# CONSERVATION LAWS IN PARAMETERIZED QUANTUM CIRCUITS: COMPLETE CHARACTERIZATIONS AND RANK SEPARATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Conservation laws provide a precise description of parameter combinations that cannot change under gradient flow. While such invariants have been studied extensively in classical neural networks, their role in parameterized quantum circuits (PQCs) is substantially less understood because the trainable unitary, the measurement observable, and the circuit architecture jointly determine which parameter directions are visible to optimization. We develop a geometric theory of local smooth data-independent conservation laws for PQCs and use it to characterize the active training dimension of widely used Basic Entangler and Strongly Entangling architectures. Using the differential geometry of the measurement map, we expose two distinct sources of inactive directions: unitary-parametrization redundancies and nontrivial unitary motions that are invisible to the readout observable. For commuting trainable blocks, we show that all conservation laws arise from linear invariants associated with a fixed matrix nullspace, yielding a complete characterization of Basic Entangler circuits at arbitrary width and depth. For noncommuting circuits, we derive stabilizer-induced laws and establish completeness for two-layer Strongly Entangling circuits and a class of two-layer circuits with Clifford entanglers. We further construct circuits whose Lie closure rank exceeds the instantaneous Jacobian rank by an amount that grows with circuit width, showing that invisible directions at a single point can overestimate the number of conservation laws. Experiments on synthetic data and CIFAR-10 confirm the predicted invariants during training.

## 1 INTRODUCTION

The dynamics of gradient-based learning is shaped not only by the objective function but also by structural constraints imposed by the model parametrization. In classical deep learning, this has motivated extensive studies of implicit bias, parameter symmetries, and conservation laws in overparameterized models (Saxe et al., 2013; Du et al., 2018; Bah et al., 2022; Marcotte et al., 2023; Peleg & Hein, 2024). Conservation laws identify initialization dependent quantities that remain invariant during training, revealing inactive parameter directions and reductions in the effective optimization dimension while motivating reparameterizations that remove redundant degrees of freedom (Stock et al., 2019; Saul, 2023; Robertson & Koyejo, 2023; Marcotte et al., 2026).

**Conservation Laws.** To characterize implicit bias in linearized models, Marcotte et al. (2023) developed a framework for identifying complete families of parameter dependent quantities conserved under Euclidean gradient flow. This framework was later extended to modern architectures (Marcotte et al., 2025; Tran et al., 2026), including skip connections He et al. (2016), attention Vaswani et al. (2017), and Mixture-of-Experts models Jacobs et al. (1991). Despite this progress in classical machine learning, conservation laws in quantum learning models remain largely unexplored.

**Quantum Neural Network (QNN).** Training dynamics in QNN differ substantially from those of classical neural networks because the behavior of PQCs is jointly determined by the circuit architecture and the measurement observable (Cerezo et al., 2021a; Du et al., 2022; You et al., 2023). Trainable parameters influence the objective through both the implemented unitary and the resulting measurements. Consequently, a parameter direction may become inactive either because it is



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

redundant in the unitary parametrization or because it changes the unitary while remaining invisible to the readout observable. This viewpoint is closely related to prior studies of effective dimension and Jacobian rank. In particular, Funcke et al. (2021) showed that the number of effective parameter directions can be smaller than the nominal parameter count, while Haug et al. (2021) used quantum geometric analysis to identify and remove redundant parameters without reducing the effective dimension. These approaches, however, do not directly characterize quantities that remain invariant throughout training or the parameter relations that induce such invariance. More recently, (Peng et al., 2026) empirically observed that some parameters can be frozen from initialization with little degradation in performance. Our work explains this phenomenon by providing a rigorous theoretical account through conservation laws.

**Challenges for PQC.** Characterizing conservation laws in PQCs is challenging because they depend on both circuit architecture and the measurement observable. For commuting trainable gates, the nullspace of the measurement Jacobian remains fixed, while Lie brackets add no new directions. Thus, identifying conservation laws reduces to a linear algebra problem. In noncommuting circuits, however, the Jacobian kernel generally varies with the trainable parameters, making pointwise kernel analysis insufficient. Locally conserved directions may be eliminated by Lie brackets, so counting pointwise kernel directions can overestimate the number of independent conservation laws. We provide a principled criterion for certifying completeness and characterize this rank gap for a class of two layer Clifford circuits.

**Contribution.** Motivated by the widespread use of Basic Entangler and Strongly Entangling Layers in practical quantum learning (Qi et al., 2023; Liu et al., 2025; Holzer & Turkalj, 2026), we study these two circuit families as defined in the PennyLane framework (Bergholm et al., 2018) and characterize functions of their trainable parameters that remain conserved throughout gradient flow. For brevity, throughout the remainder of the paper, we refer to these architectures as the basic PQC and the strong PQC, respectively. Our contributions are threefold.

(i) We derive a partial differential equation (PDE) whose solutions completely characterize all smooth conservation laws of the training dynamics.

(ii) We provide a complete characterization of conservation laws for commuting trainable blocks, including Basic PQCs of arbitrary width and depth. For noncommuting architectures, we derive stabilizer-induced invariants and establish completeness for two-layer Strong PQCs and a broader class of Clifford-entangler circuits with string-$Z$ readout. For the latter, we further prove a sharp separation between the instantaneous Jacobian rank and the Lie-closure rank, with the gap increasing linearly with circuit width.

(iii) Finally, experiments on synthetic data and CIFAR-10 confirm that the theoretically predicted invariants remain conserved throughout training.

**Notation.** We denote by $\mathrm{Herm}(d)$, $\mathrm{U}(d)$, and $\mathrm{SU}(d)$ the real vector space of $d \times d$ Hermitian matrices, the group of $d \times d$ unitary matrices, and the special unitary group, respectively. Moreover, we denote $\mathfrak{u}(d) = \{A \in \mathbb{C}^{d\times d} : A^\dagger = -A\}$ for the real Lie algebra of skew-Hermitian matrices, which is the Lie algebra of $\mathrm{U}(d)$. The Hilbert-Schmidt inner product is defined as $\langle A, B\rangle_{\mathrm{HS}} = \mathrm{tr}(AB)$. For a real linear map $T : \mathbb{R}^p \to \mathrm{Herm}(d)$, the adjoint $T^* : \mathrm{Herm}(d) \to \mathbb{R}^p$ is always taken with respect to $\langle\cdot,\cdot\rangle_{\mathrm{HS}}$ on $\mathrm{Herm}(d)$ and the standard Euclidean inner product on $\mathbb{R}^p$. For matrices $A, B$, the commutator is $[A, B] = AB - BA$. For matrices $A_1, \ldots, A_m$, $\mathrm{Lie}_{\mathbb{R}}\langle A_1, \ldots, A_m\rangle$ denotes the smallest real Lie subalgebra containing them. For a vector field $X$ and a function $h$, we write $Xh(\theta) = \mathrm{d}h_\theta(X(\theta)) = \nabla h(\theta)^\top X(\theta)$ for the directional derivative of $h$ along $X$.

**Organization.** Section 2 introduces the definition of QNN, the quantum circuit architectures, and conservation laws. Section 3 characterizes data-independent conservation laws through the measurement differential and its Lie closure, while Section 4 applies this framework to Basic and Strong PQCs. Section 5 presents the numerical results, and Section 6 concludes the paper. Proofs and additional experiments are deferred to the appendix.

## 2 PRELIMINARIES

In this section, we first provide a brief overview of QNNs and introduce the Basic and Strong PQC architectures commonly used in practical quantum learning. We then describe the Euclidean gradient dynamics considered in this work and formally define conservation laws.

### 2.1 Quantum Neural Networks

A QNN can be viewed as the quantum analogue of a classical neural network, mapping inputs to outputs through a parameterized function. Its key distinction is that it is built from parameterized quantum circuits, or *ansatz*, which can be executed on quantum hardware. For an $n$-qubit system, a quantum state in a $(d = 2^n)$-dimensional Hilbert space is represented by a density matrix $\rho \in \mathbb{C}^{d\times d}$, a positive semidefinite Hermitian matrix with unit trace. Such states may arise from physical quantum systems (Cong et al., 2019) or by encoding classical data into quantum states (Hur et al., 2022). Given an input state $\rho \in \mathrm{Herm}(d)$, a PQC is represented by a smooth map $U : \Theta \to \mathrm{U}(d)$, where $\Theta \subseteq \mathbb{R}^p$. The quantum prediction output is calculated by $F_\rho(\theta) = \mathrm{tr}(\rho\Phi(\theta))$. Here, $\Phi : \Theta \to \mathrm{Herm}(d)$, defined by $\Phi(\theta) = U(\theta)^\dagger M_0 U(\theta)$ denotes the measurement map and $M_0 \in \mathrm{Herm}(d)$ is a fixed observable. A widely used class of PQCs can be written in the form

$$U(\theta) := E_p R_p(\theta_p) \cdots E_1 R_1(\theta_1), \tag{1}$$

where each $E_a \in \mathrm{U}(d)$ denotes a fixed unitary transformation, while $R_a(\theta_a) = e^{-i\theta_a H_a/2}$ is a parameterized unitary rotation generated by a Hermitian operator $H_a \in \mathrm{Herm}(d)$. Here, $\theta = (\theta_1, \ldots, \theta_p)$ collects the trainable parameters of the circuit.

**Basic PQC.** The $L$-layered Basic PQC can be written as

$$U^B(\theta) := E_L^B R_L^B E_{L-1}^B R_{L-1}^B \cdots E_1^B R_1^B, \tag{2}$$

where $E_\ell^B := \mathrm{CNOT}_{n\to 1}\,\mathrm{CNOT}_{n-1\to n}\cdots\mathrm{CNOT}_{1\to 2}$ is the fixed CNOT ring for every layer $\ell$ and $R_\ell^B$ is a parameterized layer defined by

$$R_\ell^B = \bigotimes_{j=1}^{n} R_X(\theta_{\ell,j}) = \exp\left(-\frac{i}{2}\sum_{j=1}^{n}\theta_{\ell,j}X_j\right), \forall\, 1 \le \ell \le L, \tag{3}$$

where $\theta_{\ell,j}$ is a trainable parameter, and $X_j$ is the Pauli-$X$ operator on the $j$-th qubit. For $n = 2$, we use a single CNOT gate in each layer, namely $E_\ell^{\mathrm{B}} = \mathrm{CNOT}_{1\to 2}$.

**Strong PQC.** The $L$-layered strong PQC can also be written as

$$U^S(\theta) := E_L^S R_L^S E_{L-1}^S R_{L-1}^S \cdots E_1^S R_1^S. \tag{4}$$

However, compared with the Basic PQC, the Strong PQC adopts a richer structure in both its fixed entangling operations and trainable single-qubit rotations. For each layer $1 \le \ell \le L$, we choose an entangling range $\tau_\ell \in \{1, \ldots, n-1\}$. The corresponding fixed entangling unitary is defined as a range-$\tau_\ell$ ring of CNOT gates $E_\ell^S = \mathrm{CNOT}_{n\to n+\tau_\ell}\cdots\mathrm{CNOT}_{2\to 2+\tau_\ell}\,\mathrm{CNOT}_{1\to 1+\tau_\ell}$. The trainable block of the $\ell$-th layer is given by

$$R_\ell^S = \bigotimes_{j=1}^{n} R_{\ell,j} = \prod_{j=1}^{n} R_Z^{(j)}(\omega_{\ell,j}) R_Y^{(j)}(\beta_{\ell,j}) R_Z^{(j)}(\phi_{\ell,j}), \tag{5}$$

where $R_{\ell,j} = R_Z(\omega_{\ell,j})R_Y(\beta_{\ell,j})R_Z(\phi_{\ell,j})$ is a $2\times 2$ single-qubit block, and $R_T(\theta) := \exp(-i\theta T/2)$ for $T \in \{X, Y, Z\}$. The corresponding rotation on the full $n$-qubit system is $R_T^{(j)}(\theta) := \exp(-i\theta T_j/2)$, where $X_j$, $Y_j$, and $Z_j$ denote the Pauli-$X$, Pauli-$Y$, and Pauli-$Z$ operators acting on the $j$-th qubit and as the identity on all other qubits, respectively.

### 2.2 Conservation Laws in Gradient Flows

We now consider the quantum learning setting, where the dataset is given by $\mathcal{D} = \{(\rho_i, y_i)\}_{i=1}^N$, with $\rho_i \in \mathrm{Herm}(d)$ denoting the quantum inputs and $y_i \in \mathcal{Y} \subseteq \mathbb{R}$ representing the corresponding target outputs. To train the PQC, we minimize the following objective function:

$$L_{\mathcal{D}}(\theta) := \frac{1}{N}\cdot\sum_{i=1}^{N} l(F_{\rho_i}(\theta), y_i), \tag{6}$$

where $l : \mathbb{R}\times\mathcal{Y} \to \mathbb{R}$ is the loss function. In practice, the training dynamic is often modeled by the Euclidean gradient flow

$$\dot\theta(t) = -\eta\,\nabla_\theta L_{\mathcal{D}}(\theta(t)), \tag{7}$$

where $\eta$ is the learning rate and $\theta(0)$ is the initialized parameter. Our objective is to characterize the functions that remain conserved along the flow (7), which motivates the following definition.

**Definition 2.1.** *A function $h \in C^1(\Theta, \mathbb{R})$ is called a conservation law if it remains unchanged along the training flow* (7)*. Equivalently, $h(\theta(t)) = h(\theta(0))$ for all $t \geq 0$.*

Under the above definition, such a function $h(\cdot)$ is locally conserved on any open subset $\Omega \subset \Theta$. Accordingly, our goal is to characterize functions that remain invariant under local variations of the parameter $\theta$ induced by the flow (7).

**Remark 2.2.** *In this work, we primarily focus on regression settings, where the objective function can be rewritten as $L_{\mathcal{D}}(\theta) := 1/N \cdot \sum_{i=1}^{N}(F_{\rho_i}(\theta) - y_i)^2$. For loss functions such as cross entropy, conservation laws can be characterized in a similar way, as illustrated in (Marcotte et al., 2025).*

**Remark 2.3.** *In practical training settings, the optimization dynamics may additionally incorporate weight decay or follow second-order differential equations. Nevertheless, as shown in (Marcotte et al., 2024), the methodology for identifying conservation laws extends to these settings in a largely analogous manner. We therefore omit these extensions for simplicity, allowing us to present the essential structure underlying conservation laws in quantum machine learning.*

## 3 Data-Independent Conservation Laws in QNN Training

In this section, we characterize the conservation laws arising in QNN training that hold uniformly over arbitrary datasets and initializations. We then derive a PDE whose solution is equivalent to identifying all possible conservation laws.

### 3.1 Data-Independent Conservation Laws

Our goal is to characterize conservation laws arising in QNN training. Since the objective (6) depends explicitly on the dataset $\mathcal{D}$, a general characterization is intractable. Following (Marcotte et al., 2023), we therefore focus on data-independent conservation laws, namely functions $h(\cdot)$ that remain invariant under the training dynamics for every dataset $\mathcal{D}$. Proposition 3.1 provides the first characterization: $h$ is a conservation law if and only if its gradient is orthogonal to every attainable gradient of the one-sample loss.

**Proposition 3.1.** *A function $h \in C^1(\Theta, \mathbb{R})$ is a conservation law if and only if*

$$\nabla h(\theta) \perp \mathcal{W}_{\text{loss}}(\theta), \forall\, \theta \in \Theta, \tag{8}$$

*where $\mathcal{W}_{\text{loss}}(\theta) = \operatorname{span}_{\substack{\rho \succeq 0,\ \operatorname{tr}\rho = 1 \\ y \in \mathcal{Y}}} \left\{ \nabla_\theta l(F_\rho(\theta), y) \right\}$ is the space of all one-sample gradients.*

The proof for Proposition 3.1 is in Appendix B.2.1. For the squared loss function, we obtain that $\mathcal{W}_{\text{loss}}(\theta) = \operatorname{span}_{\substack{\rho \succeq 0,\ \operatorname{tr}\rho = 1 \\ y \in \mathcal{Y}}} \left\{ \big(F_\rho(\theta) - y\big) \nabla_\theta F_\rho(\theta) \right\}$. Since the residuals $F_\rho(\theta) - y$ range over $\mathbb{R}$, it does not affect the span of the gradients. Consequently, $\mathcal{W}_{\text{loss}}(\theta) = \mathcal{W}_Q(\theta) = \operatorname{span}_\rho \{\nabla_\theta F_\rho(\theta)\}$. Because $\rho \in \operatorname{Herm}(d)$, which is a $d^2$-dimensional real vector space, the gradient distribution $\mathcal{W}_Q(\theta)$ can be represented by $d^2$ explicit vector fields, as shown in the following proposition.

**Proposition 3.2.** *Let $B_1, \ldots, B_{d^2}$ be a real basis of $\operatorname{Herm}(d)$. We define the probe function*

$$f_\alpha(\theta) = \operatorname{tr}\big(B_\alpha \Phi(\theta)\big), \forall \alpha = 1, \ldots, d^2. \tag{9}$$

*Then, for every $\theta \in \Theta$, we have $\mathcal{W}_Q(\theta) = \operatorname{span}_\alpha \{\nabla_\theta f_\alpha(\theta)\}$.*

The proof for Proposition 3.2 is provided in Appendix B.2.2. Building on this proposition, we derive a PDE that characterizes all conservation laws through their partial derivatives with respect to the trainable parameters $\theta_i$, where $\theta_i$ denotes the $i$-th component of the parameter vector $\theta$. This PDE is stated formally in Theorem 3.3 as follows

**Theorem 3.3** (PDE for QNN conservation laws)**.** *A function $h \in C^1(\Theta, \mathbb{R})$ is a conservation law if and only if*

$$\mathrm{d}\Phi_\theta\big(\nabla h(\theta)\big) = 0, \forall\, \theta \in \Theta, \tag{10}$$

*where $\mathrm{d}\Phi_\theta$ is the Jacobian of the function $\Phi(\cdot)$ at $\theta$.*

The proof of Theorem 3.3 is provided in Appendix B.2.3. We recall that our goal is to characterize all conservation laws induced by the training dynamics. While the PDE in (10) can verify whether a given function is conserved, it does not determine the number of functionally independent conservation laws. First, we introduce the notion of independence as follows:

**Definition 3.4.** *A collection of conserved functions* $\{h_i\}_{i=1}^n \subset C^1(\Theta, \mathbb{R})$ *is called* functionally independent *if their gradients* $\{\nabla h_1(\theta), \dots, \nabla h_n(\theta)\}$ *are linearly independent for every* $\theta \in \Theta$.

It is worth noting that annihilation is preserved under Lie brackets: if $X_1 h = X_2 h = 0$, then $[X_1, X_2]h = 0$. Therefore, any conservation law must annihilate the full Lie closure generated by the measurement-gradient fields, and the number of independent conservation laws is governed by the resulting involutive distribution. The following theorem characterizes the number of independent conservation laws in terms of this Lie closure.

**Theorem 3.5.** *Suppose that the distribution* $\theta \mapsto \mathrm{Lie}(\mathcal{W}_Q)(\theta)$ *has constant rank* $r$ *on an open neighborhood* $\Omega \subseteq \Theta$ *of* $\theta_0$. *Then there is an open neighborhood* $\Omega_0 \subseteq \Omega$ *of* $\theta_0$ *on which the maximal number of functionally independent smooth data-independent conservation laws is exactly* $p - r$.

The proof of Theorem 3.5 is provided in (Marcotte et al., 2023, Theorem 3.3). This theorem allows us to verify whether the conservation laws derived in Section 4 are complete. We next study the differential structure of the measurement map $\Phi(\cdot)$ and relate it to the PDE (10).

### 3.2 Measurement Differential and Observable Stabilizers

The PDE (10) reduces the study of conservation laws to the kernel of the measurement differential $\mathrm{d}\Phi_\theta$. We now investigate the infinitesimal geometry of this differential and characterize its kernel in terms of the unitary parametrization and the stabilizer of the evolved observable. For each parameter coordinate $\theta_a$, we define the Hermitian and skew-Hermitian generators associated with the parametrized quantum circuit $U : \Theta \to \mathrm{U}(d)$ as

$$K_a(\theta) := iU(\theta)^\dagger \partial_a U(\theta), \qquad A_a(\theta) := U(\theta)^\dagger \partial_a U(\theta) = -iK_a(\theta), \forall\, \theta \in \Theta.$$

Taking the derivative of the identity $U^\dagger U = I$ with respect to $\theta_a$ shows that $A_a(\theta) \in \mathfrak{u}(d)$ and therefore $K_a(\theta) \in \mathrm{Herm}(d)$. Moreover, the measurement derivative takes the commutator form

$$\partial_a \Phi(\theta) = i\,[K_a(\theta), \Phi(\theta)] = [\Phi(\theta), A_a(\theta)], \forall\, \theta \in \Theta. \tag{11}$$

A detailed verification of equation (11) is given in Appendix B.2.4. To characterize the kernel of the measurement differential, we first express the differential of the unitary parametrization in Lie-algebra coordinates. For each $\theta \in \Theta$, define the differential map $\kappa_\theta : T_\theta\Theta \simeq \mathbb{R}^p \to \mathfrak{u}(d)$ by

$$\kappa_\theta(v) := U(\theta)^\dagger \mathrm{d}U_\theta(v) = \sum_{a=1}^{p} v_a A_a(\theta), \forall\, v = (v_1, \dots, v_p)^\top \in T_\theta\Theta,$$

where $T_\theta\Theta$ is the tangent space of $\Theta$ at $\theta$. Thus, $\kappa_\theta$ identifies an infinitesimal variation of the parameters with the corresponding tangent direction in the Lie algebra $\mathfrak{u}(d)$. Now, we establish some results that capture the differential geometry of the measurement.

**Theorem 3.6** (Kernel and rank of the measurement differential)**.** *For every* $\theta \in \Theta$, *the following statements hold:*

*(i)* $\mathrm{d}\Phi_\theta(v) = [\Phi(\theta), \kappa_\theta(v)], \forall\, v \in T_\theta\Theta$.

*(ii)* $\ker \mathrm{d}\Phi_\theta = \kappa_\theta^{-1}\big(\mathfrak{z}_{\Phi(\theta)}\big) = \Big\{v \in \mathbb{R}^p : \sum_{a=1}^p v_a A_a(\theta) \in \mathfrak{z}_{\Phi(\theta)}\Big\}$, *where we denote* $\mathfrak{z}_{\Phi(\theta)} = \{M \in \mathfrak{u}(d) : [\Phi(\theta), M] = 0\}$ *as the stabilizer Lie algebra under unitary conjugation.*

*(iii)* $\mathrm{rank}\, \mathrm{d}\Phi_\theta = \mathrm{rank}\, \kappa_\theta - \dim\big(\mathrm{im}\, \kappa_\theta \cap \mathfrak{z}_{\Phi(\theta)}\big)$.

The proof of Theorem 3.6 is provided in Appendix B.2.5. The above results characterize the local geometry of the measurement map. Statement (i) relates its differential to infinitesimal unitary variations, (ii) identifies the kernel with directions that stabilize the evolved observable, and (iii) expresses the differential rank after removing stabilizer directions. We next use this characterization to derive explicit conservation laws for concrete QNN architectures.

# 4 EXPLICIT CONSERVATION LAWS IN QNN ARCHITECTURES

This section uses the characterization of differential measurement to construct explicit conservation laws for well-known QNN architectures. We first study commuting trainable blocks and obtain a complete characterization for Basic PQCs. We then investigate noncommuting Strong PQCs and extend our analysis to two-layer Clifford circuits, establishing completeness and examining the gap between the measurement-Jacobian and Lie-closure ranks. Finally, we discuss the implications of these laws for training efficiency and gradient estimation.

## 4.1 COMPLETE CONSERVATION LAWS FOR COMMUTING CIRCUITS

We first focus on commuting trainable blocks, in which all blocks commute pairwise. Under this assumption, the kernel of the measurement Jacobian in Theorem 3.6 can be characterized explicitly. This leads to a fixed subspace of conserved directions that can be determined exactly from the circuit generators and the measurement observable.

**Theorem 4.1** (Exact laws of a commuting trainable block)**.** *Suppose the ansatz can be written as*

$$U(\theta) = E_L \exp\left(-\frac{i}{2}\sum_{a=1}^{p}\theta_a H_a\right) E_R,$$

*where the Hermitian generators $H_1, \ldots, H_p$ commute pairwise. Let $\Phi_L = E_L^\dagger M_0 E_L$ and define*

$$\mathcal{V}_{\mathrm{com}} = \left\{\lambda \in \mathbb{R}^p : \left[\sum_{a=1}^{p}\lambda_a H_a, \Phi_L\right] = 0\right\}.$$

*Let $B_1, \ldots, B_{d^2}$ be a Hilbert–Schmidt orthonormal basis of $\mathrm{Herm}(d)$, and define $\Gamma_{\mathrm{com}} \in \mathbb{R}^{d^2 \times p}$ by $(\Gamma_{\mathrm{com}})_{\alpha a} = \mathrm{tr}(B_\alpha\, i[H_a, \Phi_L])$. Then, we obtain that $\ker \mathrm{d}\Phi_\theta = \mathcal{V}_{\mathrm{com}} = \ker \Gamma_{\mathrm{com}}, \forall\theta \in \Theta$. Thus, every basis vector of $\ker(\Gamma_{\mathrm{com}})$ defines a fundamental linear conservation law, and together these vectors characterize all independent conserved functions.*

The proof of Theorem 4.1 is provided in Appendix B.3.1. This theorem reveals two sources of conservation laws: redundancies in the unitary parametrization and nontrivial generator combinations that commute with the effective observable $\Phi_L$. Both are encoded by the nullspace of $\Gamma_{\mathrm{com}}$. Now, we apply this characterization to the practical Basic PQC architecture, whose trainable block can be expressed in terms of fixed, pairwise commuting generators.

**Theorem 4.2** (Complete laws for Basic PQC)**.** *Let $C_0 = I_{2^n}$ and $C_\ell = E_\ell^{\mathrm{B}} \cdots E_1^{\mathrm{B}}$ for $1 \le \ell \le L$. For each layer $\ell$ and qubit $j$, define $\widetilde{X}_{\ell,j} = C_{\ell-1}^\dagger X_j C_{\ell-1}$. Then the ansatz in* (2) *admits the exact commuting representation*

$$U^{\mathrm{B}}(\theta) = C_L \exp\left(-\frac{i}{2}\sum_{\ell=1}^{L}\sum_{j=1}^{n}\theta_{\ell,j}\widetilde{X}_{\ell,j}\right), \tag{12}$$

*where the generators $\widetilde{X}_{\ell,j}$ commute pairwise. Consequently, Theorem 4.1 provides a complete set of local smooth data-independent laws via a single matrix-nullspace computation with generators $\widetilde{X}_{\ell,j}$ and effective observable $C_L^\dagger M_0 C_L$.*

Following Theorem 4.2, we provide explicit examples that characterize the complete set of independent conservation laws for Basic PQCs. Further details are given in Appendix B.3.2. Here, we illustrate the result using a simple two-qubit, two-layer circuit.

**Example 4.3** (Two-qubit two-layer Basic PQC)**.** *Consider the Basic PQC with $n = 2$, $L = 2$, and readout $M_0 = Z_1 Z_2$. For this circuit, the nullspace $\mathcal{V}_{\mathrm{com}}$ is explicitly given by $\mathcal{V}_{\mathrm{com}} = \{\lambda \in \mathbb{R}^4 : \lambda_{1,1} = 0,\ \lambda_{1,2} + \lambda_{2,2} = 0\}$. Accordingly, the complete set of independent laws is $h_1(\theta) = \theta_{2,1}$ and $h_2(\theta) = \theta_{1,2} - \theta_{2,2}$. Therefore, during training dynamics, $\theta_{2,1}$ remains fixed, while $\theta_{1,2}$ and $\theta_{2,2}$ must change equally, preserving their difference. Hence, despite having 4 trainable parameters, gradient flow evolves along only 2 independent optimization directions.*

### 4.2 Conservation Laws for Strong PQCs

In this part, we turn to conservation laws for non-commuting trainable blocks. Because the effective Hermitian generators depend on the preceding layers and generally fail to commute, the kernel of the measurement differential is generally parameter dependent (see Proposition B.4). For concreteness, we focus on the Strong PQC, whose non-commuting trainable blocks prevent the problem from reducing to a fixed nullspace characterization. Nevertheless, explicit conservation laws can still be derived from stabilizers associated with the final layer.

**Theorem 4.4** (Last-layer stabilizer laws). *Let $\Phi_L^{\mathrm{S}} = (E_L^S)^\dagger M_0 E_L^S$ be the effective last-layer observable of the Strong PQC. For each qubit $j$, the following statements hold:*

*(i) If $[\Phi_L^{\mathrm{S}}, Z_j] = 0$, then $\omega_{L,j}$ is a data-independent conservation law.*

*(ii) If $\Phi_L^{\mathrm{S}}$ acts as the identity on the $j$-th qubit, then $\phi_{L,j}$, $\beta_{L,j}$, and $\omega_{L,j}$ are data-independent conservation laws.*

*In particular, suppose that $\Phi_L^{\mathrm{S}} = Z(\mathcal{J}) = \prod_{j\in\mathcal{J}} Z_j$ for a subset $\mathcal{J} \subseteq \{1,\dots,n\}$, where $Z(\varnothing) = I$. Then the set $\{\omega_{L,j}\}_{j\in\mathcal{J}} \cup \{\phi_{L,j}, \beta_{L,j}, \omega_{L,j}\}_{j\notin\mathcal{J}}$ forms $3n - 2|\mathcal{J}|$ independent conservation laws.*

Theorem 4.4 identifies conservation laws directly from the effective observable of the final layer. The proof is provided in Appendix B.3.3. These conservation laws hold for any depth $L$, although additional cross-layer conservation laws may also arise. We show this in a two-qubit, two-layer Strong PQC, where an additional cross-layer conservation law is present.

**Example 4.5** (Two-qubit, two-layer Strong PQC). *Consider the Strong PQC with $n = 2$, $L = 2$, and $M_0 = Z_1 Z_2$. Its complete set of independent conservation laws is $\{\omega_{2,1}, \phi_{2,2}, \beta_{2,2}, \omega_{2,2}, \phi_{2,1} - \omega_{1,2}\}$. On an open dense subset $\Omega \subset \Theta$, the measurement differential has rank $r_0 = 6$, whereas the Lie closure has rank $r_{\mathrm{Lie}} = 7$. Equivalently, the Lie closure removes one additional direction from the pointwise kernel of the measurement Jacobian. The cross-layer law $\phi_{2,1} - \omega_{1,2}$ appears because the measurement depends on $\phi_{2,1}$ and $\omega_{1,2}$ only through their sum. The proof is given in Appendix B.3.4.*

### 4.3 Conservation Laws and Rank Gaps in Two-Layer Clifford Circuits

Example 4.5 shows that non-commuting circuits may exhibit cross-layer conservation laws and that the rank of the measurement Jacobian alone is insufficient to certify completeness. We next generalize this analysis to a broader family of two-layer circuits with Clifford entanglers and construct examples in which the rank gap increases linearly with circuit width.

Consider a class of two layer circuits on $n \geq 3$ qubits with $p = 6n$ real valued angle parameters. Suppose the unitary circuit is defined by $U = E_2 R_2^S E_1 R_1^S$, where $R_\ell^S = \bigotimes_{j=1}^n R_Z(\omega_{\ell,j}) R_Y(\beta_{\ell,j}) R_Z(\phi_{\ell,j})$, $E_1$ is Clifford, $E_2$ is fixed, and $E_2^\dagger M_0 E_2 = Z_k$ for a readout qubit $k \in \{1,\dots,n\}$. Since conjugation by a Clifford unitary maps any Pauli string to another Pauli string up to an overall sign, we define $P_T = E_1^\dagger T_k E_1$ for $T \in \{X, Y, Z\}$. For a Pauli string $P = \bigotimes_{j=1}^n P_j$, its support is $\mathrm{supp}(P) := \{j : P_j \neq I\}$. We assume that $P_X, P_Y, P_Z$ have pairwise distinct supports. Ignoring overall signs, let $P_T^{(j)} \in \{I, X, Y, Z\}$ denote the local Pauli factor of $P_T$ acting on qubit $j$. We classify each qubit according to the local triple $(P_X^{(j)}, P_Y^{(j)}, P_Z^{(j)})$.

More precisely, $S_0 = \{j : P_X^{(j)} = P_Y^{(j)} = P_Z^{(j)} = I\}$ contains qubits on which all three strings act trivially. For $T \in \{X, Y, Z\}$, $S_T = \{j : P_A^{(j)} \in \{I, T\}$ for all $A \in \{X, Y, Z\}$, with $P_A^{(j)} = T$ for at least one $A\}$ contains qubits for which every nonidentity local factor is $T$. Finally, $S_3 = \{j : P_X^{(j)}, P_Y^{(j)}, P_Z^{(j)}$ are all nonidentity and pairwise distinct$\}$. The identity $P_Y = iP_X P_Z$ implies that these cases exhaust all possible local configurations. We write $n_T = |S_T|$ for $T \in \{0, X, Y, Z, 3\}$. These quantities will be used in the following theorem to explicitly characterize the complete set of independent conservation laws.

**Theorem 4.6** (Complete laws for two-layer Clifford Circuits). *Under the preceding assumptions, a complete set of independent conservation laws in two-layer Clifford circuits is*

$$\{\omega_{2,k}\} \cup \{\phi_{2,j}, \beta_{2,j}, \omega_{2,j}\}_{j\neq k} \cup \{\phi_{1,j}, \beta_{1,j}, \omega_{1,j}\}_{j\in S_0} \cup \{\omega_{1,j}\}_{j\in S_Z}. \tag{13}$$

*Furthermore, we explicitly show that on an open dense set $\Omega_C \subseteq \Theta$, we have $r_0 = 2 + 2(n_X + n_Y + n_Z) + 3n_3$ and $r_{\mathrm{Lie}} = 2 + 3(n_X + n_Y + n_3) + 2n_Z$. Consequently, there are exactly $q = 3n - 2 + 3n_0 + n_Z$ independent conserved functions and the Lie closure removes exactly $r_{\mathrm{Lie}} - r_0 = n_X + n_Y$ directions from the kernel of the measurement Jacobian.*

The distinct support condition reduces the measurement differential to a collection of single qubit calculations, so that the conservation laws and the two relevant ranks can be evaluated in $O(g(E_1) + n)$ time, where $g(E_1)$ denotes the number of bounded locality Clifford gates in $E_1$. By applying Theorem 4.6, we then derive explicit results for two-layer Clifford circuits, including a Strong PQC with a CNOT ring and a construction that achieves the maximal rank gap.

**Corollary 4.7** (Rank gaps in two-layer Clifford Circuits)**.** *Under the hypotheses of Theorem 4.6, the following statements hold:*

*(i) (CNOT ring). For $E_1 = E_2 = E^{\mathrm{B}}$ and $M_0 = Z_1 Z_n$, the collection $\omega_{2,1}$, $(\phi_{2,j}, \beta_{2,j}, \omega_{2,j})_{j=2}^{n}$ and $(\omega_{1,j})_{j=3}^{n}$ form a complete set of $4n - 4$ independent local laws. On $\Omega_{\mathrm{C}}$, $r_0 = 2n + 3$, and $r_{\mathrm{Lie}} = 2n + 4$, giving a rank gap of one.*

*(ii) (Rank gap upper bound). On $\Omega_{\mathrm{C}}$, the rank gap satisfies $r_{\mathrm{Lie}} - r_0 \leq n - 1$.*

*(iii) (Maximal rank gap). There exists a circuit whose entangler contains $n - 1$ CNOT gates and a single Hadamard gate, together with a two qubit readout, for which the $3n - 2$ frozen parameters in the final layer constitute a complete set of local conservation laws. On $\Omega_{\mathrm{C}}$, the measurement differential and Lie closure satisfy $r_0 = 2n + 3$ and $r_{\mathrm{Lie}} = 3n + 2$ which yields a rank gap of $n - 1$. Hence, this construction attains the maximal possible gap, showing that the difference between instantaneous invisible directions and independent conservation laws can scale linearly with the circuit width.*

### 4.4 Training Benefits with Conservation Laws

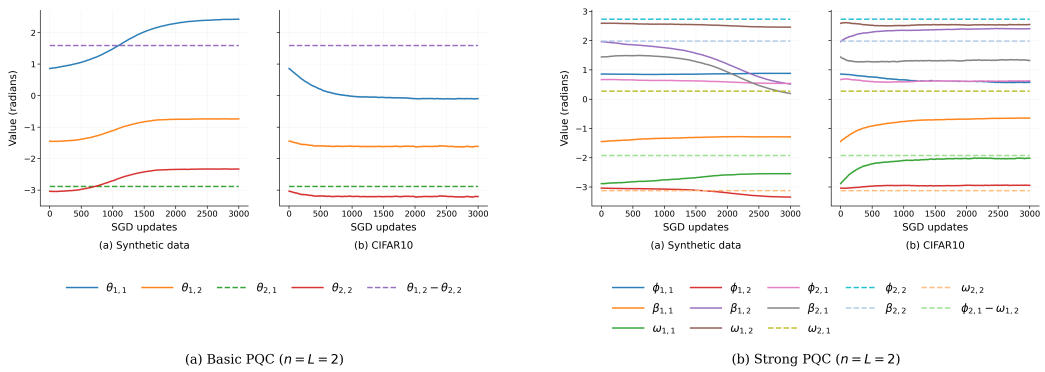
*Dimension reduction.* Similar to classical neural networks (Marcotte et al., 2023; 2026), conservation laws determine the number of effective optimization directions and thereby restrict training to a lower dimensional parameter space. Each linear law fixes one linear combination of parameters at its initial value, so $q$ independent global linear laws constrain a $p$ parameter circuit to an affine subspace of dimension $p - q$. Choosing orthonormal coordinates on this subspace preserves both the original Euclidean gradient flow and exact gradient SGD, without changing the learning rates or minibatches. For example, the five laws in Example 4.5 reduce the two qubit, two layer Strong PQC from twelve optimization variables to seven.

*Reduced gradient measurement cost.* When each parameter occurs once in a Pauli rotation, the parameter-shift rule (Schuld et al., 2019) estimates all derivatives of the measurement output $F_\rho$ using $2p$ shifted circuit settings per input. Linear laws reduce this to $2(p - q)$ settings by reconstructing the full gradient from $p - q$ selected derivatives. With exact derivatives, this reconstruction yields the same gradient and hence the same training updates. With finite-shot estimates, it enforces the conservation laws exactly, although the reconstruction may amplify sampling noise. For the Clifford circuits of Theorem 4.6, differential relations further reduce the count to $2r_0$ on $\Omega_{\mathrm{C}}$, with $O(n)$ classical cost per step once the Pauli classification is available. Writing $\Delta = r_{\mathrm{Lie}} - r_0$, and allocating $S_{\mathrm{d}}$ independent $\pm 1$ projective shots equally across these settings, the unbiased estimator $\widehat{\nabla F_\rho}$ satisfies $\mathbb{E}\|\widehat{\nabla F_\rho} - \nabla F_\rho\|^2 \leq r_0(r_0 + 2\Delta)/S_{\mathrm{d}}$. An independent prediction estimate requires one additional unshifted setting. Proofs and error analysis are given in Appendix B.3.7.

## 5 Numerical Evidence of Conservation Laws

In this section, we empirically examine whether the conservation laws derived under continuous-time gradient flow (7) remain approximately preserved under discrete-time optimization methods such as stochastic gradient descent (SGD). We have proved that if the conserved function is linear w.r.t to parameter $\theta$, its invariant is preserved exactly under discrete updates (See Appendix B.3.7 for the proof).

**Dataset.** We validate our theoretical results on both synthetic data and CIFAR 10 (Krizhevsky et al., 2009). We focus on quantum circuits with $n = 2$ qubits and $L = 2$ layers, while results for




(a) Basic PQC ($n = L = 2$) (b) Strong PQC ($n = L = 2$)

Figure 1: Evolution of the model parameters during training. Panels (a) and (b) correspond to the Basic and Strong PQCs, respectively, on synthetic data and CIFAR-10. The dashed lines indicate the theoretically predicted conservation laws, including both invariant individual parameters and conserved linear combinations of parameters. These quantities remain constant throughout SGD updates, consistently with our theoretical results.

larger and deeper circuits are provided in Appendix C. For the synthetic experiments, we sample the true parameter $\theta^\star$ from $\mathcal{U}([-\pi,\pi]^p)$ and generate $1024$ i.i.d. samples $(\rho_i, y_i)$. The observable $M_0$ is chosen as $\bigotimes_{i=1}^n Z$, where $Z$ denotes the Pauli $Z$ operator. Each input state is given by $\rho_i = z_i z_i^\dagger / \|z_i\|^2$, $z_i \sim \mathcal{CN}(0, I_{2^n})$, with labels $y_i = F_{\rho_i}(\theta^\star) + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, 0.1)$. For CIFAR 10, we consider binary classification of classes $0$ and $1$, each containing $5000$ images, and minimize the cross entropy loss. The quantum circuit output is rescaled to the interval $[0, 1]$ via the transformation $x \to (1 + x)/2$. The images are reduced using PCA and subsequently amplitude encoded following (Han et al., 2025).

**Training.** The training procedure is as follows. We first initialize the model parameters as $\theta \sim \mathcal{U}([-\pi,\pi]^p)$. We then minimize the loss function using stochastic gradient descent (SGD) with a batch size 32, without weight decay. The step size is fixed at $\tau = 0.01$ throughout training.

**Verification for basic and strong PQCs.** For the basic PQC, the conservation laws are given by $\theta_{2,1}$ and $\theta_{1,2} - \theta_{2,2}$. Likewise, for the strong PQC, the identified conservation laws are $\omega_{2,1}$, $\phi_{2,2}$, $\beta_{2,2}$, $\omega_{2,2}$, and $\phi_{2,1} - \omega_{1,2}$. As illustrated in Figure 1, these quantities remain conserved throughout discrete training. Moreover, we observe no additional parameter relations that yield independent conserved quantities, such as further linear combinations involving sums or differences of parameters. This empirical evidence supports that the identified laws constitute a maximal independent set and hence characterize all conservation laws.

## 6 Conclusion

This paper develops a theoretical framework for characterizing conservation laws in widely used parameterized quantum circuits, including the Basic Entangler and Strongly Entangling architectures, under gradient flow. For both architectures, we identify conservation laws associated with individual parameters, showing that several parameters remain frozen throughout training. We further show that, for a class of two layer Clifford circuits with at least three qubits, the gap between the Lie closure rank and the measurement Jacobian rank can grow linearly with circuit width. These results reveal structural redundancies that may lead to unnecessary optimization costs.

Several directions remain for future work. While we provide a complete characterization of conservation laws for the Basic Entangler circuit, completeness for Strongly Entangling circuits with more than two layers remains open. Our framework highlights that trainability depends jointly on the circuit architecture and the chosen observable, since the same unitary ansatz can exhibit different active dimensions and conservation laws under different readouts. This suggests designing circuits and measurements together to avoid frozen parameters, improve parameter efficiency, and reduce redundant computational cost.

### AI USE STATEMENT

In this work, we used generative AI tools for language editing, structural reorganization, and algebraic cross-checking. We have not used generative AI tools to create research content. Additionally, we used generative AI tools to check the coding experiments described in the paper. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This theoretical work uses the publicly available CIFAR-10 benchmark for numerical illustration and introduces neither new human-subject data nor claims about real-world deployment. The primary risk is that results derived under idealized, noiseless gradient assumptions could be misinterpreted as guarantees of real-world hardware performance. We therefore clearly distinguish exact structural results from proposed extensions relevant to hardware implementation.

### REPRODUCIBILITY STATEMENT

All circuit conventions, theorem assumptions, and training hyperparameters used in the reported experiments are specified in the main text, with complete proofs provided in the appendices. The full set of independent conservation laws for the two-layer Basic Entangler and Strongly Entangling layers is explicitly stated. Empirical observations are not used to establish completeness.

## REFERENCES

Amira Abbas, David Sutter, Christa Zoufal, Aurelien Lucchi, Alessio Figalli, and Stefan Woerner. The power of quantum neural networks. *Nature Computational Science*, 1:403–409, 2021.

Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA*, 11(1):307–353, 2022.

Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, Shahnawaz Ahmed, Vishnu Ajith, M Sohaib Alam, Guillermo Alonso-Linaje, Bharath AkashNarayanan, Ali Asadi, et al. Pennylane: Automatic differentiation of hybrid quantum-classical computations. *arXiv preprint arXiv:1811.04968*, 2018.

Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, 2021a.

Marco Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J Coles. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature communications*, 12(1): 1791, 2021b.

Iris Cong, Soonwon Choi, and Mikhail D Lukin. Quantum convolutional neural networks. *Nature Physics*, 15(12):1273–1278, 2019.

Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in neural information processing systems*, 31, 2018.

Yuxuan Du, Tao Huang, Shan You, Min-Hsiu Hsieh, and Dacheng Tao. Quantum circuit architecture search for variational quantum algorithms. *npj Quantum Information*, 8(1):62, 2022.

Lena Funcke, Tobias Hartung, Karl Jansen, Stefan Kühn, and Paolo Stornati. Dimensional expressivity analysis of parametric quantum circuits. *Quantum*, 5:422, 2021.

Jason Han, Nicholas S DiBrita, Younghyun Cho, Hengrui Luo, and Tirthak Patel. Enqode: Fast amplitude embedding for quantum machine learning using classical data. In *2025 62nd ACM/IEEE Design Automation Conference (DAC)*, pp. 1–7. IEEE, 2025.

Tobias Haug, Kishor Bharti, and MS Kim. Capacity and quantum geometry of parametrized quantum circuits. *PRX Quantum*, 2(4):040309, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Patrick Holzer and Ivica Turkalj. Spectral invariance and maximality properties of the frequency spectrum of quantum neural networks. *Quantum Machine Intelligence*, 8(1):33, 2026.

Tak Hur, Leeseok Kim, and Daniel K Park. Quantum convolutional neural network for classical data classification. *Quantum Machine Intelligence*, 4(1):3, 2022.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel L. K. Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. In *International Conference on Learning Representations*, 2021.

Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J. Coles, and M. Cerezo. Theory of overparametrization in quantum neural networks. *Nature Computational Science*, 2023.

Chen-Yu Liu, Chao-Han Huck Yang, Hsi-Sheng Goan, and Min-Hsiu Hsieh. A quantum circuit-based compression perspective for parameter-efficient learning. In *International Conference on Learning Representations*, volume 2025, pp. 49118–49142, 2025.

Junyu Liu, Francesco Tacchino, Jennifer R. Glick, Liang Jiang, and Antonio Mezzacapo. Representation learning via quantum neural tangent kernels. *PRX Quantum*, 2022.

Junyu Liu, Khadijeh Najafi, Kunal Sharma, Francesco Tacchino, Liang Jiang, and Antonio Mezzacapo. Analytic theory for the dynamics of wide quantum neural networks. *Physical Review Letters*, 2023.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Abide by the law and follow the flow: Conservation laws for gradient flows. *Advances in neural information processing systems*, 36:63210–63221, 2023.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Keep the momentum: Conservation laws beyond euclidean gradient flows. *arXiv preprint arXiv:2405.12888*, 2024.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Transformative or conservative? conservation laws for resnets and transformers. *arXiv preprint arXiv:2506.06194*, 2025.

Sibylle Marcotte, Gabriel Peyré, and Rémi Gribonval. Intrinsic training dynamics of deep neural networks. In *International Conference on Learning Representations*, volume 2026, pp. 14606–14646, 2026.

Yunseong Nam, Neil J. Ross, Yuan Su, Andrew M. Childs, and Dmitri Maslov. Automated optimization of large quantum circuits with continuous parameters. *npj Quantum Information*, 4:23, 2018.

Amit Peleg and Matthias Hein. Bias of stochastic gradient descent or the architecture: disentangling the effects of overparameterization of neural networks. *arXiv preprint arXiv:2407.03848*, 2024.

Yifeng Peng, Xinyi Li, Samuel Yen-Chi Chen, Kaining Zhang, Zhiding Liang, Ying Wang, and Yuxuan Du. Titan: A trajectory-informed technique for adaptive parameter freezing in large-scale vqe. *Advances in Neural Information Processing Systems*, 38:47294–47315, 2026.

Jun Qi, Chao-Han Huck Yang, Pin-Yu Chen, and Min-Hsiu Hsieh. Theoretical error performance analysis for variational quantum circuit based functional regression. *npj Quantum Information*, 9 (1):4, 2023.

Zachary Robertson and Oluwasanmi Koyejo. Implicit regularization in feedback alignment learning mechanisms for neural networks. *arXiv preprint arXiv:2306.01870*, 2023.

Lawrence K Saul. Weight-balancing fixes and flows for deep learning. *Transactions on Machine Learning Research*, 2023.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. Evaluating analytic gradients on quantum hardware. *Physical Review A*, 99:032331, 2019. doi: 10.1103/PhysRevA.99.032331.

Pierre Stock, Benjamin Graham, Rémi Gribonval, and Hervé Jégou. Equi-normalization of neural networks. *arXiv preprint arXiv:1902.10416*, 2019.

Viet-Hoang Tran, Vinh Khanh Bui, Tan Lai Ngoc, Nam Nguyen, Tuan Dam, and Tan M Nguyen. Conservation laws for modern neural architectures. *arXiv preprint arXiv:2606.17816*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Xuchen You, Shouvanik Chakrabarti, Boyang Chen, and Xiaodi Wu. Analyzing convergence in quantum neural networks: Deviations from neural tangent kernels. In *International Conference on Machine Learning*, pp. 40199–40224. PMLR, 2023.

Bo Zhao, Iordan Ganev, Robin Walters, Rose Yu, and Nima Dehmamy. Symmetries, flat minima, and the conserved quantities of gradient flow. In *International Conference on Learning Representations*, 2023.

# Appendix of "Conservation Laws in Parameterized Quantum Circuits: Complete Characterizations and Rank Separation"

## APPENDIX CONTENTS


**A Related Work** **14**

**B Theoretical Proofs** **14**

B.1 Notation and definitions . . . . . 14

B.2 Proofs for Section 3 . . . . . 16

B.2.1 Proof of Proposition 3.1 . . . . . 16

B.2.2 Proof of Proposition 3.2 . . . . . 16

B.2.3 Proof of Theorem 3.3 . . . . . 17

B.2.4 Detailed verification of equation (11) . . . . . 17

B.2.5 Proof of Theorem 3.6 . . . . . 17

B.3 Proofs for Section 4 . . . . . 18

B.3.1 Proof of Theorem 4.1 . . . . . 18

B.3.2 Proof of Theorem 4.2 and examples . . . . . 19

B.3.3 Proof of Theorem 4.4 and examples . . . . . 21

B.3.4 Example 4.5 . . . . . 24

B.3.5 Proof of Theorem 4.6 . . . . . 27

B.3.6 Proofs for the two-layer circuit families . . . . . 30

B.3.7 Dimension reduction and gradient measurement cost . . . . . 31

**C Broader Numerical Results** **38**

# A Related Work

**Conservation laws of gradient flows.** Our framework transfers the Lie-algebraic characterization of Marcotte et al. (2023) from classical networks to the measurement map of a PQC. The linear laws we obtain are translation symmetries of the measurement map, which links them to the symmetry-based analyses of gradient-flow dynamics in (Kunin et al., 2021; Zhao et al., 2023). The gap between directions that are invisible at a single parameter point and genuine conservation laws is already present in classical models: for two-layer linear networks, Marcotte et al. (2023) prove that the balancedness laws are complete and show that the differential kernel can have larger dimension than the span of gradients of independent laws. In the quantum setting, our Clifford circuit class admits a complete classification through Pauli propagation and local Lie-bracket computations, without forming any matrix of size $2^n$. Within this class, the gap between the measurement-Jacobian rank and the Lie-closure rank grows linearly with the number of qubits (Corollary 4.7).

**Circuit redundancy and light cones.** Several structural mechanisms are known to render circuit parameters irrelevant to the output: gates outside the backward light cone of the observable (Cerezo et al., 2021b), generators that commute with the evolved observable (Cerezo et al., 2021b), and consecutive rotations that can be merged across fixed gates, as exploited in circuit optimization (Nam et al., 2018). Theorem 3.6 separates these mechanisms into two first-order sources of invisibility: a parameter variation may leave the unitary unchanged, or it may change the unitary without changing the readout. Such mechanisms are useful for constructing laws, but they do not by themselves show that no further laws exist. Our completeness results (Theorems 4.2 and 4.6, and Example 4.5) close this gap by excluding any additional independent local law. For the noncommuting families, the Lie-closure analysis also explains why certain directions that are invisible at a point do not give rise to conservation laws.

**Expressivity, redundancy, and effective dimension.** For a fixed input state, (Funcke et al., 2021) uses the Jacobian of output state to count locally independent directions and describe how to remove directions that change only its global phase. Meanwhile, (Haug et al., 2021; Larocca et al., 2023) employ rank of quantum Fisher information matrix to measure the local dimension of the family of output states. Abbas et al. (2021) use classical Fisher information, which captures the sensitivity of model predictions to parameter changes, to define a sample-size-dependent measure of model capacity called the effective dimension. Our setting is complementary to those existing works: the observable is fixed and the input state is arbitrary, so we study the evolved observable and its stabilizer rather than the output state. This provides a Heisenberg-picture analogue of these state-based notions. On regions of constant rank, the measurement Jacobian gives the local dimension of the model, whereas the Lie closure gives the number of independent local data-independent laws. Our rank separation shows that these two quantities can differ: the invariant manifolds determined by the Lie closure may have larger dimension than the local image of the measurement map, although every training trajectory remains inside them (Corollary 4.7).

**Training dynamics of QNN** (Liu et al., 2022) uses quantum neural tangent kernel (QNTK) to analyze training in lazy training and estimate how training error departs from the frozen-QNTK prediction. Built on that framework, Liu et al. (2023) derive an average training-error convergence rate for wide random QNNs and give conditions under which kernel fluctuations and second-order effects are small. You et al. (2023) further show that QNN training can deviate from fixed-kernel dynamics even in an overparameterized limit. For pure state inputs, binary labels and Pauli readouts, they prove an at-most-sublinear convergence rate. Related empirical work, such as TITAN (Peng et al., 2026), uses training trajectories to predict parameters that can be frozen in variational quantum eigensolvers. Our analysis complements this approach by identifying conditions on the architecture and readout that guarantee exact parameter freezing and conserved parameter combinations.

# B Theoretical Proofs

## B.1 Notation and definitions

We collect the geometric, Pauli, and sampling notation used throughout this appendix. All linear spans are over $\mathbb{R}$, and symbols introduced for an individual calculation are local to that proof.

**Measurement gradients and Lie closure.** Let $\xi_\alpha = \nabla f_\alpha$, where the probe functions $f_\alpha$ are defined in (9), and let $\mathfrak{L}$ be the smallest real Lie algebra of smooth vector fields containing the $\xi_\alpha$. Its pointwise distribution is

$$\mathcal{L}_Q(\theta) = \operatorname{span}\{Y(\theta) : Y \in \mathfrak{L}\} = \operatorname{Lie}(\mathcal{W}_Q)(\theta),$$

the last expression being the notation of Section 3. We write $r_0(\theta) = \operatorname{rank} \mathrm{d}\Phi_\theta$ and $r_{\mathrm{Lie}}(\theta) = \dim \mathcal{L}_Q(\theta)$ for the measurement-Jacobian rank and the Lie-closure rank, respectively, and drop the argument $\theta$ whenever these ranks are constant on the domain under consideration. For smooth vector fields $X, Y$, we use the bracket convention $[X, Y](\theta) = DY(\theta)X(\theta) - DX(\theta)Y(\theta)$, where $DX(\theta)$ denotes the Jacobian of $X$.

For a real linear map $T : \mathbb{R}^p \to \mathrm{Herm}(d)$, the adjoint $T^* : \mathrm{Herm}(d) \to \mathbb{R}^p$ is characterized by $\langle Tv, A\rangle_{\mathrm{HS}} = v^\top T^* A$ for all $v \in \mathbb{R}^p$ and $A \in \mathrm{Herm}(d)$, where $\langle A, B\rangle_{\mathrm{HS}} = \operatorname{tr}(AB)$ for $A, B \in \mathrm{Herm}(d)$. As shown in the proof of Theorem 3.3, the measurement-gradient distribution satisfies

$$\mathcal{W}_Q(\theta) = \operatorname{im}(\mathrm{d}\Phi_\theta)^* = (\ker \mathrm{d}\Phi_\theta)^\perp. \tag{14}$$

**Pauli strings and support spaces.** The matrices $X, Y, Z$ act on a single qubit, and $I_2$ denotes the single-qubit identity. For a single-qubit matrix $A$, we write $A^{(j)} = I_2^{\otimes(j-1)} \otimes A \otimes I_2^{\otimes(n-j)}$ for its action on qubit $j$ of the full register; in particular, $X_j = X^{(j)}$, $Y_j = Y^{(j)}$, and $Z_j = Z^{(j)}$. A Pauli string is a tensor product of matrices from $\{I_2, X, Y, Z\}$, possibly multiplied by an overall sign, and its support is the set of positions carrying a nonidentity factor.

For $S \subseteq \{1, \dots, n\}$, let $\mathcal{V}_S^{\mathrm{supp}}$ be the real span of tensor products whose local factors are traceless Hermitian matrices at the positions in $S$ and $I_2$ elsewhere. In the Clifford support calculation we use the normalized inner product $\langle A, B\rangle = 2^{-n} \operatorname{tr}(AB)$ on $\mathrm{Herm}(2^n)$ together with its induced norm; elsewhere, the Hilbert–Schmidt inner product is the unnormalized one defined above.

**Regular domains for the two-layer circuits.** Throughout, $\theta$ collects the angle parameters $\phi_{\ell,j}$, $\beta_{\ell,j}$, and $\omega_{\ell,j}$, with $\ell$ indexing the layer and $j$ the qubit. For the two-qubit, two-layer Strong circuit, define

$$\Omega_{\mathrm{S}} = \{\theta \in \Theta : \sin\beta_{1,1} \sin\beta_{1,2} \sin\beta_{2,1} \cos\beta_{2,1} \neq 0\}. \tag{15}$$

For the Clifford class, write the three strings as $P_\alpha = \epsilon_\alpha \bigotimes_{j=1}^n p_{\alpha,j}$ with $\alpha \in \{X, Y, Z\}$, $\epsilon_\alpha \in \{\pm 1\}$, and $p_{\alpha,j} \in \{I_2, X, Y, Z\}$. The support assumption reads

$$\operatorname{supp}(P_X), \quad \operatorname{supp}(P_Y), \quad \operatorname{supp}(P_Z) \quad \text{are pairwise distinct.} \tag{16}$$

With the readout qubit $k$ of the main text, set $a_X = -\sin\beta_{2,k} \cos\phi_{2,k}$, $a_Y = \sin\beta_{2,k} \sin\phi_{2,k}$, and $a_Z = \cos\beta_{2,k}$. The regular set of the main text can then be written as

$$\Omega_{\mathrm{C}} = \{\theta \in \Theta : a_X a_Y a_Z \neq 0, \quad \sin\beta_{1,j} \neq 0 \text{ for every } j \notin S_0\}, \tag{17}$$

where $S_0$ is the set of qubits at which all three local factors $p_{X,j}, p_{Y,j}, p_{Z,j}$ equal the identity. Both $\Omega_{\mathrm{S}}$ and $\Omega_{\mathrm{C}}$ are open dense subsets of their respective parameter domains.

**Finite-shot gradient estimates.** All expectations in the finite-shot calculations are conditional on the current input $\rho$ and parameters $\theta$. We assume ideal independent projective measurements of a Hermitian observable $M_0$ with $M_0^2 = I$, so that each shot returns $\pm 1$; this covers Pauli readouts and the readouts of Theorem 4.6. Each differentiated parameter occurs in a single gate $\exp(-i\theta_j H_j/2)$ with $H_j^\dagger = H_j$ and $H_j^2 = I$, so the two-shift rule applies. In the presence of hardware noise, the reconstruction identities must be verified for the corresponding noisy measurement map.

In these calculations, write $d = \nabla F_\rho(\theta) \in \mathbb{R}^p$. Let $J \subseteq \{1, \dots, p\}$ be an ordered set of $m = |J|$ parameter indices, and suppose that a known matrix $T \in \mathbb{R}^{p \times m}$ provides the reconstruction $d = T d_J$, where $d_J$ is the subvector of $d$ indexed by $J$. Both $T$ and $J$ may depend on $\theta$ but remain fixed during sampling. For $j \in J$, let $e_j$ be the $j$-th standard basis vector of $\mathbb{R}^p$ and define

$$F_{j,\pm} = F_\rho(\theta \pm \tfrac{\pi}{2} e_j), \qquad v_{j,\pm} = 1 - F_{j,\pm}^2,$$

where $v_{j,\pm} \leq 1$ is the single-shot variance at the corresponding shifted setting. Let $\widehat{F}_{j,\pm}$ be mutually independent sample means formed from integer shot counts $N_{j,\pm} \geq 1$; collecting them in the order of $J$, define

$$\widehat{d}_J = \tfrac{1}{2}(\widehat{F}_{J,+} - \widehat{F}_{J,-}), \qquad \widehat{d} = T\widehat{d}_J.$$

The columns of $T$ are labeled by the ordered indices in $J$, and we set $w_j = \|T_{:,j}\|^2$. For vectors, $\|\cdot\|$ is the Euclidean norm; for real matrices, $\|T\|_{\mathrm{op}} = \sup_{\|v\|=1} \|Tv\|$ and $\|T\|_F = (\mathrm{tr}(T^\top T))^{1/2}$ denote the operator and Frobenius norms, respectively.

### B.2 Proofs for Section 3

We prove the gradient characterizations and the measurement-differential identities. Throughout, $\Theta \subseteq \mathbb{R}^p$ is open, $U$ is smooth, $\eta > 0$, and $l(\cdot, y)$ is $C^2$ for every $y \in \mathcal{Y}$; the inputs range over all density matrices, and the label set $\mathcal{Y}$ is nonempty. We write $\partial_1 l(s, y)$ for the derivative of the loss with respect to its first argument. We first characterize conservation laws through loss and measurement gradients. We then express the measurement differential in terms of circuit generators and observable stabilizers.

#### B.2.1 Proof of Proposition 3.1

The proof below applies the chain rule along gradient-flow trajectories. Single-sample datasets give the necessary condition, while averaging their gradients gives the converse.

*Proof.* Suppose first that $h \in C^1(\Theta, \mathbb{R})$ is a data-independent conservation law. Fix $\theta \in \Theta$, a density matrix $\rho$, and a label $y \in \mathcal{Y}$, and consider the one-sample dataset $\mathcal{D} = \{(\rho, y)\}$ with loss $L_{\mathcal{D}}(\vartheta) = l(F_\rho(\vartheta), y)$. Under the regularity assumptions, $\nabla L_{\mathcal{D}}$ is locally Lipschitz, so the Picard–Lindelöf theorem provides a unique local solution of (7) with initial condition $\theta(0) = \theta$ and initial velocity

$$\dot{\theta}(0) = -\eta\, \partial_1 l(F_\rho(\theta), y) \nabla F_\rho(\theta).$$

Since $h$ is constant along this trajectory, the chain rule gives

$$0 = \frac{\mathrm{d}}{\mathrm{d}t} h(\theta(t))\Big|_{t=0} = -\eta\, \partial_1 l(F_\rho(\theta), y) \nabla h(\theta)^\top \nabla F_\rho(\theta).$$

As $\eta > 0$ and $\rho, y$ were arbitrary, $\nabla h(\theta)$ is orthogonal to every one-sample loss gradient, and hence to their real span $\mathcal{W}_{\mathrm{loss}}(\theta)$.

Conversely, suppose that $\nabla h(\theta) \perp \mathcal{W}_{\mathrm{loss}}(\theta)$ for every $\theta \in \Theta$. For any finite dataset $\mathcal{D} = \{(\rho_i, y_i)\}_{i=1}^N$, the loss gradient satisfies

$$\nabla L_{\mathcal{D}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \partial_1 l(F_{\rho_i}(\theta), y_i) \nabla F_{\rho_i}(\theta) \in \mathcal{W}_{\mathrm{loss}}(\theta).$$

Along any solution of (7), we therefore have $\frac{\mathrm{d}}{\mathrm{d}t} h(\theta(t)) = -\eta\, \nabla h(\theta(t))^\top \nabla L_{\mathcal{D}}(\theta(t)) = 0$ for every time at which the trajectory remains in $\Theta$. Thus $h$ is constant along every training trajectory and for every dataset. The same argument applies to a local law as long as the trajectory remains in its domain. □

#### B.2.2 Proof of Proposition 3.2

*Proof.* Let $B_1, \ldots, B_{d^2}$ be the real basis of $\mathrm{Herm}(d)$ used in (9), so that $f_\alpha(\theta) = \mathrm{tr}(B_\alpha \Phi(\theta))$. Density matrices also span $\mathrm{Herm}(d)$ over $\mathbb{R}$, since by spectral decomposition every Hermitian matrix is a real linear combination of rank-one density matrices. Because both $B \mapsto \mathrm{tr}(B\Phi(\theta))$ and differentiation with respect to $\theta$ are real-linear, the two families generate the same gradient span:

$$\mathcal{W}_Q(\theta) = \mathrm{span}_\rho \{\nabla F_\rho(\theta)\} = \mathrm{span}_{\alpha=1,\ldots,d^2} \{\nabla f_\alpha(\theta)\}.$$

This proves the finite-probe representation.

We next justify its use for conservation laws under the squared loss considered in the main text. For $l(s, y) = (s - y)^2$,

$$\nabla_\theta l(F_\rho(\theta), y) = 2(F_\rho(\theta) - y) \nabla F_\rho(\theta),$$

so $\mathcal{W}_{\mathrm{loss}}(\theta) \subseteq \mathcal{W}_Q(\theta)$. For the reverse inclusion, fix $\theta$ and a label $y_0 \in \mathcal{Y}$. If $M_0$ is a scalar multiple of $I_d$, then $\Phi$ is constant and both spaces are zero. Otherwise $\Phi(\theta)$ is not scalar, and there is a density matrix $\sigma$ with $F_\sigma(\theta) \neq y_0$. Whenever $F_\rho(\theta) \neq y_0$, the preceding identity gives $\nabla F_\rho(\theta) \in \mathcal{W}_{\mathrm{loss}}(\theta)$.

If instead $F_\rho(\theta) = y_0$, consider the density matrix $\widetilde{\rho} = (\rho + \sigma)/2$; then $F_{\widetilde{\rho}}(\theta) \neq y_0$, and linearity in the input yields $\nabla F_\rho(\theta) = 2\nabla F_{\widetilde{\rho}}(\theta) - \nabla F_\sigma(\theta) \in \mathcal{W}_{\mathrm{loss}}(\theta)$. Hence $\mathcal{W}_{\mathrm{loss}}(\theta) = \mathcal{W}_Q(\theta)$, also at points where individual residuals vanish. By Proposition 3.1, a function $h \in C^1(\Theta, \mathbb{R})$ is therefore a data-independent conservation law if and only if $\xi_\alpha h = 0$ for every $\alpha$, where $\xi_\alpha = \nabla f_\alpha$. For other losses, this characterization requires the same equality between the loss-gradient and measurement-gradient spans. □

### B.2.3 Proof of Theorem 3.3

*Proof.* For $B \in \mathrm{Herm}(d)$, define $f_B(\theta) = \mathrm{tr}(B\Phi(\theta))$. By the chain rule and the definition of the adjoint, for every $v \in \mathbb{R}^p$,

$$\mathrm{d}(f_B)_\theta(v) = \langle B, \mathrm{d}\Phi_\theta(v)\rangle_{\mathrm{HS}} = v^\top (\mathrm{d}\Phi_\theta)^* B,$$

so that $\nabla f_B(\theta) = (\mathrm{d}\Phi_\theta)^* B$. Letting $B$ range over a basis of $\mathrm{Herm}(d)$ and applying Proposition 3.2, we obtain

$$\mathcal{W}_Q(\theta) = \mathrm{span}_{\alpha=1,\dots,d^2}\{\nabla f_\alpha(\theta)\} = \mathrm{im}(\mathrm{d}\Phi_\theta)^*.$$

For the squared loss, or for any loss satisfying $\mathcal{W}_{\mathrm{loss}} = \mathcal{W}_Q$, Proposition 3.1 thus shows that $h$ is a data-independent conservation law if and only if $\nabla h(\theta) \perp \mathrm{im}(\mathrm{d}\Phi_\theta)^*$ for every $\theta \in \Theta$.

For a real linear map $T : \mathbb{R}^p \to \mathrm{Herm}(d)$, the identity $(\mathrm{im}\, T^*)^\perp = \ker T$ follows directly from the adjoint relation: a vector $v$ is orthogonal to $\mathrm{im}\, T^*$ exactly when $v^\top T^* B = \langle Tv, B\rangle_{\mathrm{HS}} = 0$ for every $B \in \mathrm{Herm}(d)$, that is, exactly when $Tv = 0$. Applied to $T = \mathrm{d}\Phi_\theta$, this gives

$$\nabla h(\theta) \perp \mathrm{im}(\mathrm{d}\Phi_\theta)^* \quad \Longleftrightarrow \quad \mathrm{d}\Phi_\theta\big(\nabla h(\theta)\big) = 0.$$

Writing $\partial_a = \partial/\partial\theta_a$, this condition takes the coordinate form

$$\mathrm{d}\Phi_\theta\big(\nabla h(\theta)\big) = \sum_{a=1}^{p} \partial_a h(\theta)\, \partial_a \Phi(\theta) = 0,$$

which proves the claimed equivalence. □

### B.2.4 Detailed verification of equation (11)

*Proof.* For $a \in \{1, \dots, p\}$, recall that $A_a(\theta) = U(\theta)^\dagger \partial_a U(\theta)$ and $K_a(\theta) = iA_a(\theta)$. Differentiating the unitarity identity $U(\theta)^\dagger U(\theta) = I_d$ gives

$$(\partial_a U(\theta))^\dagger U(\theta) + U(\theta)^\dagger \partial_a U(\theta) = 0,$$

so $A_a(\theta)^\dagger = -A_a(\theta)$; thus $A_a(\theta)$ is skew-Hermitian and $K_a(\theta)$ is Hermitian. The definition of $A_a(\theta)$ also yields $\partial_a U(\theta) = U(\theta)A_a(\theta)$ and $(\partial_a U(\theta))^\dagger = -A_a(\theta)U(\theta)^\dagger$. Since $M_0$ is fixed, differentiating $\Phi(\theta) = U(\theta)^\dagger M_0 U(\theta)$ gives

$$\begin{aligned}\partial_a \Phi(\theta) = (\partial_a U(\theta))^\dagger M_0 U(\theta) + U(\theta)^\dagger M_0 \partial_a U(\theta) &= -A_a(\theta)\Phi(\theta) + \Phi(\theta)A_a(\theta) \\ &= [\Phi(\theta), A_a(\theta)] = i[K_a(\theta), \Phi(\theta)],\end{aligned}$$

where the last equality uses $A_a(\theta) = -iK_a(\theta)$. □

### B.2.5 Proof of Theorem 3.6

*Proof.* Fix $\theta \in \Theta$ and recall that the real linear map $\kappa_\theta : \mathbb{R}^p \to \mathfrak{u}(d)$ is given by

$$\kappa_\theta(v) = U(\theta)^\dagger \mathrm{d}U_\theta(v) = \sum_{a=1}^{p} v_a A_a(\theta), \qquad v = (v_1, \dots, v_p)^\top \in \mathbb{R}^p.$$

(i) By linearity of the differential and (11),

$$\mathrm{d}\Phi_\theta(v) = \sum_{a=1}^{p} v_a [\Phi(\theta), A_a(\theta)] = [\Phi(\theta), \kappa_\theta(v)].$$

(ii) The stabilizer Lie algebra of $\Phi(\theta)$ under unitary conjugation is $\mathfrak{z}_{\Phi(\theta)} = \{A \in \mathfrak{u}(d) : [\Phi(\theta), A] = 0\}$. By part (i), $\mathrm{d}\Phi_\theta(v) = 0$ exactly when $\kappa_\theta(v) \in \mathfrak{z}_{\Phi(\theta)}$, and therefore

$$\ker \mathrm{d}\Phi_\theta = \kappa_\theta^{-1}\big(\mathfrak{z}_{\Phi(\theta)}\big) = \left\{ v \in \mathbb{R}^p : \sum_{a=1}^{p} v_a A_a(\theta) \in \mathfrak{z}_{\Phi(\theta)} \right\},$$

where $\kappa_\theta^{-1}$ denotes the preimage under $\kappa_\theta$.

(iii) Define the real linear commutator map $\mathrm{ad}_{\Phi(\theta)} : \mathfrak{u}(d) \to \mathrm{Herm}(d)$ by $\mathrm{ad}_{\Phi(\theta)}(A) = [\Phi(\theta), A]$; its values are Hermitian because $\Phi(\theta)$ is Hermitian and $A$ is skew-Hermitian. By part (i), $\mathrm{d}\Phi_\theta = \mathrm{ad}_{\Phi(\theta)} \circ \kappa_\theta$. The restriction of the commutator map to $\operatorname{im} \kappa_\theta$ has kernel

$$\ker\left(\mathrm{ad}_{\Phi(\theta)}\big|_{\operatorname{im}\kappa_\theta}\right) = \operatorname{im}\kappa_\theta \cap \mathfrak{z}_{\Phi(\theta)}.$$

Since $\kappa_\theta$ maps onto its image, applying the rank–nullity theorem to this restriction gives

$$\begin{aligned}\operatorname{rank} \mathrm{d}\Phi_\theta &= \operatorname{rank}\left(\mathrm{ad}_{\Phi(\theta)}\big|_{\operatorname{im}\kappa_\theta}\right) \\ &= \dim \operatorname{im}\kappa_\theta - \dim\big(\operatorname{im}\kappa_\theta \cap \mathfrak{z}_{\Phi(\theta)}\big) \\ &= \operatorname{rank}\kappa_\theta - \dim\big(\operatorname{im}\kappa_\theta \cap \mathfrak{z}_{\Phi(\theta)}\big).\end{aligned}$$

□

### B.3 Proofs for Section 4

We first show the laws for commuting blocks and apply them to Basic circuits. We then study last-layer stabilizers, prove the two-layer classifications, and present the resulting training benefits.

#### B.3.1 Proof of Theorem 4.1

We first express the span of the measurement gradients in terms of the measurement differential. We then show that, for commuting generators, the kernel of this differential does not depend on $\theta$, and we derive the complete family of laws from this fact.

**Proposition B.1** (Visible measurement rank)**.** *Let $B_1, \dots, B_{d^2}$ be a Hilbert–Schmidt orthonormal basis of $\mathrm{Herm}(d)$, and define $f_\alpha(\theta) = \operatorname{tr}(B_\alpha \Phi(\theta))$ and $\xi_\alpha = \nabla f_\alpha$. Then, for every $\theta \in \Theta$, we have*

$$\mathcal{W}_Q(\theta) = \operatorname{span}_{\alpha=1,\dots,d^2}\{\xi_\alpha(\theta)\} = \operatorname{im}\big((\mathrm{d}\Phi_\theta)^*\big), \qquad \dim \mathcal{W}_Q(\theta) = \operatorname{rank}\mathrm{d}\Phi_\theta.$$

*Proof.* For $B \in \mathrm{Herm}(d)$, let $f_B(\theta) = \operatorname{tr}(B\Phi(\theta))$. The proof of Theorem 3.3 gives $\nabla f_B(\theta) = (\mathrm{d}\Phi_\theta)^* B$. Since $B_1, \dots, B_{d^2}$ span $\mathrm{Herm}(d)$, their probe gradients span the image of $(\mathrm{d}\Phi_\theta)^*$. This proves the first identity. The second follows because a linear map and its adjoint have the same rank. □

*Proof of Theorem 4.1.* Recall that $E_L, E_R$ are fixed unitaries, the Hermitian generators $H_1, \dots, H_p$ commute pairwise, and $\Phi_L = E_L^\dagger M_0 E_L$. Define

$$R(\theta) = \exp\left(-\frac{i}{2}\sum_{a=1}^{p}\theta_a H_a\right), \qquad H_\lambda = \sum_{a=1}^{p}\lambda_a H_a \quad \text{for } \lambda \in \mathbb{R}^p.$$

Then $U(\theta) = E_L R(\theta) E_R$ and $\Phi(\theta) = E_R^\dagger R(\theta)^\dagger \Phi_L R(\theta) E_R$. Since $H_\lambda$ commutes with $R(\theta)$, differentiation in direction $\lambda$ gives

$$\mathrm{d}\Phi_\theta(\lambda) = \frac{i}{2} E_R^\dagger [H_\lambda, R(\theta)^\dagger \Phi_L R(\theta)] E_R = \frac{i}{2} E_R^\dagger R(\theta)^\dagger [H_\lambda, \Phi_L] R(\theta) E_R.$$

Since all matrices outside the commutator are invertible, we have

$$\mathrm{d}\Phi_\theta(\lambda) = 0 \quad \Longleftrightarrow \quad [H_\lambda, \Phi_L] = 0 \quad \Longleftrightarrow \quad \lambda \in \mathcal{V}_{\mathrm{com}},$$

where $\mathcal{V}_{\mathrm{com}} = \{\lambda \in \mathbb{R}^p : [H_\lambda, \Phi_L] = 0\}$ is the coefficient space defined in the theorem. Thus $\ker \mathrm{d}\Phi_\theta = \mathcal{V}_{\mathrm{com}}$ for every $\theta \in \Theta$. Set $q = \dim \mathcal{V}_{\mathrm{com}}$, and let $N \in \mathbb{R}^{p\times q}$ have columns forming a basis of $\mathcal{V}_{\mathrm{com}}$. Rank-nullity gives

$$r_0(\theta) = \operatorname{rank} \mathrm{d}\Phi_\theta = p - q.$$

By Proposition B.1 and the identity $\operatorname{im} T^* = (\ker T)^\perp$ for a linear map between finite-dimensional inner-product spaces, one has

$$\mathcal{W}_Q(\theta) = (\ker \mathrm{d}\Phi_\theta)^\perp = \mathcal{V}_{\mathrm{com}}^\perp.$$

Every measurement-gradient field therefore takes values in the same fixed subspace $\mathcal{V}_{\mathrm{com}}^\perp$. In addition, if smooth vector fields $\xi, \zeta$ take values in this subspace, their derivatives $D\xi(\theta)u$ and $D\zeta(\theta)u$ also belong to it for every $u \in \mathbb{R}^p$, where $D\xi$ and $D\zeta$ denote their Jacobian matrices. Consequently,

$$[\xi, \zeta](\theta) = D\zeta(\theta)\xi(\theta) - D\xi(\theta)\zeta(\theta) \in \mathcal{V}_{\mathrm{com}}^\perp.$$

Thus the pointwise Lie-closure distribution satisfies $\mathcal{L}_Q(\theta) \subseteq \mathcal{V}_{\mathrm{com}}^\perp$. Conversely, it contains $\mathcal{W}_Q(\theta) = \mathcal{V}_{\mathrm{com}}^\perp$. Therefore, we have

$$\mathcal{L}_Q(\theta) = \mathcal{V}_{\mathrm{com}}^\perp, \qquad r_{\mathrm{Lie}}(\theta) = \dim \mathcal{L}_Q(\theta) = p - q.$$

Define the linear law map $\mathbf{h}_{\mathrm{com}}(\theta) = N^\top \theta$. If $n_j$ is the $j$-th column of $N$, its corresponding component $h_j(\theta) = n_j^\top \theta$ satisfies $\nabla h_j(\theta) = n_j \in \mathcal{V}_{\mathrm{com}} = \ker \mathrm{d}\Phi_\theta$ for every $\theta \in \Theta$. From Theorem 3.3, each component is a data-independent conservation law. In addition, their gradients are independent because the columns of $N$ form a basis. Moreover, one has

$$\ker \mathrm{d}(\mathbf{h}_{\mathrm{com}})_\theta = \ker N^\top = (\operatorname{im} N)^\perp = \mathcal{V}_{\mathrm{com}}^\perp = \mathcal{L}_Q(\theta).$$

By Theorem 3.5, we have $q$ is the local number of independent conservation laws. To establish the functional characterization directly, let $g$ be a local smooth law. Its differential annihilates $\mathcal{W}_Q(\theta) = \ker N^\top$. Complete $u = N^\top \theta$ to a system of linear coordinates $(u, w)$, where $w \in \mathbb{R}^{p-q}$ consists of complementary coordinates. On a sufficiently small product neighborhood, the annihilation condition gives $\partial_w g = 0$. Hence

$$g(\theta) = \psi\big(N^\top \theta\big)$$

for a smooth function $\psi$. Conversely, every smooth function of $N^\top \theta$ has gradient in $\mathcal{V}_{\mathrm{com}}$ and is therefore a law. This proves local completeness.

Finally, recall the real matrix $\Gamma_{\mathrm{com}} \in \mathbb{R}^{d^2\times p}$ defined by $(\Gamma_{\mathrm{com}})_{\alpha a} = \operatorname{tr}(B_\alpha i[H_a, \Phi_L])$. Since $i[H_\lambda, \Phi_L]$ is Hermitian, its expansion in the orthonormal basis is

$$i[H_\lambda, \Phi_L] = \sum_{\alpha=1}^{d^2} \left( \sum_{a=1}^{p} (\Gamma_{\mathrm{com}})_{\alpha a} \lambda_a \right) B_\alpha.$$

The commutator vanishes exactly when every basis coefficient vanishes, which is equivalent to $\Gamma_{\mathrm{com}} \lambda = 0$. Hence $\mathcal{V}_{\mathrm{com}} = \ker \Gamma_{\mathrm{com}}$, completing the proof. □

### B.3.2 Proof of Theorem 4.2 and examples

Conjugation by CNOT gates maps every Pauli string built from identity and Pauli-$X$ factors to a string of the same form. This property allows us to rewrite the Basic circuit as a single commuting block and to apply Theorem 4.1.

*Proof of Theorem 4.2.* For $\nu \in \mathbb{F}_2^n$, where $\mathbb{F}_2$ is the field with two elements, define $X(\nu) = \bigotimes_{j=1}^n X^{\nu_j}$, with $X^0 = I_2$ and $X^1 = X$. Let $\mathfrak{a}_X = \operatorname{span}\{X(\nu) : \nu \in \mathbb{F}_2^n\} \subseteq \mathrm{Herm}(2^n)$, where the span is over $\mathbb{R}$. Every two elements of $\mathfrak{a}_X$ commute because its generating Pauli strings contain only $I_2$ and $X$ factors.

For a CNOT gate with control qubit $j$ and target qubit $k$, the conjugation rules are

$$\mathrm{CNOT}_{j\to k}^\dagger \, X_j \, \mathrm{CNOT}_{j\to k} = X_j X_k, \qquad \mathrm{CNOT}_{j\to k}^\dagger \, X_k \, \mathrm{CNOT}_{j\to k} = X_k,$$

while $X_b$ is unchanged for $b \notin \{j,k\}$. Thus CNOT conjugation maps every $X$-type Pauli string to another $X$-type Pauli string. The fixed Basic entangler $E^{\mathrm{B}}$ is a product of CNOT gates, so $(E^{\mathrm{B}})^\dagger \mathfrak{a}_X E^{\mathrm{B}} = \mathfrak{a}_X$. Moreover, $X(\nu)X(\mu) = X(\nu+\mu)$, where addition is over $\mathbb{F}_2$. Consequently, conjugation induces an invertible $\mathbb{F}_2$-linear map $T_{\mathrm{B}} : \mathbb{F}_2^n \to \mathbb{F}_2^n$ satisfying $(E^{\mathrm{B}})^\dagger X(\nu) E^{\mathrm{B}} = X(T_{\mathrm{B}}\nu)$.

Define the cumulative entanglers and dressed generators by

$$C_0 = I_{2^n}, \qquad C_\ell = (E^{\mathrm{B}})^\ell, \qquad \widetilde{X}_{\ell,j} = C_{\ell-1}^\dagger X_j C_{\ell-1}.$$

Recall that the trainable layer is $R_\ell^{\mathrm{B}} = \exp(-\frac{i}{2}\sum_{j=1}^n \theta_{\ell,j} X_j)$. Therefore, grouping the fixed entanglers on the left gives

$$U^{\mathrm{B}}(\theta) = C_L\big(C_{L-1}^\dagger R_L^{\mathrm{B}} C_{L-1}\big)\cdots\big(C_0^\dagger R_1^{\mathrm{B}} C_0\big).$$

For each layer, one has

$$C_{\ell-1}^\dagger R_\ell^{\mathrm{B}} C_{\ell-1} = \exp\left(-\frac{i}{2}\sum_{j=1}^n \theta_{\ell,j}\widetilde{X}_{\ell,j}\right).$$

All dressed generators lie in the abelian space $\mathfrak{a}_X$, so the exponents of the dressed layers commute pairwise and their product is the exponential of their sum:

$$U^{\mathrm{B}}(\theta) = C_L \exp\left(-\frac{i}{2}\sum_{\ell=1}^L\sum_{j=1}^n \theta_{\ell,j}\widetilde{X}_{\ell,j}\right).$$

This is exactly (12).

In the notation of Theorem 4.1, take $p = Ln$, $E_L = C_L$, $E_R = I_{2^n}$, and $H_{(\ell,j)} = \widetilde{X}_{\ell,j}$, where $(\ell,j)$ indexes the parameter $\theta_{\ell,j}$. The effective observable is

$$\Phi_L^{\mathrm{B}} = C_L^\dagger M_0 C_L.$$

The corresponding coefficient space is

$$\mathcal{V}_{\mathrm{B}} = \left\{\lambda \in \mathbb{R}^{Ln} : \left[\sum_{\ell=1}^L\sum_{j=1}^n \lambda_{\ell,j}\widetilde{X}_{\ell,j}, \Phi_L^{\mathrm{B}}\right] = 0\right\}. \tag{18}$$

Set $q^{\mathrm{B}} = \dim \mathcal{V}_{\mathrm{B}}$, and let $N^{\mathrm{B}} \in \mathbb{R}^{Ln\times q^{\mathrm{B}}}$ have columns forming a basis of $\mathcal{V}_{\mathrm{B}}$. For the Basic measurement map $\Phi^{\mathrm{B}}(\theta) = (U^{\mathrm{B}}(\theta))^\dagger M_0 U^{\mathrm{B}}(\theta)$, the commuting-block theorem gives

$$\ker \mathrm{d}(\Phi^{\mathrm{B}})_\theta = \mathcal{V}_{\mathrm{B}}, \qquad r_0(\theta) = r_{\mathrm{Lie}}(\theta) = Ln - q^{\mathrm{B}}.$$

Here $r_0$ and $r_{\mathrm{Lie}}$ denote, respectively, the rank of the measurement differential and the rank of the Lie closure of this circuit. Consequently, the components of $(N^{\mathrm{B}})^\top\theta$ form a complete family of independent linear laws: near every parameter point, a smooth function is a law if and only if it is a smooth function of these components.

Finally, we express $\mathcal{V}_{\mathrm{B}}$ as a matrix nullspace. Let $d = 2^n$ and fix a Hilbert–Schmidt orthonormal basis $B_1, \ldots, B_{d^2}$ of $\mathrm{Herm}(d)$. Define $\Gamma_{\mathrm{B}} \in \mathbb{R}^{d^2\times Ln}$ by

$$(\Gamma_{\mathrm{B}})_{\alpha,(\ell,j)} = \mathrm{tr}\Big(B_\alpha\, i\big[\widetilde{X}_{\ell,j}, \Phi_L^{\mathrm{B}}\big]\Big).$$

By the nullspace characterization in Theorem 4.1, $\mathcal{V}_{\mathrm{B}} = \ker\Gamma_{\mathrm{B}}$. Hence a single matrix-nullspace computation yields the complete family of laws. □

We compute the dressed generators and their commutators with the effective observable. The resulting linear conditions yield the two stated laws.

*Verification of Example 4.3.* Take $n = L = 2$, $M_0 = Z_1Z_2$, and the two-qubit Basic convention $E^{\mathrm{B}} = \mathrm{CNOT}_{1\to2}$. Then $C_2 = I_4$ and $\Phi_2^{\mathrm{B}} = Z_1Z_2$. In the coordinate order $(1,1),(1,2),(2,1),(2,2)$, the dressed generators are $X_1, X_2, X_1X_2, X_2$. The operator $X_1X_2$ commutes with $Z_1Z_2$, whereas $i[X_1, Z_1Z_2] = 2Y_1Z_2$, $i[X_2, Z_1Z_2] = 2Z_1Y_2$. These two commutators are linearly independent, so the kernel conditions are $\lambda_{1,1} = 0$ and $\lambda_{1,2} + \lambda_{2,2} = 0$. Therefore, $\theta_{2,1}$ and $\theta_{1,2} - \theta_{2,2}$ form the stated complete family of independent linear laws. □

**Example B.2** (Three-qubit, three-layer Basic PQC)**.** *Consider $n = L = 3$, the entangler $E^{\mathrm{B}} = \mathrm{CNOT}_{3\to1}\,\mathrm{CNOT}_{2\to3}\,\mathrm{CNOT}_{1\to2}$, and the readout $M_0 = Z_1Z_2Z_3$. With $C_3 = (E^{\mathrm{B}})^3$, direct Pauli conjugation gives*

$$\Phi_3^{\mathrm{B}} = C_3^\dagger M_0 C_3 = Z_1Z_3.$$

*The dressed generators are*

$$(\widetilde{X}_{1,j})_{j=1}^3 = (X_1, X_2, X_3),$$
$$(\widetilde{X}_{2,j})_{j=1}^3 = (X_1X_2, X_2X_3, X_1X_2X_3),$$
$$(\widetilde{X}_{3,j})_{j=1}^3 = (X_1X_3, X_1, X_2).$$

*Among these generators, $X_1, X_3, X_1X_2, X_2X_3$ anticommute with $\Phi_3^{\mathrm{B}}$, and their commutators are nonzero multiples of distinct Pauli strings. Hence these four commutators are linearly independent. All remaining displayed generators commute with $\Phi_3^{\mathrm{B}}$. Since $X_1$ occurs at both $(1,1)$ and $(3,2)$, the coefficient space is*

$$\mathcal{V}_{\mathrm{B}} = \left\{\lambda \in \mathbb{R}^9 : \lambda_{1,1} + \lambda_{3,2} = 0, \quad \lambda_{1,3} = \lambda_{2,1} = \lambda_{2,2} = 0\right\}.$$

*A complete family of five independent linear laws is therefore*

$$(h_1(\theta), \dots, h_5(\theta)) = (\theta_{1,2}, \theta_{2,3}, \theta_{3,1}, \theta_{3,3}, \theta_{1,1} - \theta_{3,2}).$$

*Thus $\theta_{1,2}, \theta_{2,3}, \theta_{3,1}, \theta_{3,3}$ remain fixed during training, while $\theta_{1,1}$ and $\theta_{3,2}$ have equal updates and preserve their difference. Both ranks equal four, and gradient flow remains in the four-dimensional affine subspace determined by the initial law values. Figure 2 illustrates the training dynamics of this circuit.*

### B.3.3 Proof of Theorem 4.4 and examples

We first establish a stabilizer criterion and compute the effective generators of the strong circuit. We then identify the last-layer laws and, when the effective observable is a Pauli-$Z$ string, prove that these laws are complete for a single layer.

**Proposition B.3** (Stabilizer-redundant circuit blocks)**.** *Suppose the circuit factors as $U(\theta) = \Sigma(\chi)\check{U}(\vartheta)$, where $\theta = (\chi, \vartheta)$ partitions the Euclidean parameter coordinates and $\Sigma, \check{U}$ are unitary blocks. If $\Sigma(\chi)^\dagger M_0 \Sigma(\chi) = M_0$ for every admissible $\chi$, then $\Phi(\theta)$ is independent of $\chi$, and each coordinate function $h_j(\theta) = \chi_j$ is a data-independent conservation law.*

*Proof of Proposition B.3.* Under the assumed factorization, the measurement map becomes

$$\Phi(\theta) = \check{U}(\vartheta)^\dagger \Sigma(\chi)^\dagger M_0 \Sigma(\chi) \check{U}(\vartheta) = \check{U}(\vartheta)^\dagger M_0 \check{U}(\vartheta).$$

Hence $\Phi(\theta)$ is independent of $\chi$, and $\partial_{\chi_j}\Phi(\theta) = 0$ for every coordinate $\chi_j$. As $\nabla h_j$ is the Euclidean coordinate vector of $\chi_j$, it follows that

$$\mathrm{d}\Phi_\theta\big(\nabla h_j(\theta)\big) = \partial_{\chi_j}\Phi(\theta) = 0.$$

By Theorem 3.3, each $h_j$ is therefore a data-independent conservation law. □

**Proposition B.4** (Strong-PQC measurement differential)**.** *Consider the Strong PQC in* (4)*, with $L$ layers and $n$ qubits. Write its local Euler blocks and trainable layers as*

$$R_{\ell,j} = R_Z(\omega_{\ell,j}) R_Y(\beta_{\ell,j}) R_Z(\phi_{\ell,j}), \qquad R_\ell^{\mathrm{S}} = \bigotimes_{j=1}^n R_{\ell,j}.$$

*Let $E_\ell^{\mathrm{S}}$ denote the fixed entangler in layer $\ell$, and define the circuit before that layer by*

$$U_{<1} = I_{2^n}, \qquad U_{<\ell} = E_{\ell-1}^{\mathrm{S}} R_{\ell-1}^{\mathrm{S}} \cdots E_1^{\mathrm{S}} R_1^{\mathrm{S}} \quad (\ell \geq 2).$$

*For an angle coordinate $a$, define $K_a(\theta) = i(U^{\mathrm{S}}(\theta))^\dagger \partial_a U^{\mathrm{S}}(\theta)$. Then*

$$K_{\phi_{\ell,j}}(\theta) = \frac{1}{2} U_{<\ell}(\theta)^\dagger Z_j U_{<\ell}(\theta), \tag{19}$$

$$K_{\beta_{\ell,j}}(\theta) = \frac{1}{2} U_{<\ell}(\theta)^\dagger \big(\sin\phi_{\ell,j}\, X_j + \cos\phi_{\ell,j}\, Y_j\big) U_{<\ell}(\theta), \tag{20}$$

$$K_{\omega_{\ell,j}}(\theta) = \frac{1}{2} U_{<\ell}(\theta)^\dagger \Big(-\sin\beta_{\ell,j}\cos\phi_{\ell,j}\, X_j + \sin\beta_{\ell,j}\sin\phi_{\ell,j}\, Y_j + \cos\beta_{\ell,j}\, Z_j\Big) U_{<\ell}(\theta). \tag{21}$$

*Proof of Proposition B.4.* Fix a parameter $a$ in layer $\ell$ and factor the circuit as $U^{\mathrm{S}} = W_{>\ell} E_\ell^{\mathrm{S}} R_\ell^{\mathrm{S}} U_{<\ell}$, where $W_{>\ell}$ denotes the product of all layers after layer $\ell$ (the identity when $\ell = L$). Since $W_{>\ell}$ and $E_\ell^{\mathrm{S}}$ do not depend on $a$ and are unitary, their contributions cancel, and

$$i(U^{\mathrm{S}})^\dagger \partial_a U^{\mathrm{S}} = U_{<\ell}^\dagger \big( i(R_\ell^{\mathrm{S}})^\dagger \partial_a R_\ell^{\mathrm{S}} \big) U_{<\ell}.$$

Suppose $a$ belongs to the local block $R_{\ell,j}$. The blocks acting on the remaining qubits cancel in the same way. For a $2 \times 2$ matrix $B$, write $B^{(j)} = I_2^{\otimes(j-1)} \otimes B \otimes I_2^{\otimes(n-j)}$ for its embedding on qubit $j$. Then

$$i(U^{\mathrm{S}})^\dagger \partial_a U^{\mathrm{S}} = U_{<\ell}^\dagger \big( i R_{\ell,j}^\dagger \partial_a R_{\ell,j} \big)^{(j)} U_{<\ell},$$

so it suffices to compute the three local generators of $R_{\ell,j}$.

Recall that $R_P(t) = \exp(-itP/2)$ for $P \in \{Y, Z\}$. For the rightmost rotation, which contains $\phi_{\ell,j}$, one has

$$i R_{\ell,j}^\dagger \partial_{\phi_{\ell,j}} R_{\ell,j} = \frac{1}{2} Z.$$

For the middle rotation,

$$i R_{\ell,j}^\dagger \partial_{\beta_{\ell,j}} R_{\ell,j} = \frac{1}{2} R_Z(\phi_{\ell,j})^\dagger Y R_Z(\phi_{\ell,j}) = \frac{1}{2} \big( \sin\phi_{\ell,j}\, X + \cos\phi_{\ell,j}\, Y \big).$$

For the leftmost rotation,

$$i R_{\ell,j}^\dagger \partial_{\omega_{\ell,j}} R_{\ell,j} = \frac{1}{2} R_Z(\phi_{\ell,j})^\dagger R_Y(\beta_{\ell,j})^\dagger Z R_Y(\beta_{\ell,j}) R_Z(\phi_{\ell,j}),$$

and the identities

$$R_Y(\beta)^\dagger Z R_Y(\beta) = \cos\beta\, Z - \sin\beta\, X, \qquad R_Z(\phi)^\dagger X R_Z(\phi) = \cos\phi\, X - \sin\phi\, Y$$

give

$$i R_{\ell,j}^\dagger \partial_{\omega_{\ell,j}} R_{\ell,j} = \frac{1}{2} \big( -\sin\beta_{\ell,j} \cos\phi_{\ell,j}\, X + \sin\beta_{\ell,j} \sin\phi_{\ell,j}\, Y + \cos\beta_{\ell,j}\, Z \big).$$

Embedding these local matrices on qubit $j$ and conjugating by $U_{<\ell}$ proves (19)–(21).

It remains to derive the conservation-law condition. Let $\Phi^{\mathrm{S}}(\theta) = (U^{\mathrm{S}}(\theta))^\dagger M_0 U^{\mathrm{S}}(\theta)$ be the measurement map of the Strong PQC. By (11), $\partial_a \Phi^{\mathrm{S}}(\theta) = i[K_a(\theta), \Phi^{\mathrm{S}}(\theta)]$, and hence, for every $h \in C^1(\Theta)$,

$$\mathrm{d}(\Phi^{\mathrm{S}})_\theta \big( \nabla h(\theta) \big) = i \left[ \sum_a \partial_a h(\theta) K_a(\theta), \Phi^{\mathrm{S}}(\theta) \right],$$

where the sum runs over all $3Ln$ angle coordinates. By Theorem 3.3, $h$ is a data-independent conservation law if and only if

$$\left[ \sum_{\ell=1}^{L} \sum_{j=1}^{n} \left( \partial_{\phi_{\ell,j}} h\, K_{\phi_{\ell,j}} + \partial_{\beta_{\ell,j}} h\, K_{\beta_{\ell,j}} + \partial_{\omega_{\ell,j}} h\, K_{\omega_{\ell,j}} \right), \Phi^{\mathrm{S}} \right] = 0 \tag{22}$$

holds at every parameter point in the domain of $h$, all functions in (22) being evaluated at the same $\theta$. □

*Proof of Theorem 4.4.* We first prove the last-layer statements for arbitrary depth, and then establish local completeness in the single-layer case.

*(i) General stabilizer conditions.* Recall the effective last-layer observable $\Phi_L^{\mathrm{S}} = (E_L^{\mathrm{S}})^\dagger M_0 E_L^{\mathrm{S}}$. Since $U^{\mathrm{S}} = E_L^{\mathrm{S}} R_L^{\mathrm{S}} U_{<L}$, the measurement map factors as

$$\Phi^{\mathrm{S}}(\theta) = U_{<L}^\dagger (R_L^{\mathrm{S}})^\dagger \Phi_L^{\mathrm{S}} R_L^{\mathrm{S}} U_{<L}.$$

The prefix $U_{<L}$ does not involve any last-layer parameter, so it suffices to determine which last-layer coordinates occur in $(R_L^{\mathrm{S}})^\dagger \Phi_L^{\mathrm{S}} R_L^{\mathrm{S}}$.

Introduce the full-register rotation blocks

$$\mathsf{Z}_\omega = \bigotimes_{k=1}^{n} R_Z(\omega_{L,k}), \qquad \mathsf{Y}_\beta = \bigotimes_{k=1}^{n} R_Y(\beta_{L,k}), \qquad \mathsf{Z}_\phi = \bigotimes_{k=1}^{n} R_Z(\phi_{L,k}),$$

so that $R_L^{\mathrm{S}} = \mathsf{Z}_\omega \mathsf{Y}_\beta \mathsf{Z}_\phi$. If $[\Phi_L^{\mathrm{S}}, Z_j] = 0$, then $\Phi_L^{\mathrm{S}}$ commutes with the rotation $R_Z(\omega_{L,j})$ acting on qubit $j$. Since this rotation also commutes with every other factor of $\mathsf{Z}_\omega$, its two occurrences in $\mathsf{Z}_\omega^\dagger \Phi_L^{\mathrm{S}} \mathsf{Z}_\omega$ cancel, and consequently

$$\partial_{\omega_{L,j}} \Phi^{\mathrm{S}}(\theta) = 0.$$

By Theorem 3.3, $\omega_{L,j}$ is a data-independent conservation law.

Suppose next that $\Phi_L^{\mathrm{S}}$ acts as the identity on qubit $j$. Ordering the tensor factors so that qubit $j$ comes first, write $\Phi_L^{\mathrm{S}} = I_2 \otimes B$ with $B \in \mathrm{Herm}(2^{n-1})$ acting on the remaining qubits, and $R_L^{\mathrm{S}} = R_{L,j} \otimes R_{L,j^{\mathrm{c}}}$, where $j^{\mathrm{c}} = \{1, \dots, n\} \setminus \{j\}$ and $R_{L,j^{\mathrm{c}}}$ is the tensor product of the other local blocks. Then

$$(R_L^{\mathrm{S}})^\dagger \Phi_L^{\mathrm{S}} R_L^{\mathrm{S}} = (R_{L,j}^\dagger I_2 R_{L,j}) \otimes (R_{L,j^{\mathrm{c}}}^\dagger B R_{L,j^{\mathrm{c}}}) = I_2 \otimes (R_{L,j^{\mathrm{c}}}^\dagger B R_{L,j^{\mathrm{c}}}),$$

which does not depend on any of the three parameters of $R_{L,j}$. By the same criterion, $\phi_{L,j}$, $\beta_{L,j}$, and $\omega_{L,j}$ are laws.

*(ii) Pauli-Z string specialization.* Suppose $\Phi_L^{\mathrm{S}} = Z(\mathcal{J})$, where $\mathcal{J} \subseteq \{1, \dots, n\}$ and $Z(\mathcal{J}) = \prod_{j \in \mathcal{J}} Z_j$, the empty product being $I_{2^n}$. For $j \in \mathcal{J}$, the effective observable commutes with $Z_j$, so $\omega_{L,j}$ is a law by (i). For $j \notin \mathcal{J}$, the effective observable acts as the identity on qubit $j$, so all three last-layer coordinates of that qubit are laws. These are distinct coordinate functions, hence functionally independent, and their number is

$$|\mathcal{J}| + 3(n - |\mathcal{J}|) = 3n - 2|\mathcal{J}|.$$

This proves the stated family for arbitrary depth $L$.

It remains to establish local completeness when $L = 1$. Define the coordinate law map and the open dense parameter set

$$\mathbf{h}_{\mathcal{J},1}(\theta) = \big((\omega_{1,j})_{j \in \mathcal{J}}, (\phi_{1,j}, \beta_{1,j}, \omega_{1,j})_{j \notin \mathcal{J}}\big), \qquad \Omega_{\mathcal{J}} = \{\theta \in \Theta : \sin \beta_{1,j} \neq 0 \text{ for every } j \in \mathcal{J}\}.$$

For a single layer, the measurement map reads

$$\Phi^{\mathrm{S}}(\theta) = (R_1^{\mathrm{S}})^\dagger Z(\mathcal{J}) R_1^{\mathrm{S}}.$$

For each $j \in \mathcal{J}$, consider the local Hermitian matrix

$$\mu_j = R_{1,j}^\dagger Z R_{1,j} = -\sin \beta_{1,j} \cos \phi_{1,j}\, X + \sin \beta_{1,j} \sin \phi_{1,j}\, Y + \cos \beta_{1,j}\, Z,$$

in which $\omega_{1,j}$ does not occur because $R_Z(\omega_{1,j})$ commutes with $Z$. Hence

$$\Phi^{\mathrm{S}}(\theta) = \bigotimes_{j=1}^{n} \widehat{\mu}_j, \qquad \widehat{\mu}_j = \begin{cases} \mu_j, & j \in \mathcal{J}, \\ I_2, & j \notin \mathcal{J}. \end{cases}$$

For $j \in \mathcal{J}$, the derivatives of $\mu_j$ with respect to its two active angles are

$$\mu_j^{(\phi)} := \partial_{\phi_{1,j}} \mu_j = \sin \beta_{1,j} \big(\sin \phi_{1,j}\, X + \cos \phi_{1,j}\, Y\big),$$
$$\mu_j^{(\beta)} := \partial_{\beta_{1,j}} \mu_j = -\cos \beta_{1,j} \cos \phi_{1,j}\, X + \cos \beta_{1,j} \sin \phi_{1,j}\, Y - \sin \beta_{1,j}\, Z.$$

Using $\mathrm{tr}(P^2) = 2$ and $\mathrm{tr}(PP') = 0$ for distinct $P, P' \in \{X, Y, Z\}$, one obtains

$$\mathrm{tr}(\mu_j \mu_j^{(\phi)}) = \mathrm{tr}(\mu_j \mu_j^{(\beta)}) = \mathrm{tr}(\mu_j^{(\phi)} \mu_j^{(\beta)}) = 0 \tag{23}$$

and

$$\mathrm{tr}\big((\mu_j^{(\phi)})^2\big) = 2 \sin^2 \beta_{1,j}, \qquad \mathrm{tr}\big((\mu_j^{(\beta)})^2\big) = 2.$$

Thus $\mu_j^{(\phi)}$ and $\mu_j^{(\beta)}$ are nonzero and linearly independent whenever $\sin \beta_{1,j} \neq 0$.

The global derivatives $\partial_{\phi_{1,j}} \Phi^{\mathrm{S}}$ and $\partial_{\beta_{1,j}} \Phi^{\mathrm{S}}$, $j \in \mathcal{J}$, are pairwise Hilbert–Schmidt orthogonal: for the two derivatives at the same qubit this follows from $\mathrm{tr}(\mu_j^{(\phi)} \mu_j^{(\beta)}) = 0$, while for derivatives at different qubits the tensor-product inner product contains a factor $\mathrm{tr}(\mu_j^{(\phi)} \mu_j)$ or $\mathrm{tr}(\mu_j^{(\beta)} \mu_j)$, which vanishes by (23). Since every factor $\widehat{\mu}_j$ squares to $I_2$, their squared norms are

$$\|\partial_{\phi_{1,j}} \Phi^{\mathrm{S}}\|_{\mathrm{HS}}^2 = 2^n \sin^2 \beta_{1,j}, \qquad \|\partial_{\beta_{1,j}} \Phi^{\mathrm{S}}\|_{\mathrm{HS}}^2 = 2^n,$$

where $\|B\|_{\mathrm{HS}}^2 = \mathrm{tr}(B^2)$ for Hermitian $B$. These $2|\mathcal{J}|$ derivatives are therefore linearly independent on $\Omega_{\mathcal{J}}$. Since all remaining coordinate derivatives vanish,

$$\mathrm{rank}\, \mathrm{d}(\Phi^{\mathrm{S}})_\theta = 2|\mathcal{J}|, \qquad \dim \mathcal{W}_Q(\theta) = 2|\mathcal{J}|, \qquad \theta \in \Omega_{\mathcal{J}},$$

the second equality being Proposition B.1.

The map $\mathbf{h}_{\mathcal{J},1}$ is a coordinate projection of rank $3n - 2|\mathcal{J}|$, so its differential has a kernel of dimension $2|\mathcal{J}|$. Each of its components is a law, and is therefore annihilated by every measurement-gradient field and by all of their iterated brackets. Hence

$$\mathcal{W}_Q(\theta) \subseteq \mathcal{L}_Q(\theta) \subseteq \ker \mathrm{d}(\mathbf{h}_{\mathcal{J},1})_\theta.$$

On $\Omega_{\mathcal{J}}$ the outer two spaces have the same dimension, so all three coincide:

$$\mathcal{W}_Q(\theta) = \mathcal{L}_Q(\theta) = \ker \mathrm{d}(\mathbf{h}_{\mathcal{J},1})_\theta, \qquad r_0(\theta) = r_{\mathrm{Lie}}(\theta) = 2|\mathcal{J}|.$$

By Theorem 3.5, there are exactly $3n - 2|\mathcal{J}|$ independent local laws on this set.

Finally, we verify functional completeness directly. Let $g$ be a local smooth law on an open subset of $\Omega_{\mathcal{J}}$. Its differential annihilates $\ker \mathrm{d}\mathbf{h}_{\mathcal{J},1}$, and since $\mathbf{h}_{\mathcal{J},1}$ is a coordinate projection, $g$ does not depend on the complementary coordinates on a sufficiently small product neighborhood. Thus $g = \psi(\mathbf{h}_{\mathcal{J},1})$ locally for some smooth $\psi$. Conversely, every such function is a law by the chain rule. □

### B.3.4 Example 4.5

Consider $E = \mathrm{CNOT}_{2\to1}\,\mathrm{CNOT}_{1\to2}$, $U^{\mathrm{S}} = ER_2^{\mathrm{S}}ER_1^{\mathrm{S}}$, and $M_0 = Z_1Z_2$, with the Euler blocks defined above. The measurement map is $\Phi^{\mathrm{S}}(\theta) = (U^{\mathrm{S}}(\theta))^\dagger M_0 U^{\mathrm{S}}(\theta)$, and the open parameter domain $\Theta \subseteq \mathbb{R}^{12}$ has twelve real angle coordinates. On the open dense set $\Omega_{\mathrm{S}}$ defined in Appendix B.1, we prove

$$\mathrm{rank}\, \mathrm{d}(\Phi^{\mathrm{S}})_\theta = 6, \qquad \dim \mathcal{L}_Q(\theta) = 7. \tag{24}$$

We also prove that the five stated laws are locally complete near every parameter point, including points outside $\Omega_{\mathrm{S}}$.

*Proof.* We proceed in four steps: we factor the measurement map and read off the five laws; we compute the differential kernel and its orthogonal complement; we use one Lie bracket to determine the Lie-closure rank; and we establish local completeness.

*Factorization and the fifth law.* For the two-qubit entangler $E$, direct CNOT conjugation gives

$$E^\dagger(Z\otimes Z)E = Z\otimes I_2, \qquad E^\dagger(Z\otimes I_2)E = I_2\otimes Z, \tag{25}$$

$$E^\dagger(X\otimes I_2)E = X\otimes X, \qquad E^\dagger(Y\otimes I_2)E = X\otimes Y. \tag{26}$$

Introduce the scalar variables and single-qubit unitaries

$$\begin{aligned} t &= \beta_{2,1}, \quad & s &= \phi_{2,1} + \omega_{1,2}, \\ A &= R_{1,1}, \quad & V &= R_Z(s)R_Y(\beta_{1,2})R_Z(\phi_{1,2}), \end{aligned} \tag{27}$$

together with the single-qubit Hermitian matrices

$$a = A^\dagger X A, \qquad b = V^\dagger Z V, \qquad c = V^\dagger X V. \tag{28}$$

These matrices are traceless, satisfy $a^2 = b^2 = c^2 = I_2$, and obey $\mathrm{tr}(bc) = 0$. Combining

$$R_{2,1}^\dagger Z R_{2,1} = \cos t\, Z - \sin t \cos\phi_{2,1}\, X + \sin t \sin\phi_{2,1}\, Y$$

with (25)–(26), we obtain

$$\Phi^{\mathrm{S}}(\theta) = \cos t\, I_2 \otimes b - \sin t\, a \otimes c. \tag{29}$$

The four coordinates $\omega_{2,1}, \phi_{2,2}, \beta_{2,2}, \omega_{2,2}$ do not occur in this expression and are therefore conservation laws. Moreover, $\phi_{2,1}$ and $\omega_{1,2}$ enter only through their sum $s$, so that

$$\partial_{\phi_{2,1}}\Phi^{\mathrm{S}} = \partial_{\omega_{1,2}}\Phi^{\mathrm{S}},$$

and (10) shows that $\phi_{2,1} - \omega_{1,2}$ is a law as well. This yields the law map

$$\mathbf{h}^{\mathrm{S}}(\theta) = (\omega_{2,1}, \phi_{2,2}, \beta_{2,2}, \omega_{2,2}, \phi_{2,1} - \omega_{1,2}).$$

Writing $\mathbf{h}_i^{\mathrm{S}}$ for its $i$-th component, define

$$K = \operatorname{span}\{\nabla \mathbf{h}_1^{\mathrm{S}}, \dots, \nabla \mathbf{h}_5^{\mathrm{S}}\} \subseteq \mathbb{R}^{12}. \tag{30}$$

These gradients are constant and linearly independent, so $K$ is a fixed five-dimensional subspace. Since every component of $\mathbf{h}^{\mathrm{S}}$ is a law,

$$K \subseteq \ker \mathrm{d}(\Phi^{\mathrm{S}})_\theta, \qquad \ker \mathrm{d}(\mathbf{h}^{\mathrm{S}})_\theta = K^\perp$$

for every $\theta \in \Theta$.

*The differential has rank six.* Fix $\theta \in \Omega_{\mathrm{S}}$ and a parameter variation $\delta\theta \in \mathbb{R}^{12}$, and for any differentiable scalar- or matrix-valued function $G$ write $\delta G = \mathrm{d}G_\theta(\delta\theta)$. Suppose that $\delta\Phi^{\mathrm{S}} = 0$. Taking the partial trace over the first qubit in the variation of (29) gives

$$-\sin t\, \delta t\, b + \cos t\, \delta b = 0.$$

Differentiating $\operatorname{tr}(b^2) = 2$ yields $\operatorname{tr}(b\, \delta b) = 0$, so the Hilbert–Schmidt inner product of the last display with $b$ reduces to $\sin t\, \delta t = 0$; since $\sin t \neq 0$ on $\Omega_{\mathrm{S}}$, this forces $\delta t = 0$, and since $\cos t \neq 0$ on the same set, also $\delta b = 0$. The variation equation then reduces to

$$\delta a \otimes c + a \otimes \delta c = 0.$$

Applying the linear functional $B \mapsto \operatorname{tr}(cB)$ to the second tensor factor and using $\operatorname{tr}(c^2) = 2$, $\operatorname{tr}(c\, \delta c) = 0$ gives $2\delta a = 0$; applying $B \mapsto \operatorname{tr}(aB)$ to the first tensor factor then gives $2\delta c = 0$. Conversely, these four conditions make the variation of (29) vanish. Hence

$$\delta\Phi^{\mathrm{S}} = 0 \quad \Longleftrightarrow \quad \delta t = 0, \quad \delta a = 0, \quad \delta b = 0, \quad \delta c = 0. \tag{31}$$

We next translate the conditions on $b$ and $c$ into angle coordinates. Since $V \in \mathrm{SU}(2)$, the matrix $2i\, \delta V V^\dagger$ is traceless and Hermitian, and the conditions $\delta b = \delta c = 0$ mean that it commutes with both $Z$ and $X$. The only traceless Hermitian matrix with this property is zero, so $\delta V = 0$. In the Pauli basis $(X, Y, Z)$, the linear map from $(\delta\phi_{1,2}, \delta\beta_{1,2}, \delta s)$ to the coefficients of $2i\, \delta V V^\dagger$ has matrix

$$\begin{pmatrix} \sin\beta_{1,2}\cos s & -\sin s & 0 \\ \sin\beta_{1,2}\sin s & \cos s & 0 \\ \cos\beta_{1,2} & 0 & 1 \end{pmatrix},$$

whose determinant $\sin\beta_{1,2}$ is nonzero on $\Omega_{\mathrm{S}}$. Consequently,

$$\delta b = \delta c = 0 \quad \Longleftrightarrow \quad \delta\phi_{1,2} = \delta\beta_{1,2} = \delta s = 0. \tag{32}$$

For the block $A$, set $x = \phi_{1,1}$, $y = \beta_{1,1}$, and $z = \omega_{1,1}$. Differentiating $A = R_Z(z)R_Y(y)R_Z(x)$ gives

$$\begin{aligned} 2i\, \delta A A^\dagger = {} & \delta x\big(\sin y \cos z\, X + \sin y \sin z\, Y + \cos y\, Z\big) \\ & + \delta y\big(-\sin z\, X + \cos z\, Y\big) + \delta z\, Z. \end{aligned}$$

Because $a = A^\dagger X A$, the condition $\delta a = 0$ is equivalent to this matrix commuting with $X$, that is, to the vanishing of its $Y$ and $Z$ coefficients:

$$\sin y \sin z\, \delta x + \cos z\, \delta y = 0, \qquad \cos y\, \delta x + \delta z = 0. \tag{33}$$

When $\sin y \neq 0$ these two constraints are independent, and their common kernel is spanned by

$$k = \cos z\, \partial_x - \sin z \sin y\, \partial_y - \cos z \cos y\, \partial_z, \tag{34}$$

where $\partial_x, \partial_y, \partial_z$ denote the coordinate vector fields; we regard $k$ as a vector field on the full parameter space by setting its remaining nine components to zero.

Collecting the above, the four absent coordinates give four unconstrained directions, and the single condition $\delta s = 0$ on the pair $(\phi_{2,1}, \omega_{1,2})$ leaves the direction $\partial_{\phi_{2,1}} - \partial_{\omega_{1,2}}$ free. These five directions span $K$, and the only remaining null direction is $k(\theta)$. Hence

$$\ker \mathrm{d}(\Phi^{\mathrm{S}})_\theta = K \oplus \operatorname{span}\{k(\theta)\}, \qquad \theta \in \Omega_{\mathrm{S}}. \tag{35}$$

Since $k(\theta)$ is nonzero on $\Omega_{\mathrm{S}}$ and is supported on coordinates on which every vector of $K$ vanishes, the sum is direct and the kernel has dimension six. As the parameter space has dimension twelve, the measurement differential has rank six.

*One Lie bracket gives the seventh direction.* By (14) and (35), one has

$$\mathcal{W}_Q(\theta) = K^{\perp} \cap \big(\mathrm{span}\{k(\theta)\}\big)^{\perp}, \qquad \theta \in \Omega_{\mathrm{S}}.$$

This space contains the four constant vector fields

$$\partial_{\phi_{1,2}}, \qquad \partial_{\beta_{1,2}}, \qquad \partial_{\beta_{2,1}}, \qquad \partial_{\phi_{2,1}} + \partial_{\omega_{1,2}}, \tag{36}$$

as well as the two smooth fields supported on $(x, y, z)$,

$$P = \sin z \sin y\, \partial_x + \cos z\, \partial_y, \qquad Q = \cos y\, \partial_x + \partial_z, \tag{37}$$

which are orthogonal to both $K$ and $k(\theta)$. With the convention $[P, Q] = DQ\,P - DP\,Q$, a direct computation gives

$$[P, Q] = -2 \cos z \sin y\, \partial_x + \sin z\, \partial_y. \tag{38}$$

In the coordinates $(x, y, z)$, the determinant of the three coefficient vectors is

$$\det\big(P, Q, [P, Q]\big) = -\sin y\big(1 + \cos^2 z\big) \neq 0, \qquad \theta \in \Omega_{\mathrm{S}}, \tag{39}$$

so $P$, $Q$, and $[P, Q]$ span all three first-layer parameter directions on qubit one. Together with (36), they span $K^{\perp}$, which has dimension seven.

It remains to relate these fields to the Lie algebra generated by the measurement gradients. Fix a sufficiently small neighborhood in $\Omega_{\mathrm{S}}$. Since $\mathcal{W}_Q$ has constant rank six there, six of the probe-gradient fields form a smooth local basis of it, and every smooth field taking values in $\mathcal{W}_Q$ is a smooth linear combination of this basis. For smooth scalar functions $f, g$ and vector fields $X, Y$,

$$[fX, gY] = fg[X, Y] + f(Xg)Y - g(Yf)X,$$

so the bracket of two such combinations takes values in the pointwise span of the original gradient fields and their brackets. In particular, $[P, Q](\theta) \in \mathcal{L}_Q(\theta)$, and the same holds for the six fields already taking values in $\mathcal{W}_Q$.

Conversely, every measurement-gradient field annihilates each component of $\mathbf{h}^{\mathrm{S}}$, and so do all iterated brackets; hence $\mathcal{L}_Q(\theta) \subseteq K^{\perp}$. Combined with the seven independent directions constructed above, this gives

$$\mathcal{L}_Q(\theta) = K^{\perp} = \ker \mathrm{d}(\mathbf{h}^{\mathrm{S}})_{\theta}, \qquad \theta \in \Omega_{\mathrm{S}}. \tag{40}$$

Thus the Lie-closure rank is seven, which establishes (24).

*Completeness for $C^1$ laws, including singular points.* Let $O \subseteq \Theta$ be open and let $g \in C^1(O)$ be a conservation law. On $O \cap \Omega_{\mathrm{S}}$, the differential of $g$ annihilates every field taking values in $\mathcal{W}_Q$, in particular those in (36) and (37). We show that it also annihilates $[P, Q]$, without assuming that $g$ is twice differentiable.

Let $\varphi^P_{\varepsilon}$ and $\varphi^Q_{\varepsilon}$ be the local flows of $P$ and $Q$. Since $Pg = Qg = 0$, both flows preserve $g$ as long as they stay in $O \cap \Omega_{\mathrm{S}}$. For a fixed point $\theta$ of this set and sufficiently small $\varepsilon$, the composition

$$\Psi_{\varepsilon} = \varphi^Q_{-\varepsilon} \circ \varphi^P_{-\varepsilon} \circ \varphi^Q_{\varepsilon} \circ \varphi^P_{\varepsilon}$$

is well defined, preserves $g$, and satisfies

$$\Psi_{\varepsilon}(\theta) = \theta + \varepsilon^2 [P, Q](\theta) + O(\varepsilon^3).$$

Inserting this into the first-order Taylor expansion of the $C^1$ function $g$ at $\theta$, using $g(\Psi_{\varepsilon}(\theta)) = g(\theta)$, dividing by $\varepsilon^2$, and letting $\varepsilon \to 0$, we obtain

$$\mathrm{d}g_{\theta}\big([P, Q](\theta)\big) = 0.$$

Since the fields in (36) together with $P, Q, [P, Q]$ span $K^{\perp}$, it follows that

$$\nabla g(\theta) \in K, \qquad \theta \in O \cap \Omega_{\mathrm{S}}.$$

Because $\Omega_{\mathrm{S}}$ is dense, $K$ is a fixed closed subspace, and $\nabla g$ is continuous, this inclusion extends to all $\theta \in O$.

The linear map $\mathbf{h}^{\mathrm{S}}$ has rank five. Complete its components to linear coordinates $(u, w)$ on $\mathbb{R}^{12}$ with $u = \mathbf{h}^{\mathrm{S}}(\theta)$ and $w \in \mathbb{R}^7$. The directions that vary $w$ while keeping $u$ fixed lie in $\ker \mathrm{d}\mathbf{h}^{\mathrm{S}} = K^{\perp}$, and since $\nabla g \in K$, the derivatives of $g$ along these directions vanish. On a sufficiently small product neighborhood, $g$ therefore depends on $u$ alone:

$$g(\theta) = \psi\big(\mathbf{h}^{\mathrm{S}}(\theta)\big)$$

for some $C^1$ function $\psi$.

Conversely, by the chain rule, the gradient of any local $C^1$ function of $\mathbf{h}^{\mathrm{S}}$ lies in $K$, and since $K \subseteq \ker \mathrm{d}(\Phi^{\mathrm{S}})_\theta$ at every parameter point, Theorem 3.3 shows that every such function is a conservation law. Hence the five components of $\mathbf{h}^{\mathrm{S}}$ form a complete family of local $C^1$ laws near every parameter point. □

### B.3.5 Proof of Theorem 4.6

We use the regular set $\Omega_{\mathrm{C}}$ and the Pauli notation from Appendix B.1. The site sets $S_0, S_X, S_Y, S_Z, S_3$ and their sizes $n_T = |S_T|$ are those defined in the theorem.

The distinct-support assumption decouples the measurement differential into single-qubit calculations. We compute the local kernels and use Lie brackets to prove completeness on the regular set; a continuity argument then extends local completeness to every parameter point.

*Proof.* We first identify the coordinate laws and then separate the measurement differential according to Pauli support, which reduces the rank and completeness calculations to individual Euler blocks.

For $\alpha \in \{X, Y, Z\}$, write $P_\alpha = \epsilon_\alpha \bigotimes_{j=1}^{n} p_{\alpha,j}$ with $\epsilon_\alpha \in \{\pm 1\}$ and $p_{\alpha,j} \in \{I_2, X, Y, Z\}$, and set $\mu_{\alpha,j} = R_{1,j}^{\dagger} p_{\alpha,j} R_{1,j}$. Under the readout assumption $E_2^{\dagger} M_0 E_2 = Z_k$, the coefficients

$$a_X = -\sin\beta_{2,k}\cos\phi_{2,k}, \qquad a_Y = \sin\beta_{2,k}\sin\phi_{2,k}, \qquad a_Z = \cos\beta_{2,k}$$

give the measurement map

$$\begin{aligned} \Phi(\theta) &= \sum_{\alpha\in\{X,Y,Z\}} a_\alpha T_\alpha, \\ T_\alpha &= (R_1^{\mathrm{S}})^{\dagger} P_\alpha R_1^{\mathrm{S}} = \epsilon_\alpha \bigotimes_{j=1}^{n} \mu_{\alpha,j}. \end{aligned} \tag{41}$$

Let $\mathbf{h}^{\mathrm{C}}(\theta)$ be the vector collecting the coordinate functions listed in (13), in a fixed order. Neither the last-layer coordinates listed in (13) nor the first-layer coordinates at sites in $S_0$ occur in this expression. At a site in $S_Z$, every nonidentity local factor equals $R_{1,j}^{\dagger} Z R_{1,j}$, which does not depend on $\omega_{1,j}$. Hence every component of $\mathbf{h}^{\mathrm{C}}$ is a global coordinate law.

Writing $\mathbf{h}_i^{\mathrm{C}}$ for the $i$-th component of this law map, define

$$K = \mathrm{span}\{\nabla \mathbf{h}_i^{\mathrm{C}} : 1 \le i \le q\}, \qquad q = 3n - 2 + 3n_0 + n_Z.$$

These gradients are distinct coordinate vectors, so $K$ is a fixed subspace of dimension $q$, and

$$K \subseteq \ker \mathrm{d}\Phi_\theta, \qquad \ker \mathrm{d}(\mathbf{h}^{\mathrm{C}})_\theta = K^{\perp}$$

at every parameter point.

*Separation of the differential by support.* Fix $\theta \in \Omega_{\mathrm{C}}$ and a variation $\delta\theta \in \mathbb{R}^{6n}$, and for any differentiable quantity $G$ write $\delta G = \mathrm{d}G_\theta(\delta\theta)$. Recall that $\mathcal{V}_S^{\mathrm{supp}}$ is the real span of tensor products with traceless Hermitian factors on $S$ and identity factors on the complement of $S$. These spaces are pairwise Hilbert–Schmidt orthogonal: when two supports differ, at least one local inner product pairs a traceless matrix with the identity.

Both $T_\alpha$ and $\delta T_\alpha$ lie in $\mathcal{V}_{\mathrm{supp}(P_\alpha)}^{\mathrm{supp}}$, because local unitary conjugation preserves tracelessness and the identity factors have zero derivative. By the distinct-support assumption (16),

$$\delta\Phi = 0 \quad \Longleftrightarrow \quad \delta a_\alpha T_\alpha + a_\alpha \delta T_\alpha = 0 \quad \text{for every } \alpha \in \{X, Y, Z\}.$$

For this support calculation we use the normalized inner product $\langle B, C\rangle = 2^{-n}\operatorname{tr}(BC)$ on the full register and $\frac{1}{2}\operatorname{tr}(BC)$ on a single qubit, with the induced norms. Since $\|T_\alpha\| = 1$, differentiation gives $\langle T_\alpha, \delta T_\alpha\rangle = 0$, so taking the inner product of the preceding equation with $T_\alpha$ yields $\delta a_\alpha = 0$. As all three coefficients $a_\alpha$ are nonzero on $\Omega_{\mathrm{C}}$, it follows that $\delta T_\alpha = 0$ as well.

Each local factor $\mu_{\alpha,j}$ has normalized norm one and is orthogonal to its own derivative, so the terms in

$$\delta T_\alpha = \epsilon_\alpha \sum_{j:p_{\alpha,j}\neq I_2} \mu_{\alpha,1}\otimes\cdots\otimes\delta\mu_{\alpha,j}\otimes\cdots\otimes\mu_{\alpha,n}$$

are pairwise orthogonal. Consequently,

$$\|\delta T_\alpha\|^2 = \sum_{j:p_{\alpha,j}\neq I_2} \frac{1}{2}\operatorname{tr}\big((\delta\mu_{\alpha,j})^2\big),$$

and $\delta T_\alpha = 0$ if and only if each displayed local derivative vanishes. Altogether,

$$\delta\Phi = 0 \quad\Longleftrightarrow\quad \begin{cases}\delta a_X = \delta a_Y = \delta a_Z = 0,\\ \delta\mu_{\alpha,j} = 0 \quad \text{whenever } p_{\alpha,j}\neq I_2.\end{cases} \tag{42}$$

The first line involves only $(\phi_{2,k}, \beta_{2,k})$. The derivatives of $(a_X, a_Y, a_Z)$ with respect to these two angles are orthogonal vectors in $\mathbb{R}^3$ with squared norms $\sin^2\beta_{2,k}$ and 1; since $\sin\beta_{2,k}\neq 0$ on $\Omega_{\mathrm{C}}$, these conditions have rank two.

*The local Euler blocks.* At a first-layer site, write $(x, y, z) = (\phi_{1,j}, \beta_{1,j}, \omega_{1,j})$ and $A = R_Z(z)R_Y(y)R_Z(x)$. Its variation satisfies

$$\begin{aligned}2i\,\delta A A^\dagger = \delta x\big(\sin y\cos z\,X + \sin y\sin z\,Y + \cos y\,Z\big)\\ + \delta y\big(-\sin z\,X + \cos z\,Y\big) + \delta z\,Z.\end{aligned}$$

For a local Pauli matrix $T\in\{X, Y, Z\}$, the condition $\delta(A^\dagger T A) = 0$ is equivalent to $2i\,\delta A A^\dagger$ commuting with $T$. Throughout, recall that $\sin y\neq 0$ at every site outside $S_0$ on $\Omega_{\mathrm{C}}$.

At an $S_X$ site the $Y$ and $Z$ coefficients must vanish, and at an $S_Y$ site the $X$ and $Z$ coefficients must vanish. Solving these two linear systems gives the respective kernel fields

$$k_X = \cos z\,\partial_x - \sin z\sin y\,\partial_y - \cos z\cos y\,\partial_z, \tag{43}$$

$$k_Y = \sin z\,\partial_x + \cos z\sin y\,\partial_y - \sin z\cos y\,\partial_z, \tag{44}$$

where $\partial_x, \partial_y, \partial_z$ denote the local coordinate vector fields. At an $S_Z$ site, vanishing of the $X$ and $Y$ coefficients forces $\delta x = \delta y = 0$, while $\delta z$ is unrestricted. At an $S_3$ site, the generator must commute with two distinct Pauli matrices; being traceless and Hermitian, it must vanish, and since its coefficient matrix in $(\delta x, \delta y, \delta z)$ has determinant $\sin y$, all three variations vanish. At an $S_0$ site, every local factor is the identity and every coordinate direction lies in the kernel.

The local ranks and pointwise kernels are summarized as follows.

| Site type | Local differential rank | Local differential kernel |
|---|---|---|
| $S_0$ | 0 | $\operatorname{span}\{\partial_x, \partial_y, \partial_z\}$ |
| $S_Z$ | 2 | $\operatorname{span}\{\partial_z\}$ |
| $S_X$ | 2 | $\operatorname{span}\{k_X(x, y, z)\}$ |
| $S_Y$ | 2 | $\operatorname{span}\{k_Y(x, y, z)\}$ |
| $S_3$ | 3 | $\{0\}$ |

For $j\in S_X$, let $k_{X,j}$ be the field obtained from (43) by inserting the angles of qubit $j$ and setting all other parameter components to zero; define $k_{Y,j}$ analogously for $j\in S_Y$.

All conditions in (42) concern disjoint parameter blocks. The frozen last-layer directions, all first-layer directions on $S_0$, and the $\omega_{1,j}$ directions on $S_Z$ are precisely the directions spanning $K$. Therefore

$$\ker \mathrm{d}\Phi_\theta = K \oplus \bigoplus_{j\in S_X} \operatorname{span}\{k_{X,j}(\theta)\} \oplus \bigoplus_{j\in S_Y} \operatorname{span}\{k_{Y,j}(\theta)\}, \qquad \theta\in\Omega_{\mathrm{C}}. \tag{45}$$

Summing the local ranks and adding the two active last-layer directions gives

$$r_0(\theta) = \operatorname{rank} \mathrm{d}\Phi_\theta = 2 + 2(n_X + n_Y + n_Z) + 3n_3.$$

*Lie brackets supply the missing directions at $S_X$ and $S_Y$ sites.* At an $S_X$ site we use the pair $(F_X, G)$ and at an $S_Y$ site the pair $(F_Y, G)$, where

$$\begin{aligned} F_X &= \sin z \sin y\, \partial_x + \cos z\, \partial_y, \\ F_Y &= -\cos z \sin y\, \partial_x + \sin z\, \partial_y, \\ G &= \cos y\, \partial_x + \partial_z, \end{aligned}$$

each extended by zero outside its site. The pair $(F_X, G)$ is orthogonal to $k_X$, the pair $(F_Y, G)$ is orthogonal to $k_Y$, and both pairs are orthogonal to $K$; by (14) and (45), they therefore take values in the measurement-gradient distribution $\mathcal{W}_Q$.

With the convention $[F, G] = DG\,F - DF\,G$, a direct calculation gives

$$\begin{aligned} [F_X, G] &= -2\cos z \sin y\, \partial_x + \sin z\, \partial_y, \\ [F_Y, G] &= -2\sin z \sin y\, \partial_x - \cos z\, \partial_y, \end{aligned}$$

and the determinants of the coefficient vectors in $(x, y, z)$ are

$$\det(F_X, G, [F_X, G]) = -\sin y(1 + \cos^2 z) \neq 0, \tag{46}$$

$$\det(F_Y, G, [F_Y, G]) = -\sin y(1 + \sin^2 z) \neq 0. \tag{47}$$

Hence each pair together with its bracket spans all three coordinate directions at the corresponding site.

These bracket values lie in the pointwise Lie closure of the original measurement gradients. Indeed, on the constant-rank set $\Omega_{\mathrm{C}}$, a smooth field taking values in $\mathcal{W}_Q$ is locally a smooth linear combination of a basis chosen among the probe-gradient fields, and for smooth scalar functions $f, g$ and vector fields $X, Y$ the identity

$$[fX, gY] = fg[X, Y] + f(Xg)Y - g(Yf)X$$

places its bracket values in the pointwise span of the original gradient fields and their brackets.

Thus (46)–(47) supply the missing third direction at every $S_X$ and $S_Y$ site, while the two active last-layer directions, the two active directions at each $S_Z$ site, and all three directions at each $S_3$ site already belong to $\mathcal{W}_Q$. Together these directions span $K^\perp$. Conversely, every measurement-gradient field annihilates each component of $\mathbf{h}^{\mathrm{C}}$, and so do all iterated brackets. Therefore

$$\mathcal{L}_Q(\theta) = K^\perp = \ker \mathrm{d}(\mathbf{h}^{\mathrm{C}})_\theta, \qquad \theta \in \Omega_{\mathrm{C}},$$

and consequently

$$\begin{aligned} r_{\mathrm{Lie}}(\theta) &= \dim \mathcal{L}_Q(\theta) = 2 + 3(n_X + n_Y + n_3) + 2n_Z, \\ r_{\mathrm{Lie}}(\theta) - r_0(\theta) &= n_X + n_Y. \end{aligned}$$

Since the circuit has $6n$ parameters, these identities also give $6n - r_{\mathrm{Lie}}(\theta) = q$, as claimed.

*Completeness at all parameter points.* Let $O \subseteq \Theta$ be open and let $g \in C^1(O)$ be a conservation law. Its differential annihilates $\mathcal{W}_Q$ on $O$, and hence, on $O \cap \Omega_{\mathrm{C}}$, every local pair $(F_X, G)$ or $(F_Y, G)$ introduced above.

The same holds for their brackets, even though $g$ is only $C^1$. Write $F = F_X$ or $F = F_Y$ as appropriate. Since $Fg = Gg = 0$, the local flows of $F$ and $G$ preserve $g$, and their flow commutator displaces a point $\theta$ by $\varepsilon^2[F, G](\theta) + O(\varepsilon^3)$ for sufficiently small real $\varepsilon$. The first-order Taylor expansion of the $C^1$ function $g$ therefore gives $\mathrm{d}g_\theta([F, G](\theta)) = 0$, exactly as in Appendix B.3.4.

The gradient fields and the brackets above span $K^\perp$ on $\Omega_{\mathrm{C}}$, so

$$\nabla g(\theta) \in K, \qquad \theta \in O \cap \Omega_{\mathrm{C}}.$$

Because $\Omega_{\mathrm{C}}$ is dense, $K$ is a fixed closed subspace, and $\nabla g$ is continuous, this inclusion extends to every $\theta \in O$.

The map $\mathbf{h}^{\mathrm{C}}$ selects $q$ distinct angle coordinates, so $\nabla g \in K$ means precisely that every partial derivative of $g$ with respect to the remaining coordinates vanishes. On a sufficiently small product neighborhood, $g$ therefore depends only on the selected coordinates:

$$g(\theta) = \psi\big(\mathbf{h}^{\mathrm{C}}(\theta)\big)$$

for some $C^1$ function $\psi$. Conversely, by the chain rule the gradient of any such function lies in $K$, and since $K \subseteq \ker \mathrm{d}\Phi_\theta$ at every parameter point, (10) shows that every such function is a conservation law. This proves local completeness near every parameter point. □

### B.3.6 Proofs for the two-layer circuit families

For each family we compute the effective readout and the three conjugated Pauli strings, and then apply Theorem 4.6. We conclude with the upper bound on the rank gap.

*Proof of Corollary 4.7*(i). Let $n \geq 3$ and take

$$E_1 = E_2 = E = \mathrm{CNOT}_{n\to 1}\,\mathrm{CNOT}_{n-1\to n}\cdots\mathrm{CNOT}_{1\to 2}, \qquad M_0 = Z_1 Z_n.$$

CNOT conjugation gives $E^\dagger M_0 E = Z_1$, so the effective readout qubit is $k = 1$, and the three conjugated Pauli strings are

$$P_X = X_1 X_2, \qquad P_Y = X_1 Y_2 Z_3 \cdots Z_n, \qquad P_Z = Z_2 \cdots Z_n.$$

Their supports, $\{1,2\}$, $\{1,\dots,n\}$, and $\{2,\dots,n\}$, are pairwise distinct for $n \geq 3$, so the support assumption of Theorem 4.6 is satisfied.

The local factors of $(P_X, P_Y, P_Z)$ are $(X, X, I_2)$ at qubit one, $(X, Y, Z)$ at qubit two, and $(I_2, Z, Z)$ at every remaining qubit. Hence

$$S_X = \{1\}, \qquad S_3 = \{2\}, \qquad S_Z = \{3,\dots,n\}, \qquad S_0 = S_Y = \varnothing.$$

The theorem yields the complete family

$$\omega_{2,1}, \qquad (\phi_{2,j}, \beta_{2,j}, \omega_{2,j})_{j=2}^{n}, \qquad (\omega_{1,j})_{j=3}^{n},$$

consisting of $1 + 3(n-1) + (n-2) = 4n - 4$ independent global coordinate laws that are locally complete near every parameter point. Substituting $n_X = n_3 = 1$ and $n_Z = n - 2$ into the rank formulas gives

$$r_0(\theta) = 2n + 3, \qquad r_{\mathrm{Lie}}(\theta) = 2n + 4, \qquad r_{\mathrm{Lie}}(\theta) - r_0(\theta) = 1, \qquad \theta \in \Omega_{\mathrm{C}},$$

which proves part (i). □

*Proof of Corollary 4.7*(ii) and (iii). Let $n \geq 3$, set $k = n - 1$, and define the fixed gate blocks

$$A = \prod_{j=1}^{n-2} \mathrm{CNOT}_{k\to j}, \qquad B = \mathrm{CNOT}_{n\to k}, \qquad E = BA\,\mathsf{H}_n,$$

where $\mathsf{H}_n$ applies the Hadamard gate $\mathsf{H} = (X+Z)/\sqrt{2}$ to qubit $n$ and the identity to all other qubits. The CNOT factors in $A$ commute, so their order is immaterial, and each entangler $E$ contains $n-1$ CNOT gates and one Hadamard gate.

Take $E_1 = E_2 = E$ and the two-qubit readout $M_0 = Z_k Z_n$. Direct conjugation gives

$$\begin{aligned} E^\dagger M_0 E &= Z_k, \\ P_X &= X_1 \cdots X_{n-2} X_k, \\ P_Y &= X_1 \cdots X_{n-2} Y_k X_n, \\ P_Z &= Z_k X_n. \end{aligned}$$

The supports of $P_X, P_Y, P_Z$ are $\{1,\dots,n-1\}$, $\{1,\dots,n\}$, and $\{n-1, n\}$, respectively, and these are pairwise distinct for $n \geq 3$.

The local factors are $(X, Y, Z)$ at qubit $k$, $(X, X, I_2)$ at every qubit $j \le n-2$, and $(I_2, X, X)$ at qubit $n$. Hence

$$S_3 = \{k\}, \qquad S_X = \{1, \dots, n\} \setminus \{k\}, \qquad S_0 = S_Y = S_Z = \varnothing.$$

By Theorem 4.6, the $3n-2$ frozen last-layer coordinates

$$\omega_{2,k}, \qquad (\phi_{2,j}, \beta_{2,j}, \omega_{2,j})_{j \neq k}$$

form a family that is locally complete near every parameter point. With $n_X = n-1$ and $n_3 = 1$, the rank formulas give

$$r_0(\theta) = 2n+3, \qquad r_{\mathrm{Lie}}(\theta) = 3n+2, \qquad r_{\mathrm{Lie}}(\theta) - r_0(\theta) = n-1, \qquad \theta \in \Omega_{\mathrm{C}}.$$

It remains to show that no circuit in the classified family has a larger rank gap. For any circuit satisfying the hypotheses of Theorem 4.6, the strings $P_X$ and $P_Z$ anticommute, being unitary conjugates of $X_k$ and $Z_k$. Hence at least one qubit carries distinct nonidentity local factors in these two strings, and at that qubit the relation $P_Y = iP_XP_Z$ supplies the third nonidentity Pauli factor. This qubit belongs to $S_3$, so $n_3 \ge 1$. Since the site sets partition the $n$ qubits, the general rank formula yields

$$r_{\mathrm{Lie}}(\theta) - r_0(\theta) = n_X + n_Y = n - n_0 - n_Z - n_3 \le n-1, \qquad \theta \in \Omega_{\mathrm{C}}.$$

The family constructed above attains equality, which proves sharpness within the stated class. □

### B.3.7 Dimension reduction and gradient measurement cost

We prove the training benefits stated in Section 4.4. We first show that linear laws are preserved under SGD and derive the corresponding dimension reduction. We then reconstruct gradients from selected derivatives and analyze the resulting measurement error, including the additional reduction available for the classified Clifford circuits. Throughout, angles are real coordinates, inputs range over all density matrices, and conservation laws are data-independent. In order to infer relations among measurement gradients from arbitrary linear laws, we assume $\mathcal{W}_{\mathrm{loss}} = \mathcal{W}_Q$, as in Appendix B.2.3; this condition holds for the squared loss.

*Linear laws and dimension reduction.* The following lemma is the discrete-time counterpart of a linear conservation law.

**Lemma B.5** (Preservation of linear laws under SGD)**.** *Let $h(\theta) = v^\top \theta$ be a global linear data-independent conservation law. For any sequence of minibatches $\mathcal{S}_k$ and step sizes $\eta_k > 0$, consider the exact-gradient updates $\theta_{k+1} = \theta_k - \eta_k \nabla_\theta L_{\mathcal{S}_k}(\theta_k)$. Then $h(\theta_k) = h(\theta_0)$ for every iteration $k$ for which the iterates remain in $\Theta$.*

*Proof.* By Proposition 3.1, $v^\top \nabla_\theta L_{\mathcal{S}_k}(\theta_k) = 0$. Since $h$ is linear, one has

$$h(\theta_{k+1}) = h(\theta_k) - \eta_k v^\top \nabla_\theta L_{\mathcal{S}_k}(\theta_k) = h(\theta_k), \tag{48}$$

and the claim follows by induction. □

Consequently, independent linear laws confine both the gradient flow and exact-gradient SGD to an affine subspace, and orthonormal coordinates on this subspace yield an equivalent optimization problem.

**Proposition B.6** (Dimension reduction from linear laws)**.** *Let $h_a(\theta) = v_a^\top \theta$, $a = 1, \dots, q$, be independent global linear laws of a circuit with $p$ parameters. Set $K = \mathrm{span}\{v_1, \dots, v_q\}$ and $r = p - q$, and let $Q \in \mathbb{R}^{p \times r}$ have orthonormal columns spanning $K^\perp$. Fix an initialization $\theta_0 \in \Theta$ and define the reduced parameter domain and the reduced objective*

$$\mathcal{Z}_{\theta_0} = \{z \in \mathbb{R}^r : \theta_0 + Qz \in \Theta\}, \qquad \widetilde{L}_{\mathcal{D}}(z) = L_{\mathcal{D}}(\theta_0 + Qz), \quad z \in \mathcal{Z}_{\theta_0}. \tag{49}$$

*Then the following hold.*

(i) (Gradient flow.) *Let $z(t) \in \mathcal{Z}_{\theta_0}$ solve the reduced gradient flow*

$$\dot{z}(t) = -\eta \nabla_z \widetilde{L}_{\mathcal{D}}(z(t)), \qquad z(0) = 0. \tag{50}$$

*Then $\theta(t) = \theta_0 + Qz(t)$ solves the original gradient flow* (7) *with $\theta(0) = \theta_0$. Conversely, the solution of* (7) *with $\theta(0) = \theta_0$ satisfies $\theta(t) = \theta_0 + Qz(t)$ with $z(t) = Q^\top(\theta(t) - \theta_0)$ solving* (50)*, for as long as $\theta(t) \in \Theta$.*

(ii) (SGD.) *For the same step sizes $\eta_k$ and minibatches $\mathcal{S}_k$, the exact-gradient SGD iterates $\theta_k$ started at $\theta_0$ and the reduced iterates $z_k$ started at $z_0 = 0$ satisfy $\theta_k = \theta_0 + Qz_k$ for every $k$ such that $\theta_k \in \Theta$.*

(iii) *If $r = 0$, then all measurement and loss gradients vanish, and training remains at $\theta_0$.*

*Proof.* We first record a fact used in all three parts. By Theorem 3.3, $\mathrm{d}\Phi_\theta(v_a) = 0$ for every $a$, and hence

$$v_a^\top \nabla F_\rho(\theta) = \mathrm{tr}\big(\rho\, \mathrm{d}\Phi_\theta(v_a)\big) = 0$$

for every input $\rho$ and every $\theta \in \Theta$. Thus all measurement gradients, and therefore all dataset and minibatch loss gradients, lie in $K^\perp$. Since $QQ^\top$ is the orthogonal projection onto $K^\perp$, this gives

$$QQ^\top \nabla L_{\mathcal{D}}(\theta) = \nabla L_{\mathcal{D}}(\theta), \qquad \nabla_z \widetilde{L}_{\mathcal{D}}(z) = Q^\top \nabla L_{\mathcal{D}}(\theta_0 + Qz), \tag{51}$$

where the second identity is the chain rule applied to (49).

(i) Let $z(t)$ solve (50) and set $\theta(t) = \theta_0 + Qz(t)$. Then $\theta(0) = \theta_0$ and, by (51),

$$\dot\theta(t) = Q\dot z(t) = -\eta\, QQ^\top \nabla L_{\mathcal{D}}(\theta(t)) = -\eta\, \nabla L_{\mathcal{D}}(\theta(t)),$$

so $\theta(t)$ solves (7). Conversely, let $\theta(t)$ solve (7) with $\theta(0) = \theta_0$. Since each $h_a$ is conserved, $\theta(t) - \theta_0 \in K^\perp$, and hence $\theta(t) = \theta_0 + Qz(t)$ with $z(t) = Q^\top(\theta(t) - \theta_0)$. Differentiating and using (51),

$$\dot z(t) = Q^\top \dot\theta(t) = -\eta\, Q^\top \nabla L_{\mathcal{D}}(\theta_0 + Qz(t)) = -\eta\, \nabla_z \widetilde{L}_{\mathcal{D}}(z(t)),$$

with $z(0) = 0$, so $z(t)$ solves (50). Both directions hold for as long as $\theta(t) \in \Theta$.

(ii) We argue by induction on $k$. The claim holds for $k = 0$ since $z_0 = 0$. If $\theta_k = \theta_0 + Qz_k$, then by (51) applied to the minibatch loss $L_{\mathcal{S}_k}$,

$$\theta_0 + Qz_{k+1} = \theta_0 + Q\big(z_k - \eta_k \nabla_z \widetilde{L}_{\mathcal{S}_k}(z_k)\big) = \theta_k - \eta_k\, QQ^\top \nabla L_{\mathcal{S}_k}(\theta_k) = \theta_k - \eta_k \nabla L_{\mathcal{S}_k}(\theta_k) = \theta_{k+1}.$$

Lemma B.5 confirms that the iterates $\theta_k$ indeed remain in $\theta_0 + K^\perp$.

(iii) If $r = 0$, then $K^\perp = \{0\}$, so every measurement and loss gradient vanishes and both the flow and the SGD iterates remain at $\theta_0$. □

For instance, if $a - b$ is conserved, one may take $(a + b)/\sqrt{2}$ as the active coordinate and keep $(a - b)/\sqrt{2}$ fixed. Orthonormal coordinates preserve the original Euclidean dynamics, whereas arbitrary coordinates need not. This reduction leaves the parameter trajectories, predictions, and losses unchanged, but does not by itself establish faster convergence, and applying a dense $Q$ incurs a classical cost. The instantaneous measurement rank $r_0(\theta)$ may be smaller than $p - q$; it provides additional relations among derivatives, but not by itself a smaller parametrization that preserves the training dynamics.

*Gradient reconstruction from selected derivatives.* A shifted circuit setting specifies a shifted parameter vector and the readout, and a shot is one independent preparation, circuit execution, and measurement at that setting. For the measurement estimates below, assume $M_0^\dagger = M_0$, $M_0^2 = I$, and that each parameter occurs in exactly one gate $\exp(-i\theta_j H_j/2)$ with $H_j^\dagger = H_j$ and $H_j^2 = I$. The readout condition covers Pauli strings and the classified Clifford circuits, since $E_2^\dagger M_0 E_2 = Z_k$.

Let $e_j$ denote the $j$-th standard coordinate vector. The parameter-shift rule (Schuld et al., 2019) gives

$$\partial_j F_\rho(\theta) = \frac{1}{2}\left[F_\rho\left(\theta + \frac{\pi}{2} e_j\right) - F_\rho\left(\theta - \frac{\pi}{2} e_j\right)\right]. \tag{52}$$

Indeed, with the other angles fixed, $H_j^2 = I$ implies that $F_\rho$ is a linear combination of $1$, $\cos\theta_j$, and $\sin\theta_j$. Estimating all derivatives separately therefore requires $2p$ settings per input. All shifts act on the original gate parameters; shared parameters or other generator spectra would require a different shift rule.

**Proposition B.7** (Gradient reconstruction from linear laws)**.** *Assume the hypotheses of Proposition B.6. For an ordered set $J$ of parameter coordinates, write $d_J$ for the subvector of $d \in \mathbb{R}^p$ with entries indexed by $J$, and $Q_J$ for the submatrix of $Q$ formed by the rows indexed by $J$.*

(i) (Exact reconstruction.) *If $r > 0$, there exists an ordered set $J$ of $r$ coordinates for which $Q_J$ is invertible. For any such $J$, the matrix $T = QQ_J^{-1} \in \mathbb{R}^{p\times r}$ satisfies, for every input $\rho$ and every dataset $\mathcal{D}$,*

$$\nabla F_\rho = T(\nabla F_\rho)_J, \qquad \nabla L_{\mathcal{D}} = T(\nabla L_{\mathcal{D}})_J. \tag{53}$$

*Consequently, the measurement gradient of each input can be reconstructed from the $r$ derivatives indexed by $J$, that is, from $2r$ shifted settings, and using the reconstructed gradients in place of the exact ones leaves the gradient flow and the exact-gradient SGD iterates unchanged.*

(ii) (Estimated derivatives.) *If $\widehat{d}_J$ is an unbiased estimator of $(\nabla F_\rho)_J$, then $T\widehat{d}_J$ is an unbiased estimator of $\nabla F_\rho$ with values in $K^\perp$. Moreover, every SGD update formed from scalar multiples and minibatch sums of such reconstructed estimates preserves the laws $h_1, \dots, h_q$ for every realization, provided the iterates remain in $\Theta$.*

(iii) *If $r = 0$, no derivative measurements are needed.*

*Proof.* (i) Since $Q$ has full column rank $r$, some $r$ of its rows are linearly independent; the indices of such rows form an ordered set $J$ with $Q_J$ invertible. As shown in the proof of Proposition B.6, every measurement gradient and every loss gradient lies in $K^\perp = \operatorname{im} Q$. Hence any such gradient $d$ can be written as $d = Qc$ for some $c \in \mathbb{R}^r$, and therefore $d_J = Q_J c$ and

$$d = Qc = QQ_J^{-1}d_J = Td_J.$$

Applying this to $d = \nabla F_\rho$ and $d = \nabla L_{\mathcal{D}}$ proves (53). Each derivative in $(\nabla F_\rho)_J$ is obtained from two shifted settings by (52), so $2r$ settings suffice. Since the reconstructed gradient coincides with the exact one, substituting it into the gradient flow or into the SGD updates leaves the trajectories and iterates unchanged.

(ii) If $\mathbb{E}\widehat{d}_J = (\nabla F_\rho)_J$, then $\mathbb{E}(T\widehat{d}_J) = T(\nabla F_\rho)_J = \nabla F_\rho$ because $T$ is fixed, so $T\widehat{d}_J$ is unbiased. Every realization of $T\widehat{d}_J$ lies in $\operatorname{im} T \subseteq \operatorname{im} Q = K^\perp$, so for every $v \in K$ and every step size $\eta$,

$$v^\top(\theta - \eta T\widehat{d}_J) = v^\top\theta.$$

Thus each update preserves $h_a(\theta) = v_a^\top\theta$, $a = 1, \dots, q$. Scalar multiples and minibatch sums of vectors in $K^\perp$ remain in $K^\perp$, so the same conclusion holds for updates formed from them, as long as the iterates remain in $\Theta$.

(iii) If $r = 0$, then $K^\perp = \{0\}$, so all measurement and loss gradients vanish and no derivative measurements are needed. □

Proposition B.7(ii) concerns the measurement gradient $\nabla F_\rho$ only. The gradient of a supervised loss also involves the scalar factor $\partial l(F_\rho(\theta), y)$, and when this factor is itself estimated from measurements, the product of two unbiased estimates need not be unbiased. The squared loss, for which an explicit analysis is possible, is treated separately below.

The number $r = p - q$ in Proposition B.7 cannot be improved by any other fixed linear reconstruction. Suppose that the laws $h_1, \dots, h_q$ span all global linear laws, so that

$$K = \bigcap_{\theta\in\Theta} \ker \mathrm{d}\Phi_\theta.$$

Let $V \subseteq \mathbb{R}^p$ be the real span of the measurement gradients $\nabla F_\rho(\theta)$ over all inputs $\rho$ and all parameters $\theta$. For $v \in \mathbb{R}^p$, the identity $v^\top \nabla F_\rho(\theta) = \operatorname{tr}(\rho\, \mathrm{d}\Phi_\theta(v))$ shows that $v \perp V$ if and only if $\operatorname{tr}(\rho\, \mathrm{d}\Phi_\theta(v)) = 0$ for all $\rho$ and $\theta$; since density matrices span $\operatorname{Herm}(d)$, this holds if and only if $\mathrm{d}\Phi_\theta(v) = 0$ for all $\theta$, that is, $v \in K$. Hence $V^\perp = K$ and $V = K^\perp$. Consequently, every fixed linear subspace that contains all measurement gradients contains $K^\perp$ and therefore has dimension at least $p - q$. This lower bound applies to fixed reconstruction ranges only; it says nothing about gradient-estimation methods that are not of this form.

*Finite-shot reconstruction error.* We assume ideal projective measurements of $M_0$, with independent shots. Since $M_0^2 = I$, a single shot at parameters $\vartheta$ returns $+1$ or $-1$, with probabilities

$(1 + F_\rho(\vartheta))/2$ and $(1 - F_\rho(\vartheta))/2$, respectively, so that its expectation is $F_\rho(\vartheta)$. All expectations below are taken over the shots only, with the input $\rho$ and the parameters $\theta$ held fixed; in particular, the coordinate set $J$ and the reconstruction matrix $T$, which may depend on $\theta$, are treated as deterministic. If the device is subject to hardware noise, the measured expectation values correspond to a noisy measurement map, and the reconstruction identities must then be verified for that map.

**Lemma B.8** (Error of a reconstructed measurement gradient)**.** *Let $d = \nabla F_\rho(\theta)$, let $J$ be an ordered set of $m \geq 1$ parameter coordinates, and let $T \in \mathbb{R}^{p\times m}$ be a matrix such that $d = Td_J$. Index the columns of $T$ by the elements of $J$, in their given order, and write $T_{:,j}$ for the column associated with $j \in J$. For each $j \in J$, define the exact shifted expectations, their single-shot variances, and the column weights*

$$F_{j,\pm} = F_\rho(\theta \pm \tfrac{\pi}{2}e_j), \qquad v_{j,\pm} = 1 - F_{j,\pm}^2, \qquad w_j = \|T_{:,j}\|^2.$$

*For each $j \in J$ and each sign, let $\widehat{F}_{j,\pm}$ be the sample mean of $N_{j,\pm} \geq 1$ shots taken at the shifted parameters $\theta \pm \frac{\pi}{2}e_j$, and assume that the $2m$ sample means are mutually independent. Let $\widehat{F}_{J,\pm} \in \mathbb{R}^m$ be the vectors with entries $\widehat{F}_{j,\pm}$, $j \in J$, and define the estimated selected derivatives, the reconstructed gradient, and the total number of shots by*

$$\widehat{d}_J = \tfrac{1}{2}(\widehat{F}_{J,+} - \widehat{F}_{J,-}), \qquad \widehat{d} = T\widehat{d}_J, \qquad S_{\mathrm{d}} = \sum_{j\in J}(N_{j,+} + N_{j,-}).$$

*Then $\widehat{d}$ is an unbiased estimator of $d$, that is, $\mathbb{E}\widehat{d} = d$, and its mean-square error is*

$$V_{\mathrm{d}} := \mathbb{E}\|\widehat{d} - d\|^2 = \frac{1}{4}\sum_{j\in J} w_j \left(\frac{v_{j,+}}{N_{j,+}} + \frac{v_{j,-}}{N_{j,-}}\right). \tag{54}$$

*In particular, if all shot counts are equal, $N_{j,\pm} = N_{\mathrm{sh}}$, then $S_{\mathrm{d}} = 2mN_{\mathrm{sh}}$ and*

$$V_{\mathrm{d}} \leq \frac{\|T\|_F^2}{2N_{\mathrm{sh}}} = \frac{m\|T\|_F^2}{S_{\mathrm{d}}}. \tag{55}$$

*Proof.* Each shot at the shifted parameters $\theta \pm \frac{\pi}{2}e_j$ takes the values $\pm 1$ with mean $F_{j,\pm}$ and variance $1 - F_{j,\pm}^2 = v_{j,\pm}$. Hence the sample mean $\widehat{F}_{j,\pm}$ of $N_{j,\pm}$ such shots satisfies

$$\mathbb{E}\widehat{F}_{j,\pm} = F_{j,\pm}, \qquad \mathrm{Var}(\widehat{F}_{j,\pm}) = \frac{v_{j,\pm}}{N_{j,\pm}}.$$

By the parameter-shift rule (52), the $j$-th entry of $\widehat{d}_J$ has expectation $\frac{1}{2}(F_{j,+} - F_{j,-}) = \partial_j F_\rho(\theta)$, so $\mathbb{E}\widehat{d}_J = d_J$ and, since $T$ is deterministic, $\mathbb{E}\widehat{d} = Td_J = d$.

For the error, write $\widehat{d} - d = T(\widehat{d}_J - d_J)$. The entries of $\widehat{d}_J$ are independent, because each is a function of a distinct pair of the mutually independent sample means, and the $j$-th entry has variance

$$\mathrm{Var}\big(\tfrac{1}{2}(\widehat{F}_{j,+} - \widehat{F}_{j,-})\big) = \frac{1}{4}\left(\frac{v_{j,+}}{N_{j,+}} + \frac{v_{j,-}}{N_{j,-}}\right).$$

Thus $\mathrm{Cov}(\widehat{d}_J)$ is diagonal with these entries, and

$$V_{\mathrm{d}} = \mathbb{E}\|T(\widehat{d}_J - d_J)\|^2 = \mathrm{tr}\big(T\,\mathrm{Cov}(\widehat{d}_J)T^\top\big) = \sum_{j\in J}\|T_{:,j}\|^2\,\mathrm{Var}\big((\widehat{d}_J)_j\big),$$

which is (54) since $\|T_{:,j}\|^2 = w_j$.

If $N_{j,\pm} = N_{\mathrm{sh}}$ for all $j$, then $S_{\mathrm{d}} = 2mN_{\mathrm{sh}}$, and since $v_{j,\pm} \leq 1$, each bracket in (54) is at most $2/N_{\mathrm{sh}}$. Therefore

$$V_{\mathrm{d}} \leq \frac{1}{2N_{\mathrm{sh}}}\sum_{j\in J} w_j = \frac{\|T\|_F^2}{2N_{\mathrm{sh}}} = \frac{m\|T\|_F^2}{S_{\mathrm{d}}},$$

using $\sum_{j\in J} w_j = \sum_{j\in J}\|T_{:,j}\|^2 = \|T\|_F^2$. This proves (55). □

Lemma B.8 assumes $m \geq 1$. If $m = 0$, no derivative is selected; then $d = Td_J$ forces $d = 0$, and no derivative measurements are needed at all. For $m \geq 1$, the bound (55) shows that, for a fixed total budget $S_{\mathrm{d}}$, the error grows with $\|T\|_F^2$. This factor measures how much the reconstruction amplifies the measurement noise. A reconstruction that uses fewer settings (smaller $m$) but has a larger $\|T\|_F^2$ may therefore require more shots, not fewer, to reach a prescribed error.

Formula (54) also indicates how a given budget should be distributed among the settings. Suppose the variances $v_{j,\pm}$ are known, and relax the problem by allowing the shot counts $N_{j,\pm}$ to be arbitrary positive real numbers, without the requirement $N_{j,\pm} \geq 1$. Minimizing $V_{\mathrm{d}}$ subject to $\sum_{j\in J}(N_{j,+} + N_{j,-}) = S_{\mathrm{d}}$ gives

$$\inf_{\substack{N_{j,\pm}>0 \\ \sum_{j\in J}(N_{j,+}+N_{j,-})=S_{\mathrm{d}}}} V_{\mathrm{d}} = \frac{1}{4S_{\mathrm{d}}}\left[\sum_{j\in J}\left(\sqrt{w_j v_{j,+}} + \sqrt{w_j v_{j,-}}\right)\right]^2,$$

which follows from the Cauchy–Schwarz inequality. When all coefficients $w_j v_{j,\pm}$ are positive, the infimum is attained by choosing $N_{j,\pm}$ proportional to $\sqrt{w_j v_{j,\pm}}$. When some coefficients are zero, the corresponding shot counts may tend to zero in the relaxed problem, and if all coefficients are zero the error is zero regardless of the allocation. This relaxed optimum is only a benchmark: in practice the shot counts must be integers, and the variances $v_{j,\pm}$ are unknown in advance and may have to be estimated from additional pilot shots.

*Stable reconstruction for Clifford circuits.* Proposition B.7 reconstructs the gradient from $p - q$ derivatives, using only the global linear laws. For the circuits classified in Theorem 4.6, the kernel of the measurement differential at the current parameter $\theta$ is larger than $K$ on the regular set $\Omega_{\mathrm{C}}$: by (45), it contains one additional direction $k_{X,j}$ or $k_{Y,j}$ at each $S_X$ or $S_Y$ site. Each such direction is a linear relation among the three first-layer derivatives of that site, valid for every input, so the gradient can be reconstructed from only $r_0 = p - q - \Delta$ derivatives. Unlike the global laws, these relations depend on $\theta$, so the reconstruction matrix must be recomputed at each step, and its noise amplification must be controlled. The next proposition shows that both can be done.

**Proposition B.9** (Stable reconstruction from the instantaneous kernel)**.** *Assume the hypotheses of Theorem 4.6, and fix $\theta \in \Omega_{\mathrm{C}}$, where $\Omega_{\mathrm{C}}$ is defined in* (17)*. Let $r_0 = r_0(\theta)$ and $r_{\mathrm{Lie}} = r_{\mathrm{Lie}}(\theta)$ be the measurement-Jacobian and Lie-closure ranks at $\theta$, and let $\Delta = r_{\mathrm{Lie}} - r_0$.*

(i) (Exact reconstruction.) *There exist an ordered set $J_\theta$ of $r_0$ parameter coordinates and a matrix $T_\theta \in \mathbb{R}^{p\times r_0}$ such that*

$$\nabla F_\rho(\theta) = T_\theta\big(\nabla F_\rho(\theta)\big)_{J_\theta} \qquad \text{for every input } \rho,$$

*and*

$$\|T_\theta\|_{\mathrm{op}}^2 \leq 3, \qquad \|T_\theta\|_F^2 \leq r_0 + 2\Delta.$$

*Once the Pauli classification of Theorem 4.6 is available, computing $J_\theta$ and $T_\theta$ and applying $T_\theta$ require $O(n)$ classical operations per step.*

(ii) (Finite-shot error.) *Let $\widehat{\nabla F_\rho}$ be the estimator obtained by applying $T_\theta$ to the parameter-shift estimates of the $r_0$ selected derivatives, each shifted setting being measured with $N_{\mathrm{sh}} \geq 1$ independent shots, so that the total number of shots is $S_{\mathrm{d}} = 2r_0 N_{\mathrm{sh}}$. Then $\widehat{\nabla F_\rho}$ is unbiased and*

$$\mathbb{E}\|\widehat{\nabla F_\rho} - \nabla F_\rho\|^2 \leq \frac{r_0(r_0+2\Delta)}{S_{\mathrm{d}}} = \frac{r_{\mathrm{Lie}}^2 - \Delta^2}{S_{\mathrm{d}}}. \tag{56}$$

(iii) (Law preservation.) *Every SGD update formed from scalar multiples and minibatch sums of $\widehat{\nabla F_\rho}$ preserves the coordinate laws of Theorem 4.6 for every realization, provided the iterates remain in $\Theta$.*

*Proof.* (i) We construct $J_\theta$ and $T_\theta$ block by block, following the decomposition of the parameters into the last-layer block and the first-layer blocks of the individual qubits. Let $K$ be the span of the gradients of the coordinate laws in (13). By (14) and the kernel decomposition (45), every measurement gradient $\nabla F_\rho(\theta)$ lies in

$$\mathcal{W}_Q(\theta) = (\ker \mathrm{d}\Phi_\theta)^\perp = K^\perp \cap \bigcap_{j\in S_X} k_{X,j}(\theta)^\perp \cap \bigcap_{j\in S_Y} k_{Y,j}(\theta)^\perp.$$

The condition $\nabla F_\rho(\theta) \in K^\perp$ means that all frozen derivatives vanish; we therefore set these entries of the reconstruction to zero. The remaining conditions are one linear relation per $S_X$ or $S_Y$ site, each involving only the three first-layer derivatives of that site.

For the blocks without such a relation, we measure all active derivatives and reconstruct them by the identity: the two active last-layer derivatives (with respect to $\phi_{2,k}$ and $\beta_{2,k}$), the two active derivatives at each $S_Z$ site (with respect to $\phi_{1,j}$ and $\beta_{1,j}$), and all three derivatives at each $S_3$ site.

Now consider an $S_X$ or $S_Y$ site $j$, and write $(\phi, \beta, \omega)$ for its first-layer angles and $d_{\mathrm{loc}} \in \mathbb{R}^3$ for the corresponding three entries of $\nabla F_\rho(\theta)$. The coefficient vectors of the kernel fields (43)–(44), in the coordinate order $(\phi, \beta, \omega)$, are

$$\nu_X = (\cos\omega, -\sin\omega\sin\beta, -\cos\omega\cos\beta)^\top,$$
$$\nu_Y = (\sin\omega, \cos\omega\sin\beta, -\sin\omega\cos\beta)^\top.$$

Let $\nu$ be $\nu_X$ or $\nu_Y$ according to the site type. The condition $\nabla F_\rho(\theta) \perp k_{\cdot,j}(\theta)$ reads $\nu^\top d_{\mathrm{loc}} = 0$, and it holds for every input $\rho$. On $\Omega_{\mathrm{C}}$ we have $\nu \neq 0$, since $\|\nu_X\|^2, \|\nu_Y\|^2$ are bounded below by the squared entry of largest modulus, which is positive. Choose an index $a \in \{1,2,3\}$ with $|\nu_a| = \max_b |\nu_b|$, resolving ties by a fixed rule; then $\nu_a \neq 0$ and $|\nu_b/\nu_a| \leq 1$ for $b \neq a$. We measure the two derivatives $(d_{\mathrm{loc}})_b$, $b \neq a$, and recover the third from the relation:

$$(d_{\mathrm{loc}})_a = -\sum_{b\neq a} \frac{\nu_b}{\nu_a}(d_{\mathrm{loc}})_b.$$

Set $u = (\nu_b/\nu_a)_{b\neq a} \in \mathbb{R}^2$. Ordering the recovered coordinate last, the local reconstruction matrix is

$$T_{\mathrm{loc}} = \begin{pmatrix} I_2 \\ -u^\top \end{pmatrix} \in \mathbb{R}^{3\times 2},$$

and since $T_{\mathrm{loc}}^\top T_{\mathrm{loc}} = I_2 + uu^\top$,

$$\|T_{\mathrm{loc}}\|_{\mathrm{op}}^2 = 1 + \|u\|^2 \leq 3, \qquad \|T_{\mathrm{loc}}\|_F^2 = 2 + \|u\|^2 \leq 4.$$

Let $J_\theta$ be the ordered set of all measured coordinates, and let $T_\theta$ be the block-diagonal matrix (after a permutation of rows) whose blocks are the identity blocks above and the matrices $T_{\mathrm{loc}}$ of the $S_X$ and $S_Y$ sites, with zero rows at the frozen coordinates. By construction $\nabla F_\rho(\theta) = T_\theta(\nabla F_\rho(\theta))_{J_\theta}$ for every $\rho$. The number of measured coordinates is

$$|J_\theta| = 2 + 2(n_X + n_Y + n_Z) + 3n_3 = r_0,$$

by Theorem 4.6. Since the blocks act on disjoint sets of rows and columns, $\|T_\theta\|_{\mathrm{op}}^2$ is the maximum of the squared operator norms of the blocks, hence at most 3, and $\|T_\theta\|_F^2$ is the sum of the squared Frobenius norms of the blocks. Each identity block contributes its number of columns, and each of the $\Delta = n_X + n_Y$ blocks $T_{\mathrm{loc}}$ contributes at most $4 = 2 + 2$, that is, at most two more than its number of columns. Summing over all blocks gives $\|T_\theta\|_F^2 \leq r_0 + 2\Delta$.

For the computational cost, once the site types are known, each block requires a constant number of operations to select $a$ and to form $u$, and applying $T_\theta$ costs a constant number of operations per block. Since there are at most $n+1$ blocks, the total is $O(n)$ per step. (The site types themselves are obtained once: a decomposition of $E_1$ into $g(E_1)$ Clifford gates of bounded locality requires $O(g(E_1))$ Pauli updates and $O(n)$ inspections of local factors.)

(ii) Apply Lemma B.8 with $J = J_\theta$, $T = T_\theta$, $m = r_0$, and $N_{j,\pm} = N_{\mathrm{sh}}$. Conditional on $\theta$, both $J_\theta$ and $T_\theta$ are deterministic, so the lemma applies and gives that $\widehat{\nabla F_\rho} = T_\theta \widehat{d}_{J_\theta}$ is unbiased with

$$\mathbb{E}\|\widehat{\nabla F_\rho} - \nabla F_\rho\|^2 \leq \frac{r_0\|T_\theta\|_F^2}{S_{\mathrm{d}}} \leq \frac{r_0(r_0+2\Delta)}{S_{\mathrm{d}}}, \qquad S_{\mathrm{d}} = 2r_0 N_{\mathrm{sh}}.$$

The final equality in (56) follows from $r_{\mathrm{Lie}} = r_0 + \Delta$, since $r_0(r_0+2\Delta) = (r_{\mathrm{Lie}} - \Delta)(r_{\mathrm{Lie}} + \Delta)$.

(iii) Every realization of $\widehat{\nabla F_\rho}$ lies in $\operatorname{im} T_\theta$, and by construction $\operatorname{im} T_\theta = \mathcal{W}_Q(\theta) \subseteq K^\perp$; in particular, its entries at the frozen coordinates are zero. Scalar multiples and minibatch sums of such vectors remain in $K^\perp$. Hence, exactly as in the proof of Proposition B.7(ii), every SGD update formed from them leaves each coordinate law unchanged, as long as the iterates remain in $\Theta$. □

Proposition B.9 applies only on the regular set $\Omega_{\mathrm{C}}$. At a parameter $\theta \notin \Omega_{\mathrm{C}}$, one can fall back on Proposition B.7, which uses the global coordinate laws alone and requires $2(p-q)$ settings. Here $p-q$ is the value of the Lie-closure rank $r_{\mathrm{Lie}}$ on $\Omega_{\mathrm{C}}$; at the singular point $\theta$ itself the pointwise Lie-closure rank may be smaller, but the reconstruction remains valid because the coordinate laws hold globally. More generally, at any parameter $\theta$ at which a basis of $(\ker \mathrm{d}\Phi_\theta)^\perp$ is known, the row-selection argument in the proof of Proposition B.7 yields a reconstruction of $\nabla F_\rho(\theta)$ from $r_0(\theta)$ coordinate derivatives. What the regular Clifford case adds to this general fact is an explicit and efficient choice of the selected coordinates together with the norm bounds of Proposition B.9.

*Examples and measurement counts.* We now make the reductions explicit for the circuits studied in Section 4. Consider first the Clifford class of Theorem 4.6, for which $p = 6n$ and, on $\Omega_{\mathrm{C}}$, $p - q = r_{\mathrm{Lie}}$. Three procedures are available for estimating the measurement gradient of one input: estimating every derivative separately, which requires $2p$ settings; using the global linear laws as in Proposition B.7, which requires $2r_{\mathrm{Lie}}$ settings; and using the instantaneous kernel as in Proposition B.9, which requires $2r_0$ settings. The third procedure therefore saves $2\Delta = 2(r_{\mathrm{Lie}} - r_0)$ settings beyond the second. For the two families of Corollary 4.7, the counts are

| Circuit family | Optimization variables | Derivative settings |
|---|---|---|
| CNOT ring | $2n+4$ | $12n \to 4n+8 \to 4n+6$ |
| Maximal-gap family | $3n+2$ | $12n \to 6n+4 \to 4n+6$ |

where the three numbers in the last column correspond to the three procedures in order. The variable counts in the second column follow from the global laws via Proposition B.6 and hold everywhere, whereas the final setting counts in the third column hold only for $\theta \in \Omega_{\mathrm{C}}$.

For the two-qubit, two-layer Strong circuit of Example 4.5, the five laws reduce the twelve optimization variables to seven. Its derivative settings decrease as $24 \to 16 \to 14 \to 12$: the first reduction removes the four frozen derivatives; the second uses the relation $\partial_{\phi_{2,1}} F_\rho = \partial_{\omega_{1,2}} F_\rho$, so that only one of the two derivatives needs to be measured; and the third uses the additional kernel direction in (35), which is available on $\Omega_{\mathrm{S}}$. This circuit is not covered by Theorem 4.6, so the last reduction relies on the separate kernel calculation in Appendix B.3.4.

Settings and shots must be distinguished. When the $q$ laws are frozen coordinates, the reconstruction only inserts zeros at those coordinates, so the reconstruction matrix $T$ is a coordinate injection and $\|T\|_F^2 = p - q$. If the same number of shots is kept for each retained setting, the retained derivative estimates have exactly the same distribution as in the baseline that measures all derivatives, and the total number of derivative shots decreases by the fraction $q/p$. For the CNOT ring, this fraction equals $2/3 - 2/(3n)$; at $n = 8$, the derivative cost decreases from $96N_{\mathrm{sh}}$ to $40N_{\mathrm{sh}}$ shots. For a general linear relation, however, reusing an estimate may amplify its noise. As an example, suppose a model depends on two parameters $a$ and $b$ only through $s = a+b$, so that $\partial_a F_\rho = \partial_b F_\rho$. Measuring the two derivatives independently, each with variance $v$, gives variance $2v$ for their sum. Measuring only one of them and reusing the estimate for both gives variance $4v$ for the sum. Restoring the variance $2v$ requires doubling the shots of the reused estimate, so the number of settings is halved but the number of shots is unchanged. A reduction in the number of settings alone therefore does not guarantee a reduction in the number of shots at a fixed error.

*Prediction and minibatch costs.* The results above concern the measurement gradient $\nabla F_\rho$. The gradient of a supervised loss $l(F_\rho(\theta), y)$ is the product $\partial l(F_\rho(\theta), y) \nabla F_\rho(\theta)$, so it also requires the scalar factor $\partial l(F_\rho(\theta), y)$, which in turn depends on the prediction $F_\rho(\theta)$. When this prediction is estimated from measurements as well, the resulting loss-gradient estimate is a product of two random quantities. For the squared loss, the factor is linear in the prediction, and an independent prediction estimate yields an unbiased loss-gradient estimator whose error can be computed exactly.

**Lemma B.10** (Squared-loss gradient estimation)**.** *Fix an input $\rho$, a label $y$, and parameters $\theta$, and write $F = F_\rho(\theta)$ and $d = \nabla F_\rho(\theta)$. Consider the loss $l(F, y) = \frac{1}{2}(F-y)^2$, whose gradient with respect to $\theta$ is $(F-y)d$. Let $\widehat{d}$ be an unbiased estimator of $d$ with mean-square error $V_{\mathrm{d}} = \mathbb{E}\|\widehat{d} - d\|^2$, for instance the estimator of Lemma B.8. Let $\widehat{F}_0$ be the sample mean of $N_{\mathrm{pred}} \geq 1$ shots taken at the unshifted parameters $\theta$, independent of the shots used for $\widehat{d}$, so that $\mathbb{E}\widehat{F}_0 = F$ and $\mathrm{Var}(\widehat{F}_0) = a_0$ with*

$$a_0 = \frac{1 - F^2}{N_{\mathrm{pred}}}.$$

*Define* $\widehat{g} = (\widehat{F}_0 - y)\widehat{d}$*. Then* $\widehat{g}$ *is an unbiased estimator of the loss gradient* $(F - y)d$*, and its mean-square error is*

$$\mathbb{E}\|\widehat{g} - (F-y)d\|^2 = \big((F-y)^2 + a_0\big)V_{\mathrm{d}} + a_0\|d\|^2. \tag{57}$$

*Proof.* Since $\widehat{F}_0$ and $\widehat{d}$ are independent, $\mathbb{E}\widehat{g} = \mathbb{E}(\widehat{F}_0 - y)\,\mathbb{E}\widehat{d} = (F-y)d$, which proves unbiasedness. For the error, set $\varepsilon_0 = \widehat{F}_0 - F$ and $e = \widehat{d} - d$, so that $\widehat{F}_0 - y = (F-y) + \varepsilon_0$ and $\widehat{d} = d + e$. Expanding the product,

$$\widehat{g} - (F-y)d = (F-y)e + \varepsilon_0 d + \varepsilon_0 e.$$

The scalar $\varepsilon_0$ and the vector $e$ are independent with $\mathbb{E}\varepsilon_0 = 0$, $\mathbb{E}e = 0$, $\mathbb{E}\varepsilon_0^2 = a_0$, and $\mathbb{E}\|e\|^2 = V_{\mathrm{d}}$. Taking the squared norm and the expectation, every cross term contains a factor $\mathbb{E}\varepsilon_0 = 0$ or $\mathbb{E}e = 0$ and therefore vanishes, while the three squared terms give

$$(F-y)^2\,\mathbb{E}\|e\|^2 = (F-y)^2 V_{\mathrm{d}}, \qquad \mathbb{E}\varepsilon_0^2\,\|d\|^2 = a_0\|d\|^2, \qquad \mathbb{E}\varepsilon_0^2\,\mathbb{E}\|e\|^2 = a_0 V_{\mathrm{d}}.$$

Summing these three contributions gives (57). □

If the loss is taken without the factor $\frac{1}{2}$, its gradient is $2(F-y)d$; the corresponding estimator is $2\widehat{g}$, and its mean-square error is four times the right-hand side of (57). For other losses, the factor $\partial_1 l(F, y)$ is nonlinear in $F$, and the unbiasedness of a plug-in estimate must be examined separately. Regardless of the loss, however, any scalar multiple of a reconstructed measurement gradient still lies in $K^{\perp}$, so the linear laws are preserved.

We can now count the total measurement cost per input. If $m \geq 1$ derivatives are selected, the gradient estimate uses $2m$ shifted settings and $S_{\mathrm{d}}$ shots, and the prediction estimate uses one unshifted setting and $N_{\mathrm{pred}}$ shots; in total, $2m + 1$ settings and $S_{\mathrm{d}} + N_{\mathrm{pred}}$ shots per input, unless an independent prediction estimate is already available from elsewhere. For the eight-qubit CNOT ring with $N_{\mathrm{pred}} = N_{\mathrm{sh}}$, the frozen-coordinate reduction of the previous paragraph reduces the total from $97N_{\mathrm{sh}}$ to $41N_{\mathrm{sh}}$ shots.

For a minibatch $\{(\rho_i, y_i)\}_{i=1}^{b}$ of size $b$, the same procedure is applied to each sample, giving $b(2m+1)$ input–setting pairs. Let $g_i$ be the exact loss gradient of the $i$-th sample and $\widehat{g}_i$ its unbiased estimate. If the shots used for different samples are independent, the errors $\widehat{g}_i - g_i$ are independent with zero mean, and hence

$$\mathbb{E}\left\|\frac{1}{b}\sum_{i=1}^{b}(\widehat{g}_i - g_i)\right\|^2 = \frac{1}{b^2}\sum_{i=1}^{b}\mathbb{E}\|\widehat{g}_i - g_i\|^2.$$

This expectation is conditional on the chosen minibatch and on the current parameters; it therefore measures the shot noise of the gradient estimate and not the randomness of minibatch selection. We stress that the results of this appendix establish exact dimension reduction and control of the measurement error; they do not establish improved prediction accuracy, nor optimality among all quantum gradient estimators.

## C Broader Numerical Results

In this appendix, we validate the conservation laws for the Basic PQC and Strong PQC architectures with $n = 3$ qubits and $L = 3$ layers. The evolution of all trainable parameters under SGD for the Basic PQC and Strong PQC is reported in Figures 2 and 3, respectively.

For the three-qubit, three-layer Basic PQC, we validate the complete set of independent conservation laws, with the explicit derivation provided in Appendix B.2. For the three-qubit, three-layer Strong PQC considered here, we validate only the conservation laws identified by Theorem 4.4, which are associated with the final layer. We do not claim that these laws form a complete set.

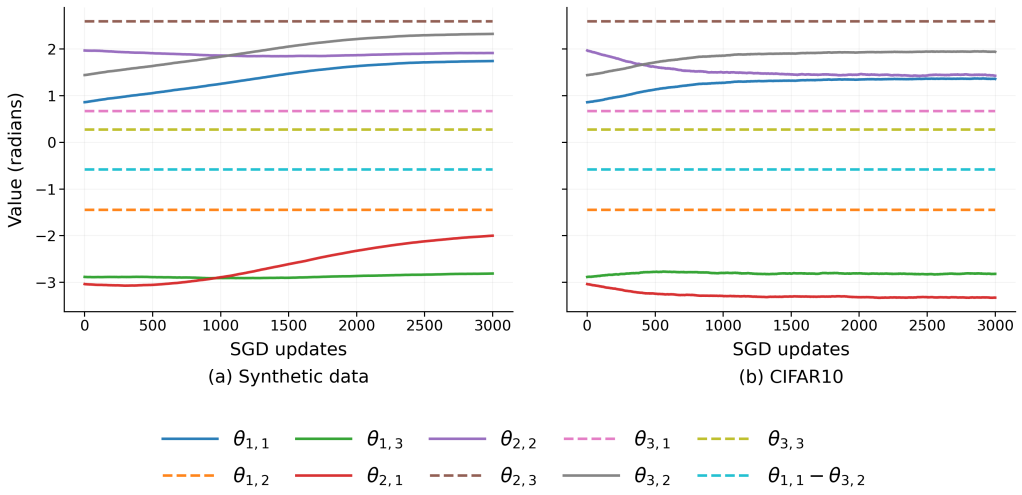



Figure 2: Evolution of the Basic PQC parameters during training for $n = L = 3$ on both synthetic and CIFAR-10 datasets. The dashed lines represent the theoretically predicted conservation laws, comprising both individually invariant parameters and conserved linear combinations of parameters.

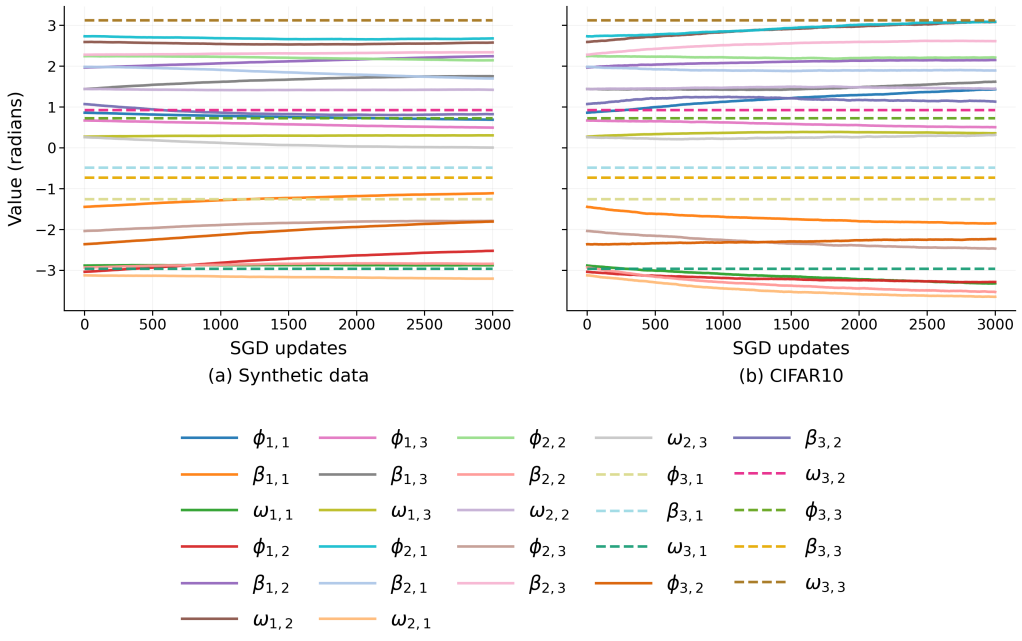



Figure 3: Evolution of the Strong PQC parameters during training for $n = L = 3$ on both synthetic and CIFAR-10 datasets. The dashed lines indicate the individually frozen parameters predicted by Theorem 4.4.