuel Rispler, Markus Müller, and Thomas Monz. Demonstration of fault-tolerant steane quantum error correction. *PRX Quantum*, 5(3):030326, 2024.

[PCOG+24] Emilio Pelaez Cisneros, Victory Omole, Pranav Gokhale, Rich Rines, Kaitlin N Smith, Michael A Perlin, and Akel Hashim. Average circuit eigenvalue sampling on nisq devices. *arXiv e-prints*, pages arXiv–2403, 2024.

[PDF+21] Juan M Pino, Jennifer M Dreiling, Caroline Figgatt, John P Gaebler, Steven A Moses, MS Allman, CH Baldwin, M Foss-Feig, D Hayes, K Mayer, et al. Demonstration of the trapped-ion quantum ccd computer architecture. *Nature*, 592(7853):209–213, 2021.

[PK22] Pavel Panteleev and Gleb Kalachev. Asymptotically good quantum and locally testable classical ldpc codes. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 375–388, 2022.

[PLLP19] Stefano Pirandola, Riccardo Laurenza, Cosmo Lupo, and Jason L Pereira. Fundamental limits to quantum channel discrimination. *npj Quantum Information*, 5(1):1–8, 2019.

[PLOB17] Stefano Pirandola, Riccardo Laurenza, Carlo Ottaviani, and Leonardo Banchi. Fundamental limits of repeaterless quantum communications. *Nature communications*, 8(1):1–15, 2017.

[Pre18] John Preskill. Quantum computing in the NISQ era and beyond. *Quantum*, 2:79, 2018.

[PRY+17] Timothy Proctor, Kenneth Rudinger, Kevin Young, Mohan Sarovar, and Robin Blume-Kohout. What randomized benchmarking actually measures. *Physical review letters*, 119(13):130502, 2017.

[PVSS20] Emanuele Polino, Mauro Valeri, Nicolò Spagnolo, and Fabio Sciarrino. Photonic quantum metrology. *AVS Quantum Science*, 2(2):024703, 2020.

[PYBBK24] Timothy Proctor, Kevin Young, Andrew D Baczewski, and Robin Blume-Kohout. Benchmarking quantum computers. *arXiv preprint arXiv:2407.08828*, 2024.

[QSFK+24] Yihui Quek, Daniel Stilck França, Sumeet Khatri, Johannes Jakob Meyer, and Jens Eisert. Exponentially tighter bounds on limitations of quantum error mitigation. *Nature Physics*, 20(10):1648–1658, 2024.

[RABL+21] Ciaran Ryan-Anderson, Justin G Bohnet, Kenny Lee, Daniel Gresh, Aaron Hankin, JP Gaebler, David Francois, Alexander Chernoguzov, Dominic Lucchetti, Natalie C Brown, et al. Realization of real-time fault-tolerant quantum error correction. *Physical Review X*, 11(4):041058, 2021.

[RF88] Halsey Lawrence Royden and Patrick Fitzpatrick. *Real analysis*, volume 32. Macmillan New York, 1988.

[RF23] Cambyse Rouzé and Daniel Stilck França. Efficient learning of the structure and parameters of local pauli noise channels. *arXiv preprint arXiv:2307.02959*, 2023.

[RKK+18] Ingo Roth, Richard Kueng, Shelby Kimmel, Y-K Liu, David Gross, Jens Eisert, and Martin Kliesch. Recovering quantum gates from few average gate fidelities. *Physical review letters*, 121(17):170502, 2018.

[Ros22] Matteo Rosati. A learning theory for quantum photonic processors and beyond. *arXiv preprint arXiv:2209.03075*, 2022.

[RVH12] László Ruppert, Dániel Virosztek, and Katalin Hangos. Optimal parameter estimation of pauli channels. *Journal of Physics A: Mathematical and Theoretical*, 45(26):265305, 2012.

[RYCS21] Zane M Rossi, Jeffery Yu, Isaac L Chuang, and Sho Sugiura. Quantum advantage for noisy channel discrimination. *arXiv preprint arXiv:2105.08707*, 2021.

[RYCS22] Zane M Rossi, Jeffery Yu, Isaac L Chuang, and Sho Sugiura. Quantum advantage for noisy channel discrimination. *Physical Review A*, 105(3):032401, 2022.

[SAS11] Alexandre M Souza, Gonzalo A Alvarez, and Dieter Suter. Robust dynamical decoupling for quantum computing and quantum memory. *Physical review letters*, 106(24):240501, 2011.

[SBA+23] K Singh, CE Bradley, S Anand, V Ramesh, R White, and H Bernien. Mid-circuit correction of correlated phase errors using an array of spectator qubits. *Science*, 380(6651):1265–1269, 2023.

[Sch11] Wolfgang P Schleich. *Quantum optics in phase space*. John Wiley & Sons, 2011.

[SCM+24] Alireza Seif, Senrui Chen, Swarnadeep Majumder, Haoran Liao, Derek S Wang, Moein Malekakhlagh, Ali Javadi-Abhari, Liang Jiang, and Zlatko K Minev. Entanglement-enhanced learning of quantum processes at scale. *arXiv preprint arXiv:2408.03376*, 2024.

[SEFT22] Yasunari Suzuki, Suguru Endo, Keisuke Fujii, and Yuuki Tokunaga. Quantum error mitigation as a universal error reduction technique: Applications from the nisq to the fault-tolerant quantum computing eras. *PRX Quantum*, 3(1):010345, 2022.

[Ser17] Alessio Serafini. *Quantum continuous variables: a primer of theoretical methods*. CRC press, 2017.

[Sho96] Peter W Shor. Fault-tolerant quantum computation. In *Proceedings of 37th Conference on Foundations of Computer Science*, pages 56–65. IEEE, 1996.

[Sho99] Peter W Shor. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM review*, 41(2):303–332, 1999.

[SHS+07] DI Schuster, Andrew Addison Houck, JA Schreier, A Wallraff, JM Gambetta, A Blais, L Frunzio, J Majer, B Johnson, MH Devoret, et al. Resolving photon number states in a superconducting circuit. *Nature*, 445(7127):515–518, 2007.

[Sik83] Pierre Sikivie. Experimental tests of the" invisible" axion. *Physical Review Letters*, 51(16):1415, 1983.

[SLT+24] Alireza Seif, Haoran Liao, Vinay Tripathi, Kevin Krsulich, Moein Malekakhlagh, Mirko Amico, Petar Jurcevic, and Ali Javadi-Abhari. Suppressing correlated noise in quantum computers via context-aware compiling. In *Proceedings of the 51st Annual International Symposium on Computer Architecture*. IEEE, 2024.

[SPR+20] Mohan Sarovar, Timothy Proctor, Kenneth Rudinger, Kevin Young, Erik Nielsen, and Robin Blume-Kohout. Detecting crosstalk errors in quantum information processors. *Quantum*, 4:321, 2020.

[SZ19] Xiaoming Sun and Yufan Zheng. Hybrid decision trees: Longer quantum time is strictly more powerful. *arXiv preprint arXiv:1911.13091*, 2019.

[SZ23] Haowei Shi and Quntao Zhuang. Ultimate precision limit of noise sensing and dark matter search. *npj Quantum Information*, 9(1):27, 2023.

[TBF18] David K Tuckett, Stephen D Bartlett, and Steven T Flammia. Ultrahigh error threshold for surface codes with biased noise. *Physical review letters*, 120(5):050505, 2018.

[TBG17] Kristan Temme, Sergey Bravyi, and Jay M Gambetta. Error mitigation for short-depth quantum circuits. *Physical review letters*, 119(18):180509, 2017.

[Ter15] Barbara M Terhal. Quantum error correction for quantum memories. *Reviews of Modern Physics*, 87(2):307, 2015.

[TSY23] Kento Tsubouchi, Takahiro Sagawa, and Nobuyuki Yoshioka. Universal cost bound of quantum error mitigation based on quantum estimation theory. *Physical Review Letters*, 131(21):210601, 2023.

[TY24] Xuan Du Trinh and Nengkun Yu. Adaptivity is not helpful for pauli channel learning. *arXiv preprint arXiv:2403.09033*, 2024.

[VDBMKT23] Ewout Van Den Berg, Zlatko K Minev, Abhinav Kandala, and Kristan Temme. Probabilistic error cancellation with sparse pauli–lindblad models on noisy quantum processors. *Nature Physics*, 2023.

[vdBW24] Ewout van den Berg and Pawel Wocjan. Techniques for learning sparse pauli-lindblad noise models. *Quantum*, 8:1556, 2024.

[VKL99] Lorenza Viola, Emanuel Knill, and Seth Lloyd. Dynamical decoupling of open quantum systems. *Physical Review Letters*, 82(12):2417, 1999.

[VNBT24] Christophe H Valahu, Tomas Navickas, Michael J Biercuk, and Ting Rei Tan. Benchmarking bosonic modes for quantum information with randomized displacements. *arXiv preprint arXiv:2405.15237*, 2024.

[VPRK97] Vlatko Vedral, Martin B Plenio, Michael A Rippin, and Peter L Knight. Quantifying entanglement. *Physical Review Letters*, 78(12):2275, 1997.

[WBC+21] Yulin Wu, Wan-Su Bao, Sirui Cao, Fusheng Chen, Ming-Cheng Chen, Xiawei Chen, Tung-Hsun Chung, Hui Deng, Yajie Du, Daojin Fan, et al. Strong quantum computational advantage using a superconducting quantum processor. *arXiv preprint arXiv:2106.14734*, 2021.

[WBD+19] K Wright, KM Beck, Sea Debnath, JM Amini, Y Nam, N Grzesiak, J-S Chen, NC Pisenti, M Chmielewski, C Collins, et al. Benchmarking an 11-qubit quantum computer. *Nature communications*, 10(1):1–6, 2019.

[WCL23] Ya-Dong Wu, Giulio Chiribella, and Nana Liu. Quantum-enhanced learning of continuous-variable quantum states. *arXiv preprint arXiv:2303.05097*, 2023.

[WE16] Joel J Wallman and Joseph Emerson. Noise tailoring for scalable quantum computation via randomized compiling. *Physical Review A*, 94(5):052325, 2016.

[WF89] William K Wootters and Brian D Fields. Optimal state-determination by mutually unbiased measurements. *Annals of Physics*, 191(2):363–381, 1989.

[WFH18] Joel Wallman, Steven Flammia, and Ian Hincks. Quantum characterization, verification, and validation. In *Oxford Research Encyclopedia of Physics*. 2018.

[Wil13] Mark M Wilde. *Quantum information theory*. Cambridge University Press, 2013.

[WKBK23] Thomas Wagner, Hermann Kampermann, Dagmar Bruß, and Martin Kliesch. Learning logical pauli noise in quantum error correction. *Physical review letters*, 130(20):200601, 2023.

[WPGP+12] Christian Weedbrook, Stefano Pirandola, Raúl García-Patrón, Nicolas J Cerf, Timothy C Ralph, Jeffrey H Shapiro, and Seth Lloyd. Gaussian quantum information. *Reviews of Modern Physics*, 84(2):621, 2012.

[WS19] Ya-Dong Wu and Barry C Sanders. Efficient verification of bosonic quantum channels via benchmarking. *New Journal of Physics*, 21(7):073026, 2019.

[XBAP+24] Qian Xu, J Pablo Bonilla Ataides, Christopher A Pattison, Nithin Raveendran, Dolev Bluvstein, Jonathan Wurtz, Bane Vasić, Mikhail D Lukin, Liang Jiang, and Hengyun Zhou. Constant-overhead fault-tolerant quantum computation with reconfigurable atom arrays. *Nature Physics*, 20(7):1084–1090, 2024.

[XLC+20] Yi Xia, Wei Li, William Clark, Darlene Hart, Quntao Zhuang, and Zheshen Zhang. Demonstration of a reconfigurable entangled radio-frequency photonic sensor network. *Physical Review Letters*, 124(15):150502, 2020.

[Yu97] Bin Yu. Assouad, fano, and le cam. *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pages 423–435, 1997.

[ZCLJ25] Zhihan Zhang, Senrui Chen, Yunchao Liu, and Liang Jiang. Generalized cycle benchmarking algorithm for characterizing midcircuit measurements. *PRX Quantum*, 6(1):010310, 2025.

[ZDQ+21] Han-Sen Zhong, Yu-Hao Deng, Jian Qin, Hui Wang, Ming-Cheng Chen, Li-Chao Peng, Yi-Han Luo, Dian Wu, Si-Qiu Gong, Hao Su, et al. Phase-programmable Gaussian boson sampling using stimulated squeezed light. *Physical review letters*, 127(18):180502, 2021.

[Zha13] Xingzhi Zhan. *Matrix theory*, volume 147. American Mathematical Soc., 2013.

[ZWD+20] Han-Sen Zhong, Hui Wang, Yu-Hao Deng, Ming-Cheng Chen, Li-Chao Peng, Yi-Han Luo, Jian Qin, Dian Wu, Xing Ding, Yi Hu, et al. Quantum computational advantage using photons. *Science*, 370(6523):1460–1463, 2020.

[ZZPJ18] Sisi Zhou, Mengzhen Zhang, John Preskill, and Liang Jiang. Achieving the heisenberg limit in quantum metrology using quantum error correction. *Nature communications*, 9(1):78, 2018.

[ZZS18] Quntao Zhuang, Zheshen Zhang, and Jeffrey H Shapiro. Distributed quantum sensing using continuous-variable multipartite entanglement. *Physical Review A*, 97(3):032329, 2018.