On the Principles of Differential Quantum Programming Languages

Shaopeng Zhu Shih-Han Hung Shouvanik Chakrabarti Xiaodi Wu
University of Maryland, College Park, USA

Abstract

Variational Quantum Circuits (VQCs), or the so-called quantum neural-networks, are predicted to be one of the most important near-term quantum applications, not only because of their similar promises as classical neural-networks, but also because of their feasibility on near-term noisy intermediate-size quantum (NISQ) machines. The need for gradient information in the training procedure of VQC applications has stimulated the interest and development of auto-differentiation techniques on quantum circuits. We propose the first formalization of this technique, not only in the context of quantum circuits but also for imperative quantum programs (e.g., with controls), inspired by the success of differential programming languages in classical machine learning. In particular, we overcome a few unique difficulties caused by exotic quantum features (such as quantum no-cloning) and provide a rigorous formulation of differentiation applied to bounded-loop imperative quantum programs, its code-transformation rules, as well as a sound logic to reason about their correctness.

Moreover, we have implemented our code transformation in OCaml and demonstrated the resource-efficiency of our scheme both analytically and empirically. We also conduct a case study of training a VQC instance with controls, which shows the advantage of our scheme over existing auto-differentiation on quantum circuits without controls.

Keywords quantum programming languages, differential programming languages, auto differentiation, quantum machine learning, collection of semantics

1 Introduction

Background. Recent years have witnessed the rapid development of quantum computing, with practical advances coming from both research and industry. Quantum programming is one topic that has been actively investigated. Early work on language design [22, 32, 39, 40, 44] has been followed up recently by several implementations of these languages, including Quipper [24], Scaffold [1], LIQUi|⟩ [49], Q# [46], and QWIRE [33]. Extensions of program logics have also been proposed for verification of quantum programs [3, 9, 10, 17, 27, 52, 55]. See also surveys [18, 43, 53].

With the availability of prototypes of quantum machines, especially the recent establishment of quantum supremacy [2], the research of quantum computing has entered a new stage where near-term Noisy Intermediate-Scale Quantum (NISQ) computers [37], e.g., the 53-qubit quantum machine from Google [2] and IBM [20], becomes the important platform for demonstrating quantum applications. Variational quantum circuits (VQCs) [15, 16, 35], or the so-called *quantum neural networks*, are predicted to be one of the most important applications on NISQ machines. This is not only because VQCs bear a lot of similar promises like classical neural networks as well as potential quantum speed-ups from the perspective of machine learning (e.g., see the survey [8]), but also because VQC is, if not the only, one of the few candidates that can be implemented on NISQ machines. Because of this, a lot of study has already been devoted to the design, analysis, and small-scale implementation of VQC (e.g., see the survey [6]).

Typical VQC applications replace classical neural networks, which are just parameterized classical circuits, by quantum circuits with *classically parameterized unitary gates*. Namely, one will have a "quantum" mapping from input to output replacing classical mapping in machine learning applications. An important component of these applications is a training procedure which optimizes a *loss function* the now depends on the *read-outs* and the *parameters* of VQCs.

It is well-known that gradient-based approaches are the best for the training procedure. However, computing the gradients of loss functions that depend on the read-outs of quantum circuits has a similar complexity of simulating quantum circuits, which is infeasible for classical computation. Thus, the ability of evaluating these "quantum" gradients efficiently by quantum computation is critical for the scalability of VQC applications, e.g., on the recent 53-qubit machines.

Fortunately, analytical formulas of gradients used in VQCs have been studied by [16, 19, 29, 41, 42]. In particular, Schuld et al. [41] proposed the so-called *phase-shift rule* that uses two quantum circuits to compute the partial derivative respective to one parameter for quantum circuits. One of the very successful tools for quantum machine learning, called PennyLane [7], implemented the phase-shift rule to achieve *auto differentiation* (AD) for the read-outs of quantum circuits. It also integrated automatic differentiation from established machine learning libraries such as TensorFlow or PyTorch for any additional classical calculation in the training procedure. However, none of these research was conducted from the perspective of programming languages and no rigorous foundation or principles have been formalized.

Motivations. An important motivation of this paper is to provide a *rigorous formalization* of the auto-differentiation

Draft paper, November 22, 2019,
2020.

technique applied to quantum circuits. In particular, we will provide a formal formulation of quantum programs, their semantics, and the meaning of differentiation on them. We will also study the code-transformation rules for auto-differentiation and prove their correctness.

As we will highlight below, research on the formalization will encounter many new challenges that have not been considered and addressed by existing results [16, 19, 29, 41, 42]. Consider one of the basic requirements, e.g., compositionality. As we will show, differentiating the composition of quantum programs will necessarily involve running multiple quantum programs on copies of initial quantum states. How to represent the collection of quantum programs succinctly and also bound the number of required copies is a totally new question. Among our techniques to address this question, we also need to change the previously proposed construct, e.g., the phase-shift rule [41], to something else.

Moreover, we want to go beyond the restriction to quantum circuits in [16, 19, 29, 41, 42]. We are inspired by classical machine learning examples that demonstrate the advantage of neural-networks with program features (e.g. controls) over the plain ones (e.g., classical circuits) [23, 25], which is also the major motivation of promoting the the paradigm shift from deep learning towards *differential programming* by prominent machine learning researchers.

Augmenting VQCs with controls is not only feasible on NISQ machines (at least for simple ones), but also now a logical step for the study of their applications in machine learning. (We indeed conduct one case study in Section 8.)

Thus, we are inspired to investigate the *principles of differential quantum imperative languages* beyond circuits.

Research challenges & Solutions. We will rely on a little notation that should be self-explanatory. Please refer to a detailed preliminary on quantum information in Section 2.

Let us start with a simple classical program

$$\mathrm{MUL} \equiv v_3 = v_1 \times v_2, \tag{1.1}$$

where v_3 is the product of v_1 and v_2. Consider the differentiation with respect to θ, we have

$$\frac{\partial}{\partial\theta}(\mathrm{MUL}) \equiv \quad v_3 = v_1 \times v_2; \tag{1.2}$$

$$\dot{v}_3 = \dot{v}_1 \times v_2 + v_1 \times \dot{v}_2, \tag{1.3}$$

where MUL keeps track of variables v_1, v_2, v_3 and their derivatives $\dot{v}_1, \dot{v}_2, \dot{v}_3$ at the same time. One simple yet important observation is that classical variables v_1, v_2, v_3 are real-valued and can be naturally differentiated, while quantum variables are quantum states represented by matrices.

What are hence reasonable quantities to compute derivatives about quantumly? One natural choice from the principles of quantum mechanics is the (classical) read-outs of quantum systems through measurements, which we formulate as the *observable semantics* of quantum programs. This natural choice also serves the purpose of gradients computation of loss functions in quantum machine learning applications, which are typically phrased in terms of these read-outs. We also directly model the parameterization of quantum programs after VQCs, i.e., each unitary gate becomes classically parameterized. The above modeling of quantum programs is very different from classical ones. It is also unclear whether any reasonable analogues of classical chain-rule and forward/backward mode can exist within quantum programs.

To model the meaning of one quantum program computing the derivative of another, we define the *differential semantics* of programs. There is a subtle quantum-unique design choice. The observable semantics of any quantum program will depend on the observable and its input state. Thus, any program computing its derivative could potentially depend on these two extra factors. We find out this potential dependence is undesirable and propose the strongest possible definition: i.e., one derivative computing program should work for *any pair* of observable and input states. We demonstrate that this strong requirement is not only achievable but also critical for the composition of auto differentiation.

We are ready to describe the technical challenges for the compositionality. Consider the following quantum program:

$$\mathrm{QMUL} \equiv U_1(\theta); U_2(\theta), \tag{1.4}$$

which performs $U_1(\theta)$ and $U_2(\theta)$ gates sequentially. Note that gate application is matrix multiplication in quantum. Roughly speaking, if the product rule of differentiation (as exhibited in (1.3)) remains in the quantum setting, at least symbolically, then one should expect $\frac{\partial}{\partial\theta}(\mathrm{QMUL})$ contains

$$\frac{\partial}{\partial\theta}(U_1(\theta)); U_2(\theta) \text{ and } U_1(\theta); \frac{\partial}{\partial\theta}(U_2(\theta)) \tag{1.5}$$

two different parts as sub-programs similarly in (1.3).

However, we cannot run $\frac{\partial}{\partial\theta}(U_1(\theta)); U_2$ and $U_1(\theta); \frac{\partial}{\partial\theta}(U_2(\theta))$ together due to the quantum *no-cloning* theorem [51]. This is because they share the same initial state and we cannot clone two copies of it and maintain the correct correlation among different parts. Note that this is not an issue classically as we can store all $v_i, \dot{v}_i$ at the same time as in (1.3). As a result, quantum differentiation needs to run *multiple* (sub-)programs on *multiple* copies of the initial state.

This poses a unique challenge for differentiation of quantum composition: (1) we hope to have a simple scheme of code transformation, ideally close-to-classical, for intuition and easy implementation of the compiler, whereas it needs to express correctly the collection of quantum programs during code transformation; (2) for the purpose of efficiency, we also want to reasonably bound the number of required copies of the initial states, which roughly refers to the number of different quantum programs in the collection.

We develop a few techniques to achieve both goals at the same time. First, we propose the so-called *additive* quantum programs as a *succinct* intermediate representation for the

collection of programs during the code transformation. Now the entire differentiation procedure will be divided into two steps: (1) all code transformations happen on additive programs and are very similar to classical ones (see Figure 4) ; (2) the collection of programs can be recovered by a compilation procedure from any additive program. Additive quantum programs are equipped with a new *sum* operation that models the multiple choices as exhibited in (1.5), which resembles a similar idea in differential lambda-calculus [12].

Second, we also design a new rule for $\frac{\partial}{\partial\theta}(U(\boldsymbol{\theta}))$ which is slightly different from [16, 19, 29, 41, 42]. The existing phase-shift rule makes use of two quantum circuits for one differentiation, which causes a lot of inconvenience in the formulation and potential trouble for efficiency. We replace it by one circuit with one additional ancilla register for each parameter. We conduct a careful resource analysis of our differentiation procedure and show the number of required copies of initial states is reasonable comparing to the classical setting. The correctness of the code transformation of composition critically relies on our design choice as well as the strong definition related to the differential semantics.

With the previous setup, we can naturally build the differentiation for controls (i.e., the quantum condition statement). Note that a general solution for classical controls is still unknown [5] due to the non-smoothness of the guard. Similar to the classical setting [36], we do not have a solution to deal with general unbounded loop. We instead provide a solution to bounded loops which can be deemed as a macro consisting of simple quantum condition statements.

Contributions. We formulate the parameterized quantum bounded while-programs with classically parameterized unitary gates modelled after VQCs [15, 30, 35] and their realistic examples on ion-trap machines , e.g. [56], in Section 3.

In Section 4, we illustrate our design of *additive* quantum programs. Specially, we add the syntax $P_1 + P_2$ to represent the either-or choice between P_1 and P_2 in (1.5). We formulate its semantics and the compilation rules that map any additive program into a collection of normal programs in the way to serve our differentiation purpose.

In Section 5, we formulate the observable and the differential semantics of quantum programs and formally define the meaning of program $\mathsf{S}'(\boldsymbol{\theta})$ computing the differential semantics of $\mathsf{S}(\boldsymbol{\theta})$ in the strongest possible sense. In particular, we require that $\mathsf{S}'(\boldsymbol{\theta})$ should work for any pair of observable and input state, which is critical for the compositionality.

In Section 6, we show that such a strong requirement is indeed achievable by demonstrating the code-transformation rules for the differentiation procedure. Thanks to the use of additive quantum programs, the code transformation is much simplified and as intuitive as classical ones. We also develop a logic with the judgement $\mathsf{S}'(\boldsymbol{\theta})|\mathsf{S}(\boldsymbol{\theta})$ that states $\mathsf{S}'(\boldsymbol{\theta})$ computes the differential semantics of $\mathsf{S}(\boldsymbol{\theta})$. We prove the soundness of our logic and use it to show the correctness of our code-transformation rules.

In Section 7, we conduct a resource analysis to further justify our design. We show that the *occurrence count* of parameters capture the extra resource required in both the classical auto-differentiation and our scheme. In this sense, our resource cost is reasonable comparing to classical.

Finally, in Section 8, we demonstrate the implementation of our code transformation in OCaml and apply it to the training of one VQC instance with controls via classical simulation. Specifically, this instance shows an advantage of controls in machine learning tasks, which implies the advantage of our scheme over previous ones that cannot handle controls. We have also empirically verified the resource-efficiency of our scheme on representative VQC instances.

Related classical work. There is an extensive study of automatic differentiation (AD) or differential programming in the classical setting (e.g., see books [11, 26]). The most relevant to us are those studies from the programming language perspective. AD has traditionally been applied to imperative programs in both the forward mode, e.g. [28, 50], and the reverse mode, e.g., [45]. The famous *backpropagation* algorithm [38] is also a special case of reserve-mode AD used to compute the gradient of a multi-layer perceptron. AD has also been recently applied to functional programs [13, 14, 34]. Motivated by the success of deep learning, there is significant recent interest to develop both the theory and the implementation of AD techniques. Please refer to the survey [4] and the keynote talk at POPL'18 [36] for more details.

2 Quantum Preliminaries

We present basic quantum preliminaries here (a summary of notation in Table 1). Details are deferred to Appendix A.

2.1 Math Preliminaries

Let n be a natural number. We refer to the complex vector space $\mathbb{C}^n$ as an *n-dimensional Hilbert space* $\mathcal{H}$. We use $|\psi\rangle$ to denote a complex vector in $\mathbb{C}^n$. The Hermitian conjugate of $|\psi\rangle \in \mathbb{C}^n$ is denoted by $\langle\psi|$. The *inner product* of $|\psi\rangle$ and $|\phi\rangle$, defined as the product of $\langle\psi|$ and $|\phi\rangle$, is denoted by $\langle\psi|\phi\rangle$. The norm of a vector $|\psi\rangle$ is denoted by $\||\psi\rangle\| = \sqrt{\langle\psi|\psi\rangle}$.

We define (linear) *operators* as linear maps between Hilbert spaces, which can be represented by matrices for finite dimension. Let A be an operator; its Hermitian conjugate is denoted by $A^\dagger$. A is *Hermitian* if $A = A^\dagger$. The *trace* of A is the sum of the entries on the main diagonal, i.e., $\mathrm{tr}(A) = \sum_i A_{ii}$. We write $\langle\psi|A|\psi\rangle$ to mean the inner product of $|\psi\rangle$ and $A|\psi\rangle$. A Hermitian operator A is *positive semidefinite* if for all vectors $|\psi\rangle \in \mathcal{H}$, $\langle\psi|A|\psi\rangle \geq 0$.

2.2 Quantum States and Operations

The state space of a *qubit* is a 2-dimensional Hilbert space. Two important orthonormal bases of a qubit system are: the

Table 1. A brief summary of notation used in this paper

Spaces	$\mathcal{H}, \mathcal{A}$	$L(\mathcal{H})$ (Linear operators)
States	(pure states)	$\lvert\psi\rangle, \lvert\phi\rangle$ (metavariables); $\lvert 0\rangle, \lvert 1\rangle, \lvert +\rangle, \lvert -\rangle$
	(density)	ρ, σ (meta); $\lvert\psi\rangle\langle\psi\rvert$
Operations	(unitaries)	$U, V, \boldsymbol{\sigma}$ (meta); $H, X, Z, X \otimes X$
	(superoperators)	$\mathcal{E}, \mathcal{F}$ (general); Φ (quantum channels)
Measurements	M	$\{M_m\}_m$ (general); $\{\lvert 0\rangle\langle 0\rvert, \lvert 1\rangle\langle 1\rvert\}$
Observables	O	$\sum_m \lambda_m \lvert\phi_m\rangle\langle\phi_m\rvert$ (general); $\lvert 0\rangle\langle 0\rvert - \lvert 1\rangle\langle 1\rvert$ (example)
Programs	(no parameters)	P, Q (meta)
	(parameterized)	$P(\boldsymbol{\theta}), S(\boldsymbol{\theta})$ (meta); $R_{\boldsymbol{\sigma}}(\theta)$(notable examples)
	(collection)	$\overline{P(\boldsymbol{\theta})}, \overline{S(\boldsymbol{\theta})}$ (meta); $\overline{\frac{\partial}{\partial\theta}(QCT_{(7,6)}(\boldsymbol{\theta}))}$ (example)
Semantics	(operational)	$\langle P, \rho\rangle \to \langle Q, \rho'\rangle$ (steps)
	(denotational)	$[\![P]\!]\rho = \rho'$ (example)
	(observable)	$[\![(O, \rho) \to P(\boldsymbol{\theta})]\!]$ (example)
Code Process	(Transform)	$\frac{\partial}{\partial\theta}(\overline{S(\boldsymbol{\theta})})$ (example)
	(Compile)	$\texttt{Compile}(\overline{S'(\boldsymbol{\theta})})$ (example)
Logic	(Judgement)	$\overline{S'(\boldsymbol{\theta})} \lvert \overline{S(\boldsymbol{\theta})}$ (example)
Count	(Non-Abort)	$\lvert\#\overline{\frac{\partial}{\partial\theta_j}(P(\boldsymbol{\theta}))}\rvert$ (example)
	(Occurrence)	$\mathrm{OC}_j(P(\boldsymbol{\theta}))$ (example)

computational basis with $|0\rangle = (1,0)^\dagger$ and $|1\rangle = (0,1)^\dagger$; the $\pm$ basis, consisting of $|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$ and $|-\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$.

A *pure* quantum state is a vector $|\psi\rangle$ with $\||\psi\rangle\| = 1$. A *mixed* state can be represented by a classical distribution over an ensemble of pure states $\{(p_i, |\psi_i\rangle)\}_i$, i.e., the system is in state $|\psi_i\rangle$ with probability p_i. One can use *density operators* to represent states: a density operator representing the ensemble $\{(p_i, |\psi_i\rangle)\}_i$ is a positive semidefinite operator $\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|$. A positive semidefinite operator ρ on $\mathcal{H}$ is said to be a *partial* density operator if $\mathrm{tr}(\rho) \leq 1$. The set of partial density operators on $\mathcal{H}$ is denoted by $\mathcal{D}(\mathcal{H})$.

Operations on quantum systems can be characterized by unitary operators. Denoting the set of linear operators on $\mathcal{H}$ as $L(\mathcal{H})$, an operator $U \in L(\mathcal{H})$ is *unitary* if $U^\dagger U = UU^\dagger = I_{\mathcal{H}}$. A unitary *evolves* a pure state $|\psi\rangle$ to $U|\psi\rangle$, or a density operator ρ to $U\rho U^\dagger$. Common unitary operators include: the *Hadamard* operator H, which transforms between the computational and the $\pm$ basis via $H|0\rangle = |+\rangle$ and $H|1\rangle = |-\rangle$; the *Pauli* X operator which performs a bit flip, i.e., $X|0\rangle = |1\rangle$ and $X|1\rangle = |0\rangle$; *Pauli* Z which performs a phase flip, i.e., $Z|0\rangle = |0\rangle$ and $Z|1\rangle = -|1\rangle$; *Pauli* Y mapping $|0\rangle$ to $i|1\rangle$ and $|1\rangle$ to $-i|0\rangle$; *CNOT* gate mapping $|00\rangle \mapsto |00\rangle, |01\rangle \mapsto |01\rangle, |10\rangle \mapsto |11\rangle, |11\rangle \mapsto |10\rangle$.

More generally, evolution of a quantum system can be characterized by an *admissible superoperator* $\mathcal{E}$, namely a *completely-positive* and *trace-non-increasing* (See Appendix A) linear map from $\mathcal{D}(\mathcal{H})$ to $\mathcal{D}(\mathcal{H}')$ for Hilbert spaces $\mathcal{H}, \mathcal{H}'$.

For every superoperator $\mathcal{E}$, there exists a set of *Kraus operators* $\{E_k\}_k$ such that $\mathcal{E}(\rho) = \sum_k E_k \rho E_k^\dagger$ for any input $\rho \in \mathcal{D}(\mathcal{H})$. The *Kraus form* of $\mathcal{E}$ is therefore $\mathcal{E} = \sum_k E_k \circ E_k^\dagger$. An identity operation refers to the superoperator $\mathcal{I}_{\mathcal{H}} = I_{\mathcal{H}} \circ I_{\mathcal{H}}$. The Schrödinger-Heisenberg *dual* of a superoperator $\mathcal{E} = \sum_k E_k \circ E_k^\dagger$, denoted by $\mathcal{E}^*$, is defined as follows: for every state $\rho \in \mathcal{D}(\mathcal{H})$ and any operator A, $\mathrm{tr}(A\mathcal{E}(\rho)) = \mathrm{tr}(\mathcal{E}^*(A)\rho)$. The Kraus form of $\mathcal{E}^*$ is $\sum_k E_k^\dagger \circ E_k$.

2.3 Quantum Measurements

Quantum *measurements* extracts classical information out of quantum systems. A quantum measurement on a system over Hilbert space $\mathcal{H}$ can be described by a set of linear operators $\{M_m\}_m$ with $\sum_m M_m^\dagger M_m = I_{\mathcal{H}}$ (identity matrix on $\mathcal{H}$). If we perform a measurement $\{M_m\}_m$ on a state ρ, the outcome m is observed with probability $p_m = \mathrm{tr}(M_m \rho M_m^\dagger)$ for each m, and the post-measurement state collapses to $M_m \rho M_m^\dagger / p_m$.

3 Parameterized Quantum Bounded While-Programs

We adopt the *bounded-loop* variant of the quantum **while**-language developed by Ying [53], and augment it by parameterizing the unitaries, as this provides sufficient expressibility for parameterized quantum operations: indeed, **abort**, **skip** and initialization behave independently of parameters, while "parameterized measurements" can be implemented with a regular measurement followed by a parameterized unitary.

From here onward, $\overline{v}$ is a finite set of variables, and $\boldsymbol{\theta}$ a length-k vector of real-valued parameters.

3.1 Syntax

Define *Var* as the set of quantum variables. We use the symbol q as a metavariable ranging over quantum variables and define a *quantum register* $\overline{q}$ to be a finite set of distinct variables. For each $q \in Var$, its state space is denoted by $\mathcal{H}_q$. The quantum register $\overline{q}$ is associated with the Hilbert space $\mathcal{H}_{\overline{q}} = \bigotimes_{q \in \overline{q}} \mathcal{H}_q$. [1] A *$T$-bounded, k-parameterized quantum **while**-program* is generated by the following syntax:

$$
\begin{aligned}
P(\boldsymbol{\theta}) \;::=\;& \textbf{abort}[\overline{q}] \mid \textbf{skip}[\overline{q}] \mid q := |0\rangle \mid \overline{q} := U(\boldsymbol{\theta})[\overline{q}] \mid \\
& P_1(\boldsymbol{\theta}); P_2(\boldsymbol{\theta}) \mid \textbf{case } M[\overline{q}] = \overline{m \to P_m(\boldsymbol{\theta})} \textbf{ end} \mid \\
& \textbf{while}^{(T)}\, M[\overline{q}] = 1 \textbf{ do } P_1(\boldsymbol{\theta}) \textbf{ done},
\end{aligned}
$$

[1] If $type(q) =$ **Bool** then $\mathcal{H}_q = \mathrm{span}\{|0\rangle, |1\rangle\}$. If $type(q) =$ **Bounded Int** then $\mathcal{H}_q$ is with basis $\{|n\rangle : n \in [-N, N]\}$ ($N \in \mathbb{Z}^+$) for some finite N. We require the Hilbert space to be finite dimensional for implementation.

where

$$\begin{aligned} &\textbf{while}^{(1)}\ M[\overline{q}] = 1\ \textbf{do}\ P_1(\boldsymbol{\theta})\ \textbf{done} \\ \equiv\ &\textbf{case}\ M[\overline{q}] = \{0 \to \textbf{skip},\ \ 1 \to P_1(\boldsymbol{\theta}); \textbf{abort}\}, \end{aligned} \tag{3.1}$$

$$\begin{aligned} &\textbf{while}^{(T\geq 2)}\ M[\overline{q}] = 1\ \textbf{do}\ P_1(\boldsymbol{\theta})\ \textbf{done} \\ \equiv\ &\textbf{case}\ M[\overline{q}] = \{0 \to \textbf{skip},\ \ 1 \to P_1(\boldsymbol{\theta}); \textbf{while}^{(T-1)}\} \end{aligned}$$

Unparameterized programs can be obtained by fixing $\boldsymbol{\theta}^* \in \mathbb{R}^k$ in some $P(\boldsymbol{\theta})$. We denote the set of variables accessible to $P(\boldsymbol{\theta})$ as $\mathrm{qVar}(P(\boldsymbol{\theta}))$; the collection of all "T-bounded while-programs $P(\boldsymbol{\theta})$ s.t. $\mathrm{qVar}(P(\boldsymbol{\theta})) = \overline{v}$" as $\textbf{q-while}^{(T)}_{\overline{v}}(\boldsymbol{\theta})$, and similarly, the unparameterized one as $\textbf{q-while}^{(T)}_{\overline{v}}$.

Now let us formally define *parameteriztion* of unitaries: let $\boldsymbol{\theta} := (\theta_1, \cdots, \theta_k)$ $(k \geq 1)$. A *k-parameterized unitary* $U(\boldsymbol{\theta})$ is a function $\mathbb{R}^k \to L(\mathcal{H})$ s.t. (1) given any $\boldsymbol{\theta}^* \in \mathbb{R}^k$, $U(\boldsymbol{\theta}^*)$ is an unitary on $\mathcal{H}$, and (2) the parameterized-matrix representation of $U(\boldsymbol{\theta})$ is entry-wise smooth.

A important family of examples is the *single-qubit rotations* about the Pauli axis X, Y, Z with angle θ:

$$R_{\boldsymbol{\sigma}}(\theta) := \exp(\frac{-i\theta}{2}\boldsymbol{\sigma}), \boldsymbol{\sigma} \in \{X, Y, Z\} \tag{3.2}$$

One can also extend Pauli rotations to multiple qubits. For example, consider *two-qubit coupling* gates $\{R_{\sigma\otimes\sigma} := \exp(\frac{-i\theta}{2}\sigma \otimes \sigma)\}_{\sigma\in\{X,Y,Z\}}$. Note that these two-qubit gates can generate entanglement between two qubits. Combined with single-qubit rotations, they form a *universal gate set* for quantum computation. Another important feature of these gates is that they can already be reliably implemented in experiments, such as ion-trap quantum computers [56].

As a result, we will work mostly with these gates in the rest of this paper. However, note that one can easily add and study other parameterized gates in our framework as well.

The language constructed above is similar to their classical counterparts. (0) **abort** terminates the program, outputting $\mathbf{0} \in \mathcal{D}(\mathcal{H}_{\overline{q}})$. (1) **skip** does nothing to states in $\mathcal{D}(\mathcal{H}_{\overline{q}})$. (2) $q := |0\rangle$ sets quantum variable q to the basis state $|0\rangle$. (3) for any $\boldsymbol{\theta}^* \in \mathbb{R}^k, \overline{q} := U(\boldsymbol{\theta}^*)[\overline{q}]$ applies the unitary $U(\boldsymbol{\theta}^*)$ to the qubits in $\overline{q}$. (4) Sequencing has the same behavior as its classical counterpart. (5) for $\boldsymbol{\theta}^* \in \mathbb{R}^k$, **case** $M[\overline{q}] = \overline{m \to P_m(\boldsymbol{\theta}^*)}$ **end** performs the measurement $M = \{M_m\}$ on the qubits in $\overline{q}$, and executes program $P_m(\boldsymbol{\theta}^*)$ if the outcome of the measurement is m. The bar over $\overline{m \to P_m}$ indicates that there may be one or more repetitions of this expression. (6) $\textbf{while}^{(T)}\ M[\overline{q}] = 1\ \textbf{do}\ P_1(\boldsymbol{\theta}^*)\ \textbf{done}$ performs the measurement $M = \{M_0, M_1\}$ on $\overline{q}$, and terminates if the outcome corresponds to M_0, or executes $P_1(\boldsymbol{\theta}^*)$ then reiterates ($T \geq 2$) / aborts ($T = 1$) otherwise. The program iterates at most T times.

We highlight two differences between quantum and classical while languages: (1) Qubits may only be initialized to the state $|0\rangle$. There is no quantum analogue for initialization to any expression (i.e. $x := e$) due to the no-cloning theorem of quantum states. Any state $|\psi\rangle \in \mathcal{H}_q$, however, can be constructed by applying some unitary U to $|0\rangle$. (2) Evaluating the guard of a case statement or loop, which performs a measurement, potentially disturbs the state of the system.

3.2 Operational and Denotational Semantics

We present the *operational semantics* of parameterized programs in Figure 1a. Transition rules are represented as $\langle P,\ \rho\rangle \to \langle P',\ \rho'\rangle$, where $\langle P,\ \rho\rangle$ and $\langle P',\ \rho'\rangle$ are quantum *configurations*.[2] In configurations, P (or P') could be a quantum program or the empty program $\downarrow$, and ρ and ρ' are partial density operators representing the current state. Intuitively, in one step, we can evaluate program P on input state ρ to program P' (or $\downarrow$) and output state ρ'. In order to present the rules in a non-probabilistic manner, the probabilities associated with each transition are encoded in the output partial density operator.[3] For m, the superoperator $\mathcal{E}_m$ is defined by $\mathcal{E}_m(\rho) = M_m \rho M_m^\dagger$, yielding the post-measurement state.

In the *Initialization* rule, the superoperators $\mathcal{E}^{\mathrm{bool}}_{q\to 0}(\rho)$ and $\mathcal{E}^{\mathrm{int}}_{q\to 0}(\rho)$, which initialize the variable q in ρ to $|0\rangle\langle 0|$, are defined by $\mathcal{E}^{\mathrm{bool}}_{q\to 0}(\rho) = |0\rangle_q\langle 0|\rho|0\rangle_q\langle 0| + |0\rangle_q\langle 1|\rho|1\rangle_q\langle 0|$ and $\mathcal{E}^{\mathrm{B-int}}_{q\to 0}(\rho) = \sum_{n=-N}^{N} |0\rangle_q\langle n|\rho|n\rangle_q\langle 0|$ $(N \in \mathbb{Z}^+)$. Here, $|\psi\rangle_q\langle\phi|$ denotes the outer product of states $|\psi\rangle$ and $|\phi\rangle$ associated with variable q; that is, $|\psi\rangle$ and $|\phi\rangle$ are in $\mathcal{H}_q$ and $|\psi\rangle_q\langle\phi|$ is a matrix over $\mathcal{H}_q$. It is conventional in the quantum information literature that when operations or measurements only apply to part of the quantum system (e.g., a subset of quantum variables of the program), one should assume that an identity is applied to the rest of quantum variables. For example, applying $|\psi\rangle_q\langle\phi|$ to ρ means applying $|\psi\rangle_q\langle\phi| \otimes I_{\mathcal{H}_{\overline{q}}}$ to ρ, where $\overline{q}$ denotes the set of all variables except q.

We present the *denotational semantics* of parameterized programs in 1b, defining $[\![P]\!]$ as a superoperator on $\rho \in \mathcal{H}_{\overline{v}}$ [53]. For more details we refer the reader to Ying [52, 53].

We have the following connection between the denotational semantics and operational for parameterized programs: in short, the meaning of running program $P(\boldsymbol{\theta}^*)$ on input state ρ and any $\boldsymbol{\theta}^* \in \mathbb{R}^k$ is the sum of all possible output states with multiplicity, weighted by their probabilities.

Proposition 3.1 ([53]). *$\forall P(\boldsymbol{\theta}) \in \textbf{q-while}^{(T)}_{\overline{v}}(\boldsymbol{\theta})$, and any specific $\boldsymbol{\theta}^* \in \mathbb{R}^k$, $\rho \in \mathcal{D}(\mathcal{H}_{\overline{v}})$,*

$$[\![P(\boldsymbol{\theta}^*)]\!](\rho) = \sum \{|\rho' : (P(\boldsymbol{\theta}^*), \rho) \to^* (\downarrow, \rho')|\}. \tag{3.3}$$

*Here $\to^*$ is the reflexive, transitive closure of $\to$ and $\{|\cdot|\}$ denotes a **multi-set**.*

We close the section with a notion arising from the following observation: some programs, while syntactically not

[2] Recall that, fixing arbitrary $\boldsymbol{\theta}^* \in \mathbb{R}^k$, both semantics reduce to those of unparameterized programs, so for compactness we write P for $P(\boldsymbol{\theta}^*)$, etc.

[3] If we had instead considered a probabilistic transition system, then one could write $\langle \textbf{case}\ M[\overline{q}] = \overline{m \to P_m}\ \textbf{end}, \rho\rangle \xrightarrow{p_m} \langle P_m, \rho_m\rangle$ where $p_m = \mathrm{tr}(M_m \rho M_m^\dagger)$ and $\rho_m = M_m \rho M_m^\dagger / p_m$.

$$\text{(Abort)}\ \overline{\langle \mathbf{abort}[\overline{q}],\ \rho\rangle \to \langle \downarrow,\ \mathbf{0}\rangle}$$

$$\text{(Skip)}\ \overline{\langle \mathbf{skip}[\overline{q}],\ \rho\rangle \to \langle \downarrow,\ \rho\rangle}$$

$$\text{(Initialization)}\ \overline{\langle q := |0\rangle,\ \rho\rangle \to \langle \downarrow,\ \rho_0^q\rangle}$$

$$\text{where } \rho_0^q = \begin{cases} \mathcal{E}^{\text{bool}}_{q\to 0}(\rho) & \text{if } type(q) = \mathbf{Bool} \\ \mathcal{E}^{\text{B-int}}_{q\to 0}(\rho) & \text{if } type(q) = \mathbf{Bdd\ Int} \end{cases}$$

$$\text{(Unitary)}\ \overline{\langle \overline{q} := U(\boldsymbol{\theta}^*)[\overline{q}],\ \rho\rangle \to \langle \downarrow,\ U(\boldsymbol{\theta}^*)\rho U^\dagger(\boldsymbol{\theta}^*)\rangle}$$

$$\text{(Sequence)}\ \frac{\langle P_1(\boldsymbol{\theta}^*),\ \rho\rangle \to \langle P_1'(\boldsymbol{\theta}^*),\ \rho'\rangle}{\langle P_1(\boldsymbol{\theta}^*); P_2(\boldsymbol{\theta}^*),\ \rho\rangle \to \langle P_1'(\boldsymbol{\theta}^*); P_2(\boldsymbol{\theta}^*),\ \rho'\rangle}$$

$$\text{(Case } m\text{)}\ \overline{\langle \mathbf{case}\ M[\overline{q}] = \overline{m \to P_m(\boldsymbol{\theta}^*)}\ \mathbf{end}, \rho\rangle \to \langle P_m(\boldsymbol{\theta}^*), \mathcal{E}_m(\rho)\rangle,\ \forall \text{ outcome } m \text{ of } M = \{M_m\}}$$

$$\text{(While}^{(T)}\ 0)\ \overline{\langle \mathbf{while}^{(T)}\ M[\overline{q}] = 1\ \mathbf{do}\ P_1(\boldsymbol{\theta}^*)\ \mathbf{done},\ \rho\rangle \to \langle \downarrow, \mathcal{E}_0(\rho)\rangle}$$

$$\text{(While}^{(T)}\ 1)\ \overline{\langle \mathbf{while}^{(T)}\ M[\overline{q}] = 1\ \mathbf{do}\ P_1(\boldsymbol{\theta}^*)\ \mathbf{done}, \rho\rangle \to \langle P_1(\boldsymbol{\theta}^*); \mathbf{while}^{(T-1)}, \mathcal{E}_1(\rho)\rangle}$$

(a)

$$\begin{aligned}
[\![\mathbf{abort}[\overline{q}]]\!]\rho &= \mathbf{0} \\
[\![\mathbf{skip}[\overline{q}]]\!]\rho &= \rho \\
[\![q := |0\rangle]\!]\rho &= \mathcal{E}^{\text{bool}}_{q\to 0}\rho \text{ or } \mathcal{E}^{\text{B-int}}_{q\to 0}(\rho) \\
[\![\overline{q} := U(\boldsymbol{\theta}^*)[\overline{q}]]\!]\rho &= U(\boldsymbol{\theta}^*)\rho U^\dagger(\boldsymbol{\theta}^*) \\
[\![P_1(\boldsymbol{\theta}^*); P_2(\boldsymbol{\theta}^*)]\!]\rho &= [\![P_2(\boldsymbol{\theta}^*)]\!]([\![P_1(\boldsymbol{\theta}^*)]\!]\rho) \\
[\![\mathbf{case}\ M[\overline{q}] = \overline{m \to P_m(\boldsymbol{\theta}^*)}\ \mathbf{end}]\!]\rho &= \textstyle\sum_m [\![P_m(\boldsymbol{\theta}^*)]\!]\mathcal{E}_m\rho \\
[\![\mathbf{while}^{(T)}\ M[\overline{q}] = 1\ \mathbf{do}\ P_1(\boldsymbol{\theta}^*)\ \mathbf{done}]\!]\rho &= \\
\textstyle\sum_{n=0}^{T-1} \mathcal{E}_0 \circ ([\![P_1(\boldsymbol{\theta}^*)]\!] \circ \mathcal{E}_1)^n \rho &
\end{aligned}$$

(b)

Figure 1. Parameterized T-bounded quantum **while** programs: (a) operational semantics (b) denotational semantics. We instantiate $\boldsymbol{\theta}^* \in \mathbb{R}^k$ since choice of parameters affects choices made at control gates. "**Bdd**" stands for **Bounded**.

"$\mathbf{abort}[\overline{q}]$", semantically aborts. Simple examples include $U(\boldsymbol{\theta})$; **abort** or a case sentence that has **abort** on each branch.

We formalize this concept (essential-abortion for unparameterized programs may be analogously defined):

Definition 3.2 ("Essentially Abort"). *Let* $P(\boldsymbol{\theta}) \in \boldsymbol{q\text{-}while}^{(T)}_{\overline{v}}(\boldsymbol{\theta})$. $P(\boldsymbol{\theta})$ *"essentially aborts" if one of the following holds:*

1. $P(\boldsymbol{\theta}) \equiv \mathbf{abort}[\overline{q}]$;
2. $P(\boldsymbol{\theta}) \equiv P_1(\boldsymbol{\theta}); P_2(\boldsymbol{\theta})$, *and either* $P_1(\boldsymbol{\theta})$ *or* $P_2(\boldsymbol{\theta})$ *essentially aborts;*
3. $P \equiv \mathbf{case}\ M[\overline{q}] = \overline{m \to P_m(\boldsymbol{\theta})}\ \mathbf{end}$, *and each* $P_m(\boldsymbol{\theta})$ *essentially aborts.*

$$\text{(Sum Components)}\ \overline{\langle \underline{P_1(\boldsymbol{\theta}^*)} + \underline{P_2(\boldsymbol{\theta}^*)},\ \rho\rangle \to \langle \underline{P_1(\boldsymbol{\theta}^*)},\ \rho\rangle,\ \langle \underline{P_1(\boldsymbol{\theta}^*)} + \underline{P_2(\boldsymbol{\theta}^*)},\ \rho\rangle \to \langle \underline{P_2(\boldsymbol{\theta}^*)},\ \rho\rangle}$$

Figure 2. additive parameterized quantum bounded while-programs: operational semantics. We fix $\boldsymbol{\theta}^* \in \mathbb{R}^k$ and inherit all the other rules from parameterized programs in Fig. 1a.

4 Additive Parameterized Quantum Bounded While-Programs

As we discussed in the introduction, we will introduce a variant of *additive* quantum programs as a succinct way to describe the collection of programs that are necessary to compute the derivatives. To that end, we introduce our design of the syntax and the semantics of additive quantum programs as well as a compilation method that turns any additive quantum program into a collection of normal programs for the actual computation of derivatives.

4.1 Syntax

We adopt the convention to use underlines to indicate additive programs, such as $\underline{P(\boldsymbol{\theta})}$, to distinguish from normal program $P(\boldsymbol{\theta})$. The syntax of $\underline{P(\boldsymbol{\theta})}$ is given by

$$\begin{aligned}
\underline{P(\boldsymbol{\theta})} ::= \ & \underline{\mathbf{abort}[\overline{q}]} \mid \underline{\mathbf{skip}[\overline{q}]} \mid \underline{q := |0\rangle} \mid \underline{\overline{q} := U(\boldsymbol{\theta})[\overline{q}]} \mid \\
& \underline{P_1(\boldsymbol{\theta}); P_2(\boldsymbol{\theta})} \mid \underline{\mathbf{case}\ M[\overline{q}] = \overline{m \to P_m(\boldsymbol{\theta})}\ \mathbf{end}} \mid \\
& \underline{\mathbf{while}^{(T)}\ M[\overline{q}] = 1\ \mathbf{do}\ P_1(\boldsymbol{\theta})\ \mathbf{done}} \mid \underline{P_1(\boldsymbol{\theta})} + \underline{P_2(\boldsymbol{\theta})},
\end{aligned}$$

where the only new syntax $+$ is the *additive* choice. Intuitively, $\underline{P_1(\boldsymbol{\theta})} + \underline{P_2(\boldsymbol{\theta})}$ refers to the program that either executes $\underline{P_1(\boldsymbol{\theta})}$ or $\underline{P_2(\boldsymbol{\theta})}$. This intuition will be implemented by the definition of its operational and denotational semantics. We assume $+$ has lower precedence order, and is left associative .[4] If $\underline{P(\boldsymbol{\theta})} = \underline{P_1(\boldsymbol{\theta})} + \underline{P_2(\boldsymbol{\theta})}$, then $\text{qVar}(\underline{P(\boldsymbol{\theta})}) \equiv \text{qVar}(\underline{P_1(\boldsymbol{\theta})}) \cup \text{qVar}(\underline{P_2(\boldsymbol{\theta})})$. Denote the collection of all non-determinisic $\underline{P(\boldsymbol{\theta})}$ s.t. $\text{qVar}(\underline{P(\boldsymbol{\theta})}) = \overline{v}$ as $\mathbf{add\text{-}q\text{-}while}^{(T)}_{\overline{v}}(\boldsymbol{\theta})$.

4.2 Operational and Denotational Semantics

We exhibit operational semantics in Figure 2 and define a similar denotational semantics for any $\underline{P(\boldsymbol{\theta}^*)} \in \mathbf{add\text{-}q\text{-}while}^{(T)}_{\overline{v}}(\boldsymbol{\theta})$.

Definition 4.1 (Denotational Semantics). *Fix* $\boldsymbol{\theta}^* \in \mathbb{R}^k$, $\rho \in \mathcal{D}(\mathcal{H}_{\overline{v}})$,

$$[\![\underline{P(\boldsymbol{\theta}^*)}]\!](\rho) \equiv \{|\rho' : \langle \underline{P(\boldsymbol{\theta}^*)}, \rho\rangle \to^* \langle \downarrow, \rho'\rangle|\} \tag{4.1}$$

We can now explain the name of $+$. Intuitively, the (S-C) rule in Figure 2 provides different choices of paths which will be eventually summed up in (4.1). This resembles the idea of the sum operator in differential lambda-calculus [12]. Interestingly, some slight change of our (S-C) rule, e.g., in the study of non-deterministic quantum programs [54], can

[4] E.g., $\underline{X} + \underline{Y}; \underline{Z} = \underline{X} + (\underline{Y}; \underline{Z})$, $X + Y + Z := (X + Y) + Z$.

also be used to describe the interlacing behavior between different components in parallel quantum programs.

4.3 Compilation Rules

We exhibit the compilation rules in Figure 3 as a way to transform a additive program $\underline{P(\boldsymbol{\theta})}$ into a multiset of normal programs. The compiled set of programs will be later used in the actual implementation of the differentiation procedure. We will show, in Section 7, that our compilation procedure is also efficient in the sense that the total number of non-essentially-aborting program in $\texttt{Compile}(\underline{P_m(\boldsymbol{\theta})})$ is reasonably bounded, in particular for the "fill and break" algorithm ($\texttt{FB}(\bullet)$) as the compilation of the case statement.

Our compilation rule is well-defined as it is compatible with the denotational semantics and operational semantics of $\underline{P(\boldsymbol{\theta})}$ in the following sense:

Proposition 4.2. *Denoting with $\coprod$ the union of multisets, then for any $\rho \in \mathcal{D}(\mathcal{H}_{\overline{v}})$,*

$$\{|\rho' : \rho' \neq \mathbf{0}, \rho' \in [\![\underline{P(\boldsymbol{\theta}^*)}]\!]\rho|\} = \coprod_{Q(\boldsymbol{\theta}) \in \texttt{Compile}(\underline{P(\boldsymbol{\theta})})} \{|\rho' \neq \mathbf{0} : \langle Q(\boldsymbol{\theta}^*), \rho\rangle \to^* \langle\downarrow, \rho'\rangle|\}. \quad (4.2)$$

Proof. Structural Induction. See Appendix C.1 for details. □

Note that (4.2) removes $\mathbf{0}$ from the multi-set as we are only interested in non-trivial final states. Moreover, in $\texttt{Compile}(\underline{P(\boldsymbol{\theta})})$, some programs may essentially abort (Definition 3.2). For implementation, we are interested in the number of $Q(\boldsymbol{\theta}) \in \texttt{Compile}(\underline{P(\boldsymbol{\theta})})$ that do not essentially abort:

Definition 4.3 (Number of Non-Aborting Programs)**.** *The* number of non-aborting programs *of $\underline{P(\boldsymbol{\theta})}$, denoted as $|\#\underline{P(\boldsymbol{\theta})}|$, is defined as*

$$|\#\underline{P(\boldsymbol{\theta})}| = |\texttt{Compile}(\underline{P(\boldsymbol{\theta})}) \setminus \{|Q(\boldsymbol{\theta}) \in \texttt{Compile}(\underline{P(\boldsymbol{\theta})}) : Q(\boldsymbol{\theta}) \textit{ essentially aborts.}|\}|$$

where $|C|$ is the cardinality of a multiset C and $C_0 \setminus C_1$ denotes the multiset difference of C_0 and C_1.

Example 4.1 (Generic-Case)**.** *Consider the following simple program with the case statement*

$$\underline{P(\boldsymbol{\theta})} \equiv \underline{\textbf{case}}\ M[\overline{q}] = 0 \to \underline{P_1(\boldsymbol{\theta})} + \underline{P_2(\boldsymbol{\theta})},\quad 1 \to \underline{P_3(\boldsymbol{\theta})}$$

where $P_1(\boldsymbol{\theta}), P_2(\boldsymbol{\theta}), P_3(\boldsymbol{\theta}) \in \underline{\boldsymbol{q}\textbf{-while}^{(T)}_{\overline{v}}(\boldsymbol{\theta})}$, none of them essentially aborts, and each of $P_1(\boldsymbol{\theta}), P_2(\boldsymbol{\theta}), P_3(\boldsymbol{\theta})$ contains no control gates. Then for any $\rho \in \mathcal{D}(\mathcal{H}_{\overline{v}})$, fixing $\boldsymbol{\theta}^*$ we have

$$\begin{aligned}
\langle \underline{P(\boldsymbol{\theta}^*)}, \rho\rangle &\overset{(Case\ m)}{\to} \langle \underline{P_1(\boldsymbol{\theta}^*)} + \underline{P_2(\boldsymbol{\theta}^*)}, M_0\rho M_0^\dagger\rangle \\
&\overset{(Sum)}{\to} \langle \underline{P_1(\boldsymbol{\theta}^*)}, M_0\rho M_0^\dagger\rangle \\
&\to^* \langle\downarrow, [\![P_1(\boldsymbol{\theta}^*)]\!](M_0\rho M_0^\dagger)\rangle; \\
\langle \underline{P(\boldsymbol{\theta}^*)}, \rho\rangle &\overset{(Case\ m)}{\to} \langle \underline{P_1(\boldsymbol{\theta}^*)} + \underline{P_2(\boldsymbol{\theta}^*)}, M_0\rho M_0^\dagger\rangle \\
&\overset{(Sum)}{\to} \langle \underline{P_2(\boldsymbol{\theta}^*)}, M_0\rho M_0^\dagger\rangle \to^* \langle\downarrow, [\![P_2(\boldsymbol{\theta}^*)]\!](M_0\rho M_0^\dagger)\rangle; \\
\langle \underline{P(\boldsymbol{\theta}^*)}, \rho\rangle &\overset{(Case\ m)}{\to} \langle \underline{P_3(\boldsymbol{\theta}^*)}, M_1\rho M_1^\dagger\rangle \\
&\to^* \langle\downarrow, [\![P_3(\boldsymbol{\theta}^*)]\!](M_1\rho M_1^\dagger)\rangle
\end{aligned}$$

(Atomic) $\texttt{Compile}(\underline{P(\boldsymbol{\theta})}) \equiv \{|\underline{P(\boldsymbol{\theta})}|\}$, if $\underline{P(\boldsymbol{\theta})} \equiv \underline{\textbf{abort}[\overline{v}]} \mid \underline{\textbf{skip}[\overline{v}]} \mid \underline{q := |0\rangle} \mid \underline{\overline{v} := U(\boldsymbol{\theta})[\overline{v}]}$.

(Sequence) $\texttt{Compile}(\underline{P_1(\boldsymbol{\theta}); P_2(\boldsymbol{\theta})}) \equiv$

$$\begin{cases} \{|\textbf{abort}|\}, \text{ if } \texttt{Compile}(\underline{P_1(\boldsymbol{\theta})}) = \{|\textbf{abort}|\}; \\ \{|\textbf{abort}|\}, \text{ if } \texttt{Compile}(\underline{P_2(\boldsymbol{\theta})}) = \{|\textbf{abort}|\}; \\ \{|Q_1(\boldsymbol{\theta}); Q_2(\boldsymbol{\theta}) : Q_b(\boldsymbol{\theta}) \in \texttt{Compile}(\underline{P_b(\boldsymbol{\theta})})|\}, \\ \qquad\qquad \text{otherwise.} \end{cases}$$

(Case m) $\texttt{Compile}(\underline{\textbf{case}}) \equiv \texttt{FB}(\underline{\textbf{case}})$, described in Fig.3*b*.

(While$^{(T)}$) $\texttt{Compile}(\underline{\textbf{while}^{(T)}})$: use (Case m) and (Sequence).

(Sum) $\texttt{Compile}(\underline{P_1(\boldsymbol{\theta}) + P_2(\boldsymbol{\theta})}) \equiv$

$$\begin{cases} \texttt{Compile}(\underline{P_1(\boldsymbol{\theta})}) \coprod \texttt{Compile}(\underline{P_2(\boldsymbol{\theta})}), \text{ if } \forall b \in \{1, 2\}, \texttt{Compile}(\underline{P_b(\boldsymbol{\theta})}) \neq \{|\textbf{abort}|\}; \\ \texttt{Compile}(\underline{P_1(\boldsymbol{\theta})}), \text{ if } \texttt{Compile}(\underline{P_2(\boldsymbol{\theta})}) = \{|\textbf{abort}|\}, \texttt{Compile}(\underline{P_1(\boldsymbol{\theta})}) \neq \{|\textbf{abort}|\}; \\ \texttt{Compile}(\underline{P_2(\boldsymbol{\theta})}), \text{ if } \texttt{Compile}(\underline{P_1(\boldsymbol{\theta})}) = \{|\textbf{abort}|\}, \texttt{Compile}(\underline{P_2(\boldsymbol{\theta})}) \neq \{|\textbf{abort}|\}; \\ \{|\textbf{abort}|\}, \text{otherwise} \end{cases}$$

(a)

1. $\forall m \in [0, w]$, let C_m denote the sub-multiset of $\texttt{Compile}(\underline{P_m(\boldsymbol{\theta})})$ composed of programs that do *not* essentially abort; without loss of generality, assume $|C_0| \geq |C_1| \geq \cdots \geq |C_w|$.
2. If all C_m's are empty, return $\texttt{FB}(\textbf{case}) \equiv \{|\textbf{abort}[\overline{v}]|\}$; else, pad each C_m to size $|C_0|$ by adding "$\textbf{abort}[\overline{v}]$".
3. $\forall m \in [0, w]$, index programs in C_m as $\{|Q_{m,0}(\boldsymbol{\theta}), \cdots, Q_{m,|C_0|-1}(\boldsymbol{\theta})|\}$.
4. Return $\texttt{FB}(\textbf{case}) \equiv \{|\ \textbf{case}\ M[\overline{q}] = \overline{m \to Q_{m,j^*}}\ \textbf{end}\ |\}_{j^*}$ with $0 \leq j^* \leq |C_0| - 1$.

(b)

Figure 3. nondeterministic programs: (a) compilation rules. (b) "Fill and Break" ("$\texttt{FB}(\bullet)$") procedure for computing $\texttt{Compile}(\underline{\textbf{case}})$. $\underline{\textbf{case}}$ stands for $\underline{\textbf{case}\ M[\overline{q}] = \overline{m \to P_m(\boldsymbol{\theta})}\ \textbf{end}}$; $\underline{\textbf{while}^{(T)}}$ stands for $\underline{\textbf{while}^{(T)}\ M[\overline{q}] = 1\ \textbf{do}\ P_1(\boldsymbol{\theta})\ \textbf{done}}$. Here $\coprod$ denotes union of multisets. One may observe from a routine structural induction and the definition of "essentially abort" that: for all $\underline{P(\boldsymbol{\theta})}$, either $\texttt{Compile}(\underline{P(\boldsymbol{\theta})}) = \{|\textbf{abort}|\}$, or $\texttt{Compile}(\underline{P(\boldsymbol{\theta})})$ does not contain essentially abort programs.

Hence by Definition 4.1.

$$[\![\underline{P(\boldsymbol{\theta}^*)}]\!]\rho = \{|[\![P_1(\boldsymbol{\theta}^*)]\!](M_0\rho M_0^\dagger), [\![P_2(\boldsymbol{\theta}^*)]\!](M_0\rho M_0^\dagger), [\![P_3(\boldsymbol{\theta}^*)]\!](M_1\rho M_1^\dagger)|\}$$

We verify computation results from the compilation rules are consist with this. Writing "compilation rule" as "CP" for short, one observes $\texttt{Compile}(\underline{P_1(\boldsymbol{\theta}) + P_2(\boldsymbol{\theta})}) \overset{CP,Sum}{=} \{|P_1(\boldsymbol{\theta}), P_2(\boldsymbol{\theta})|\}$, *while* $\texttt{Compile}(\underline{P_3(\boldsymbol{\theta})}) = \{|P_3(\boldsymbol{\theta})|\}$ *since we assumed non-essentially-abortness. Apply our "fill and break" procedure to obtain* $C_0 = \{|P_1(\boldsymbol{\theta}), P_2(\boldsymbol{\theta})|\}, C_1 = \{|P_3(\boldsymbol{\theta}), \textbf{abort}[\overline{v}]|\}$.

$$\texttt{Compile}(\underline{P(\boldsymbol{\theta})}) = \left\{\!\!\left| \begin{array}{rl} \underline{\textbf{case}}\ M[\overline{q}] = 0 \to & P_1(\boldsymbol{\theta}), \\ 1 \to & P_3(\boldsymbol{\theta}), \\ \underline{\textbf{case}}\ M[\overline{q}] = 0 \to & P_2(\boldsymbol{\theta}), \\ 1 \to & \textbf{abort}[\overline{v}] \end{array} \right|\!\!\right\}$$

Evolving pursuant to the normal programs operational semantics (Fig 1) leads to the follow that agrees with $[\![\underline{P(\boldsymbol{\theta}^*)}]\!]\rho$.

$$\begin{aligned} & \coprod_{Q(\boldsymbol{\theta}) \in \texttt{Compile}(\underline{P(\boldsymbol{\theta})})} \{|\rho' : \langle Q(\boldsymbol{\theta}^*), \rho\rangle \to^* \langle\downarrow, \rho'\rangle|\} \\ = \ & \{|[\![P_1(\boldsymbol{\theta}^*)]\!](M_0\rho M_0^\dagger), [\![P_2(\boldsymbol{\theta}^*)]\!](M_0\rho M_0^\dagger), \\ & [\![P_3(\boldsymbol{\theta}^*)]\!](M_1\rho M_1^\dagger), \mathbf{0}|\}. \end{aligned}$$

5 Observable and Differential Semantics

To capture physically observable quantities from quantum systems, physicists propose the notation of *observable* which is a Hermitian matrix over the same space. Any observable O is a combination of information about quantum measurements and classical values for each measurement outcome. To see why, let us take its spectral decomposition of $O = \sum_m \lambda_m |\psi_m\rangle\langle\psi_m|$. Then $\{|\psi_m\rangle\langle\psi_m|\}_m$ form a projective measurement. We can design an experiment to perform this projective measurement and output λ_m when the outcome is m. The *expectation* of the output is exactly given by

$$\mathrm{tr}(O\rho) = \sum_m a_m \mathrm{tr}(M_m \rho M_m^\dagger). \tag{5.1}$$

Comparing to the random outcome of any quantum measurement, the expectation $\mathrm{tr}(O\rho)$ given by any observable O represents meaningful classical information of quantum systems. In applications of quantum machine learning, it is also the quantity used in the loss function. Thus, given any observable O, we will define the *observable* semantics of quantum programs both the mathematical object to take derivatives from the original programs and the read-out of the programs that compute these derivatives.

Moreover, one should note the extra cost to approximate $\mathrm{tr}(O\rho)$ to high precision. For example, one can repeat the aforementioned experiment with $\{|\psi_m\rangle\langle\psi_m|\}_m$ measurement and use the statistical information to recover $\mathrm{tr}(O\rho)$. The number of iterations will depend on the additive precision δ and the norm of O. To simplify our presentation, also to make a precise resource count as detailed in Section 7, throughout the paper, we assume that[5]

$$-I_{\mathcal{H}} \sqsubseteq O \sqsubseteq I_{\mathcal{H}}. \tag{5.2}$$

By statistically concentration bounds (e.g. the Chernoff bound), to approximate $\mathrm{tr}(O\rho)$ with additive error δ, one needs to repeat $O(1/\delta^2)$ times with $O(1/\delta^2)$ copies of initial states.

5.1 Observable Semantics

We define the *observable semantics* of both normal (denoted by $P(\boldsymbol{\theta}), P'(\boldsymbol{\theta})$) and additive (denoted by $\underline{S(\boldsymbol{\theta})}, \underline{S'(\boldsymbol{\theta})}$) parameterized programs as follows.

Definition 5.1 (Observable Semantics). $\forall P(\boldsymbol{\theta}) \in \boldsymbol{q\text{-}while}_{\overline{v}}^{(T)}(\boldsymbol{\theta})$, *any observable* $O \in \mathcal{O}_{\overline{v}}$ *and input state* $\rho \in \mathcal{D}(\mathcal{H}_{\overline{v}})$, *the* observable semantics *of* P, *denoted* $[\![(O,\rho) \to P(\boldsymbol{\theta})]\!]$, *is*

$$[\![(O,\rho) \to P(\boldsymbol{\theta})]\!](\boldsymbol{\theta}^*) \equiv \mathrm{tr}(O[\![P(\boldsymbol{\theta}^*)]\!]\rho), \forall \boldsymbol{\theta}^* \in \mathbb{R}^k. \tag{5.3}$$

Namely, $[\![(O,\rho) \to P(\boldsymbol{\theta})]\!]$ *is a function from* $\mathbb{R}^k$ *to* $\mathbb{R}$ *whose value per point is given by (5.3).*

Similarly, for any $\underline{S(\boldsymbol{\theta})} \in \boldsymbol{add\text{-}q\text{-}while}_{\overline{v}}^{(T)}(\boldsymbol{\theta})$ *with* $\texttt{Compile}(\underline{S(\boldsymbol{\theta})}) = \{|P_i(\boldsymbol{\theta})|\}_{i=1}^t$ *where* $P_i(\boldsymbol{\theta}) \in \boldsymbol{q\text{-}while}_{\overline{v}}^{(T)}(\boldsymbol{\theta})$, *its* observable semantics *is given by*

$$[\![(O,\rho) \to \underline{S(\boldsymbol{\theta})}]\!](\boldsymbol{\theta}^*) \equiv \sum_{i\in[1,t]} [\![(O,\rho) \to P_i(\boldsymbol{\theta})]\!](\boldsymbol{\theta}^*), \forall \boldsymbol{\theta}^* \in \mathbb{R}^k. \tag{5.4}$$

To compute gradients of quantum observables for each parameter, one needs an ancilla variable as hinted by results in quantum information theory about gradient calculations for simple unitaries (e.g., Bergholm et al. [7], Schuld et al. [41]). To that end, we can easily extend quantum programs with ancilla variables. For each $j \in [1,k]$, the *j-th ancilla of* $\boldsymbol{q\text{-}while}_{\overline{v}}^{(T)}(\boldsymbol{\theta})$ is a quantum variable denoted by $A_{j,\overline{v}}$ disjoint from $\overline{v}$. We write A instead of $A_{j,\overline{v}}$ when $j, \overline{v}$ are clear from context. Ancilla A could consists of any number of qubits while we will mostly use *one-qubit* A in this paper.

We will only consider programs augmented with *one ancilla variable* A_j at any time. (So let us fix j for the following discussion). We will then consider programs that operate on the larger space $\mathcal{D}(\mathcal{H}_{\overline{v}\cup\{A\}})$ and an additional observable A to define the *observable semantics with ancilla.*

Definition 5.2 (Observable Semantics with Ancilla). *Given any* $P'(\boldsymbol{\theta}) \in \boldsymbol{q\text{-}while}_{\overline{v}\cup\{A_{j,\overline{v}}\}}^{(T)}(\boldsymbol{\theta})$, *any observable* $O \in \mathcal{O}_{\overline{v}}$, *input state* $\rho \in \mathcal{D}(\mathcal{H}_{\overline{v}})$, *and moreover the observable* O_A *on ancilla* A, *the* observable semantics *with ancilla of* P, *overloading the notation* $[\![(O,\rho) \to P'(\boldsymbol{\theta})]\!]$, *is*

$$[\![((O,O_A),\rho) \to P'(\boldsymbol{\theta})]\!](\boldsymbol{\theta}^*) \equiv \mathrm{tr}\Big((O_A \otimes O)[\![P'(\boldsymbol{\theta}^*)]\!](((|\overline{0}\rangle_{\{A\}}\langle\overline{0}|) \otimes \rho)\Big), \forall \boldsymbol{\theta}^* \in \mathbb{R}^k.$$

Again, $[\![((O,O_A),\rho) \to P(\boldsymbol{\theta})]\!]$ *is a function from* $\mathbb{R}^k$ *to* $\mathbb{R}$ *whose value per point is given by (5.5).*

[5] $\sqsubseteq$ is defined by $A \sqsubseteq B \iff B - A$ positive semidefinite. See Appendix C.1.

Similarly, for $\underline{S'(\boldsymbol{\theta})} \in \textbf{\textit{add-q-while}}^{(T)}_{\overline{v}\cup\{A_{j,\overline{v}}\}}(\boldsymbol{\theta})$ *such that* $\texttt{Compile}\,(\underline{S'(\boldsymbol{\theta})}) = \{|P'_i(\boldsymbol{\theta})|\}_{i=1}^{t}$ *where* $P'_i(\boldsymbol{\theta}) \in \textbf{\textit{q-while}}^{(T)}_{\overline{v}\cup\{A_{j,\overline{v}}\}}(\boldsymbol{\theta})$*, its* observable semantics *is:* $\forall \boldsymbol{\theta}^* \in \mathbb{R}^k$,

$$[\![((O,O_A),\rho)\to \underline{S'(\boldsymbol{\theta})}]\!](\boldsymbol{\theta}^*) \equiv \sum_{i\in[1,t]} [\![((O,O_A),\rho)\to P'_i(\boldsymbol{\theta})]\!](\boldsymbol{\theta}^*) \tag{5.5}$$

The only difference from the normal observable semantics lies in (5.5), where we initialize the ancilla with $|0\rangle$, which is a natural choice and evaluate the observable $O_A \otimes O$. As we will see in the technique, the independence between O_A and O in the form of $O_A \otimes O$ will help use obtain the strongest guarantee of our differentiation procedure.

5.2 Differential Semantics

Given the definition of observable semantics, its differential semantics can be naturally defined by

Definition 5.3 (Differential Semantics)**.** *Given additive program* $\underline{S(\boldsymbol{\theta})} \in \textbf{\textit{add-q-while}}^{(T)}_{\overline{v}}(\boldsymbol{\theta})$*, its* j*-th* differential semantics *is defined by*

$$\frac{\partial}{\partial\theta_j}([\![(O,\rho)\to \underline{S(\boldsymbol{\theta})}]\!]), \tag{5.6}$$

which is again a function from $\mathbb{R}^k$ *to* $\mathbb{R}$*. Moreover, for any* $\underline{S'(\boldsymbol{\theta})} \in \textbf{\textit{add-q-while}}^{(T)}_{\overline{v}\cup\{A\}}(\boldsymbol{\theta})$ *with ancilla* A*, we say that "*$\underline{S'(\boldsymbol{\theta})}$ computes the j-th differential semantics of $\underline{S(\boldsymbol{\theta})}$*" if and only if there exists an observable* O_A *on ancilla* A *for* $\underline{S'(\boldsymbol{\theta})}$ *such that* $\forall O \in \mathcal{O}_{\overline{v}}, \rho \in \mathcal{D}(\mathcal{H}_{\overline{v}})$,

$$[\![((O,O_A),\rho)\to \underline{S'(\boldsymbol{\theta})}]\!] = \frac{\partial}{\partial\theta_j}([\![(O,\rho)\to \underline{S(\boldsymbol{\theta})}]\!]). \tag{5.7}$$

We remark that (5.6) is well defined because $[\![(O,\rho)\to \underline{S(\boldsymbol{\theta})}]\!]$ is a function from $\mathbb{R}^k$ to $\mathbb{R}$. It is also a smooth function because we assume that parameterized unitaries are entry-wise smooth, and the observable semantics is obtained by multiplication and addition of such entries.

We also remark that the order of quantifiers in (5.7) makes the predicate the strongest that one can hope for. This is because the observable semantics of $\underline{S(\boldsymbol{\theta})}$ will depend on O and ρ in general. Thus, the program to compute its differential semantics could also depend on O and ρ in general. However, in our definition, $\underline{S'(\boldsymbol{\theta})}$ is a single fixed program that works for any O and ρ regardless of the seemingly complicated relationship. We remark that this definition is consistent with the classical case where a single program is used to compute the derivatives for any input. Indeed, we can achieve this definition in the quantum setting and it is also critical in the proof of Theorem 6.2 (item (5)).

6 Code Transformations and the Differentiation Logic

We describe the code transformation rules of the differentiation operator $\frac{\partial}{\partial\theta}(\cdot)$ in Section 6.1. We also define a logic and prove its soundness for reasoning about the correctness of these code transformations, with the following judgement

$$\underline{S'(\boldsymbol{\theta})}|\underline{S(\boldsymbol{\theta})}, \tag{6.1}$$

which states that $\underline{S'(\boldsymbol{\theta})}$ computes the differential semantics of $\underline{S(\boldsymbol{\theta})}$ in sense of Definition 5.3. We fix $\theta = \theta_j$ and hence A stands for $A_{j,\overline{v}}$ and $\frac{\partial}{\partial\theta}$ for $\frac{\partial}{\partial\theta_j}$ through this section.[6]

6.1 Code Transformations

We first define some gates associated with the single-qubit rotation and the two-qubit coupling gates, which will appear in the code transformation rules. Let A be a single qubit.

Definition 6.1. *1. Consider* $|q_1\rangle := R_\sigma(\theta)|q_1\rangle$*, where* $\sigma \in \{X,Y,Z\}$*. Define*

$$C - R_\sigma(\theta)|A,q_1\rangle \equiv \Big((|0\rangle_A\langle 0|)\otimes R_\sigma(\theta)\Big) + \tag{6.2}$$

$$\Big((|1\rangle_A\langle 1|)\otimes R_\sigma(\theta+\pi)\Big) \tag{6.3}$$

$$R'_\sigma(\theta)|A,q_1\rangle \equiv A := H|A\rangle; C - R_\sigma(\theta)|A,q_1\rangle; A := H|A\rangle. \tag{6.4}$$

2. Substituting $\sigma\otimes\sigma$ *for* σ *and* q_1,q_2 *for* q_1 *in Eqns (6.3,6.4), one defines* $C - R_{\sigma\otimes\sigma}(\theta), R'_{\sigma\otimes\sigma}(\theta)$.

For 1-qubit rotation $R_\sigma(\theta)$, the "controlled-rotation" gate $C-R_\sigma(\theta)$ maps $|0,q_1\rangle \mapsto |0\rangle\otimes R_\sigma(\theta)\,|q_1\rangle$, and $|1,q_1\rangle \mapsto |1\rangle\otimes R_\sigma(\theta+\pi)\,|q_1\rangle$; $R'_\sigma(\theta)$ conjugates $C - R_\sigma(\theta)$ with Hadamard. Similarly for corresponding two-qubit coupling gates.

We exhibit our code transformation rules in Figure 4. For Unitary rules we only include 1-qubit rotations and two-qubit coupling gates, since they form a universal gate set and are easy to implement on quantum machines. It is also possible to include more unitary rules (e.g., by following the calculations in [41]), which we will leave as future directions.

6.2 The differentiation logic and its soundness

We develop the differentiation logic given in Figure 5 to reason about the correctness of code transformations. It suffices to show that our logic is sound. For ease of notation, in future analysis we write $\frac{\partial}{\partial\theta}(P(\boldsymbol{\theta}))$ in place of $\frac{\partial}{\partial\theta}(\underline{P(\boldsymbol{\theta})})$ when $P(\boldsymbol{\theta}) \in \textbf{q-while}^{(T)}_{\overline{v}}(\boldsymbol{\theta})$.

Theorem 6.2 (Soundness)**.** *Let* $\underline{S(\boldsymbol{\theta})} \in \textbf{\textit{add-q-while}}^{(T)}_{\overline{v}}(\boldsymbol{\theta})$*,* $\underline{S'(\boldsymbol{\theta})} \in \textbf{\textit{add-q-while}}^{(T)}_{\overline{v}\cup\{A\}}(\boldsymbol{\theta})$*. Then,*

$$\underline{S'(\boldsymbol{\theta})}|\underline{S(\boldsymbol{\theta})} \implies \underline{S'(\boldsymbol{\theta})} \textit{ computes the differential semantics of } \underline{S(\boldsymbol{\theta})}. \tag{6.5}$$

[6] If A already exists, i.e., $\underline{S(\boldsymbol{\theta})} \in \textbf{add-q-while}^{(T)}_{\overline{v}\cup\{A\}}(\boldsymbol{\theta})$, we treat $\overline{v}_{\text{new}}$ as $\overline{v}_{\text{old}} \cup A_{\text{old}}$ and add A_{new}. Any observable O on $\overline{v}_{\text{old}}$ becomes $O_{A_{\text{old}}} \otimes O$ on $\overline{v}_{\text{new}}$. Both A_{old} and A_{new} are initialized to $|0\rangle$ in observable semantics.

(Trivial) $\frac{\partial}{\partial\theta}(\underline{\mathbf{abort}[\overline{v}]}), \frac{\partial}{\partial\theta}(\underline{\mathbf{skip}[\overline{v}]}), \frac{\partial}{\partial\theta}(\underline{q := |0\rangle}) \equiv \underline{\mathbf{abort}[\overline{v} \cup \{A\}]}.$

(Trivial-Unitary) $\frac{\partial}{\partial\theta}(\underline{|\overline{v}\rangle := U(\boldsymbol{\theta})|\overline{v}\rangle}) \equiv \underline{\mathbf{abort}[\overline{v} \cup \{A\}]}$, if $\theta_j \notin \boldsymbol{\theta}$.

(1-qb Rotation) $\frac{\partial}{\partial\theta}(\underline{|q_1\rangle := R_\sigma(\theta)|q_1\rangle}) \equiv \underline{|A, q_1\rangle := R'_\sigma(\theta)|A, q_1\rangle}.$

(2-qb Coupling) $\frac{\partial}{\partial\theta}(\underline{|q_1, q_2\rangle := R_{\sigma\otimes\sigma}(\theta)|q_1\rangle}) \equiv \underline{|A, q_1, q_2\rangle := R'_{\sigma\otimes\sigma}(\theta)|A, q_1, q_2\rangle}.$

(Sequence) $\frac{\partial}{\partial\theta}(\underline{S_1(\boldsymbol{\theta}); S_2(\boldsymbol{\theta})}) \equiv (\underline{S_1(\boldsymbol{\theta})}; \underline{\frac{\partial}{\partial\theta}(S_2(\boldsymbol{\theta}))}) + (\underline{\frac{\partial}{\partial\theta}(S_1(\boldsymbol{\theta}))}; \underline{S_2(\boldsymbol{\theta})}).$

(Case) $\frac{\partial}{\partial\theta}(\underline{\mathbf{case}\ M[\overline{q}] = \overline{m \to S_m(\boldsymbol{\theta})}\ \mathbf{end}}) \equiv \underline{\mathbf{case}\ M[\overline{q}] = \overline{m \to \frac{\partial}{\partial\theta}(S_m(\boldsymbol{\theta}))}\ \mathbf{end}}.$

(while$^{(T)}$) Use (Case) and (Sequence).

(Sum Components) $\frac{\partial}{\partial\theta}(\underline{S_1(\boldsymbol{\theta})} + \underline{S_2(\boldsymbol{\theta})}) \equiv \underline{\frac{\partial}{\partial\theta}(S_1(\boldsymbol{\theta}))} + \underline{\frac{\partial}{\partial\theta}(S_2(\boldsymbol{\theta}))}$

Figure 4. Code Transformation Rules. For (1-qb Rotation) and (2-qb Coupling), $(\sigma \in \{X, Y, Z\})$; $R'_\sigma(\theta), R'_{\sigma\otimes\sigma}(\theta)$ are as in Definition 6.1. $\theta_j \notin \boldsymbol{\theta}$ means "the unitary $U(\boldsymbol{\theta})$ trivially uses θ_j": for example in $P(\boldsymbol{\theta})|q_1\rangle \equiv R_X(\theta_1); R_Z(\theta_2)$, $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and $R_X(\theta_1)$ trivially uses θ_2.

(Abort) $\overline{\frac{\partial}{\partial\theta}(\underline{\mathbf{abort}[\overline{v}]})|\underline{\mathbf{abort}[\overline{v}]}}$ (Skip) $\overline{\frac{\partial}{\partial\theta}(\underline{\mathbf{skip}[\overline{v}]})|\underline{\mathbf{skip}[\overline{v}]}}$

(Initialization) $\overline{\frac{\partial}{\partial\theta}(\underline{q := |0\rangle})|\underline{(q := |0\rangle)}}$ $\theta_j \notin \boldsymbol{\theta}$

(Trivial-Unitary) $\overline{\frac{\partial}{\partial\theta}(\underline{\overline{q} = U(\boldsymbol{\theta})[\overline{v}]})|\underline{\overline{q} = U(\boldsymbol{\theta})[\overline{v}]}}$

(Rot-Couple) $\overline{\frac{\partial}{\partial\theta}(\underline{\overline{q} = R_{\boldsymbol{\sigma}}(\boldsymbol{\theta})[\overline{v}]})|\underline{(\overline{q} = R_{\boldsymbol{\sigma}}(\boldsymbol{\theta})[\overline{v}])}}$ $\overline{\frac{\partial}{\partial\theta}(\underline{S_0(\boldsymbol{\theta})})|\underline{S_0(\boldsymbol{\theta})}}$ $\overline{\frac{\partial}{\partial\theta}(\underline{S_1(\boldsymbol{\theta})})|\underline{S_1(\boldsymbol{\theta})}}$

(Sequence) $\overline{\frac{\partial}{\partial\theta}(\underline{S_0(\boldsymbol{\theta}); S_1(\boldsymbol{\theta})})|\underline{(S_0(\boldsymbol{\theta}); S_1(\boldsymbol{\theta}))}}$ $\forall m, \overline{\frac{\partial}{\partial\theta}(\underline{S_m(\boldsymbol{\theta})})|\underline{S_m(\boldsymbol{\theta})}}$

(Case) $\overline{\frac{\partial}{\partial\theta}(\mathbf{case}\ M[\overline{q}] = \overline{m \to S_m(\boldsymbol{\theta})}\ \mathbf{end})|\mathbf{case}\ M[\overline{q}] = \overline{m \to S_m(\boldsymbol{\theta})}\ \mathbf{end}}$ $\overline{\frac{\partial}{\partial\theta}(\underline{S_1(\boldsymbol{\theta})})|\underline{S_1(\boldsymbol{\theta})}}$

(While$^{(T)}$) $\overline{\frac{\partial}{\partial\theta}(\underline{\mathbf{while}^{(T)}\ M[\overline{q}] = 1\ \mathbf{do}\ S_1(\boldsymbol{\theta})\ \mathbf{done}})|\underline{\mathbf{while}^{(T)}\ M[\overline{q}] = 1\ \mathbf{do}\ S_1(\boldsymbol{\theta})\ \mathbf{done}}}$ $\overline{\frac{\partial}{\partial\theta}(\underline{S_0(\boldsymbol{\theta})})|\underline{S_0(\boldsymbol{\theta})}}$ $\overline{\frac{\partial}{\partial\theta}(\underline{S_1(\boldsymbol{\theta})})|\underline{S_1(\boldsymbol{\theta})}}$

(Sum Component) $\overline{\frac{\partial}{\partial\theta}(\underline{S_0(\boldsymbol{\theta})} + \underline{S_1(\boldsymbol{\theta})})|\underline{(S_0(\boldsymbol{\theta})} + \underline{S_1(\boldsymbol{\theta}))}}$

Figure 5. The differentiation logic. Wherever applicable, $q \in \overline{v}, \overline{q} \subseteq \overline{v}, S_i(\boldsymbol{\theta}) \in \mathbf{add\text{-}q\text{-}while}^{(T)}_{\overline{v}}(\boldsymbol{\theta})$. In (Rot-Couple), $\boldsymbol{\sigma} \in \{X, Y, Z, X\otimes X, Y\otimes Y, Z\otimes Z\}$.

Let us highlight the ideas behind the proof of the soundness and all detailed proofs are deferred to Appendix D. First remember that $\theta = \theta_j$ and for all the proofs we can choose $Z_A = |0\rangle\langle 0| - |1\rangle\langle 1|$ as the observable on the one-qubit ancilla A. Thus, we will omit Z_A and *overload* the notation, $\forall P'(\boldsymbol{\theta}) \in \mathbf{q\text{-}while}^{(T)}_{\overline{v}\cup\{A\}}(\boldsymbol{\theta})$:

$$[\![(O, \rho) \to P'(\boldsymbol{\theta})]\!] \text{ means } [\![((O, Z_A), \rho) \to P'(\boldsymbol{\theta})]\!], \tag{6.6}$$

to simplify the presentation. We make similar overloading convention for $\underline{S'(\boldsymbol{\theta})} \in \mathbf{add\text{-}q\text{-}while}^{(T)}_{\overline{v}\cup\{A\}}(\boldsymbol{\theta})$. Let us go through these logic rules one by one.

1. *Abort, Skip, Initialization, Trivial-Unitary* rules work because these statements do not depend on θ.
2. Since While$^{(T)}$ can be deemed as a macro of other statements, the correctness of *While*$^{(T)}$ rule follows by unfolding while$^{(T)}$ and applying other rules.
3. The *Sum Component* rule is due to the property of observable semantics ($[\![\cdot]\!]$) and additive operator ($+$):
$$\frac{\partial}{\partial\theta}([\![P_1 + P_2]\!]) = [\![\frac{\partial}{\partial\theta}(P_1)]\!] + [\![\frac{\partial}{\partial\theta}(P_2)]\!], \tag{6.7}$$
which follows from our definition design.
4. Our *Rot-Couple* rule is different from the phase-shift rule in [41] by using only one circuit in derivative computing. However, the proof of the *Rot-Couple* rule is largely inspired by the one of the phase-shift rule.
5. The proof of the *Sequence* rule relies very non-trivially on our design of the observable semantics with ancilla (Definition 5.2) and the strong requirement of computing differential semantics in Definition 5.3. Firstly,
$$[\![(O, \rho) \to \frac{\partial}{\partial\theta}(\underline{S_0(\boldsymbol{\theta}); S_1(\boldsymbol{\theta})})]\!] = [\![(O, \rho) \to \frac{\partial}{\partial\theta}(\underline{S_0(\boldsymbol{\theta})}); \underline{S_1(\boldsymbol{\theta})}]\!] + [\![(O, \rho) \to \underline{S_0(\boldsymbol{\theta})}; \frac{\partial}{\partial\theta}(\underline{S_1(\boldsymbol{\theta})})]\!]$$
We use the induction hypothesis to reason about each term above. Consider the case $\underline{S_0(\boldsymbol{\theta})} = S_0(\boldsymbol{\theta})$ and $\underline{S_1(\boldsymbol{\theta})} = S_1(\boldsymbol{\theta})$. We show the following for the second step:
$$[\![(O, \rho) \to S_0(\boldsymbol{\theta}); \frac{\partial}{\partial\theta}(S_1(\boldsymbol{\theta}))]\!] = [\![(O, [\![S_0(\boldsymbol{\theta})]\!](\rho)) \to \frac{\partial}{\partial\theta}(S_1(\boldsymbol{\theta}))]\!] \tag{6.8}$$
$$[\![(O, \rho) \to \frac{\partial}{\partial\theta}(S_0(\boldsymbol{\theta})); S_1(\boldsymbol{\theta})]\!] = [\![([\![S_1(\boldsymbol{\theta})]\!]^*(O), \rho) \to \frac{\partial}{\partial\theta}(S_0(\boldsymbol{\theta}))]\!], \tag{6.9}$$
where we make critical use of the fact $S_0(\boldsymbol{\theta}), S_1(\boldsymbol{\theta}) \in \mathbf{q\text{-}while}^{(T)}_{\overline{v}}(\boldsymbol{\theta})$ and $\frac{\partial}{\partial\theta}(S_0(\boldsymbol{\theta})), \frac{\partial}{\partial\theta}(S_1(\boldsymbol{\theta})) \in \mathbf{add\text{-}q\text{-}while}^{(T)}_{\overline{v}\cup\{A\}}(\boldsymbol{\theta})$.
For (6.8), since $\frac{\partial}{\partial\theta}(S_1(\boldsymbol{\theta}))$ computes the derivative for any input state and observable, we choose the input state $[\![S_0(\boldsymbol{\theta})]\!](\rho)$ and observable O.
For (6.9), we don't change the state ρ but change the observable O by applying the *dual* super-operator $[\![S_1(\boldsymbol{\theta})]\!]^*$. Since $\frac{\partial}{\partial\theta}(S_0(\boldsymbol{\theta}))$ computes the derivative for

any input state and any observable, we choose the input state ρ and observable $[\![S_1(\boldsymbol{\theta})]\!]^*(O)$.
The dual super-operator $[\![S_1(\boldsymbol{\theta})]\!]^*$ has the property that $\mathrm{tr}(O[\![S_1(\boldsymbol{\theta})]\!](\rho)) = \mathrm{tr}([\![S_1(\boldsymbol{\theta})]\!]^*(O)\rho)$, which corresponds to the Schrodinger picture (evolving states) and Heisenberg picture (evolving observables) respectively in quantum mechanics.

6. The proof of the *Case* rule basically follows from the linearity of the observable semantics and the smooth semantics of Case. It is interesting to compare with the classical case [5] where the non-smoothness of the guard causes an issue for auto differentiation.

Example 6.1 (Simple-Case)**.** *Consider the following simple instantiating of Ex. 4.1*

$$P(\theta)|q_1\rangle \equiv \textbf{case } M[q_1] = 0 \to R_X(\theta)|q_1\rangle; R_Y(\theta)|q_1\rangle,\quad 1 \to R_Z(\theta)|q_1\rangle$$

Let us apply code transformation and compilation. Let CT, CP to denote "code transformation" and "compilation", and "Seq" and "Rot" denote Sequence and Rotation rules resp.

$$\begin{aligned}
\frac{\partial}{\partial\theta}(\underline{P(\theta)}) &\overset{CT,\textbf{case}}{=} \textbf{case } M[q_1] = 0 \to \tfrac{\partial}{\partial\theta}(R_X(\theta)|q_1\rangle; R_Y(\theta)|q_1\rangle),\ 1 \to \tfrac{\partial}{\partial\theta}(R_Z(\theta)|q_1\rangle)\\
&\overset{CT,Seq+Rot}{=} \textbf{case } M[q_1] = 0 \to (R'_X(\theta)|A,q_1\rangle; R_Y(\theta)|q_1\rangle) + (R_X(\theta)|q_1\rangle; R'_Y(\theta)|A,q_1\rangle),\ 1 \to R'_Z(\theta)|A,q_1\rangle\\
&\overset{\texttt{Compile}(\bullet)}{\longmapsto} \left\{\!\!\left|\ \textbf{case } M[q_1] = 0 \to R'_X(\theta)|A,q_1\rangle; R_Y(\theta)|q_1\rangle,\ 1 \to R'_Z(\theta)|A,q_1\rangle,\right.\right.\\
&\qquad\left.\left.\ \textbf{case } M[q_1] = 0 \to R_X(\theta)|q_1\rangle; R'_Y(\theta)|A,q_1\rangle,\ |\ 1 \to \textbf{abort}.\right|\!\!\right\}
\end{aligned}$$

7 Execution and Resource Analysis

In this section we illustrate the execution of the entire differentiation procedure and analyze its resource cost. Consider any program $P(\boldsymbol{\theta}) \in \textbf{q-while}^{(T)}_{v}(\boldsymbol{\theta})$ and the parameter θ.

Execution. The first step in differentiation is to apply the code transformation rules (in Section 6) to $P(\boldsymbol{\theta})$ and obtain an additive program $\frac{\partial}{\partial\theta}(P(\boldsymbol{\theta}))$. Then one needs to compile $\frac{\partial}{\partial\theta}(P(\boldsymbol{\theta}))$ into a multiset $\{|P'_i(\boldsymbol{\theta})|\}_{i=1}^m$ of normal non-aborting programs $P'_i(\boldsymbol{\theta})$. The total count of these programs is given by $m = |\#\frac{\partial}{\partial\theta}(P(\boldsymbol{\theta}))|$. Note that the above procedure could be done at the compilation time.

Given any pair of O and ρ, the real execution to compute the derivative of $[\![(O,\rho) \to P(\boldsymbol{\theta})]\!]$ is to approximate the observable semantics $[\![(O,\rho) \to \frac{\partial}{\partial\theta}(P(\boldsymbol{\theta}))]\!]$. By Definition 5.2, we need to approximate

$$\sum_{i=1}^{m} \mathrm{tr}\Big((Z_A \otimes O)[\![P'_i(\boldsymbol{\theta})]\!](((|\overline{0}\rangle_{\{A\}}\langle\overline{0}|) \otimes \rho)\Big), \tag{7.1}$$

where each term is the observable $Z_A \otimes O$ on the output state of $P'_i(\theta)$ given input state ρ and the ancilla qubit $|0\rangle$.

To approximate the sum in (7.1) to precision δ , one could first treat the sum divided by m as the observable applied on the program that starts with a uniformly random choice of i from $1, \cdots, m$ and then execute $P'_i(\boldsymbol{\theta})$. By Chernoff bound, one only needs to repeat this procedure $O(m^2/\delta^2)$ times.

Resource count. We are only interested in non-trivial (extra) resource that is something that you wouldn't need if you only run the original program. Ancilla qubits count as the non-trivial resource. However, for our scheme, the number of required ancilla is 1 qubit when applying $\frac{\partial}{\partial\theta}(\cdot)$ once.

The more non-trivial resource is the number of the copies of input state, which is directly related to the number of repetitions in the procedure, which again connects to $m = |\#\frac{\partial}{\partial\theta}(P(\boldsymbol{\theta}))|$. We argue that our code transformation is efficient so that m is reasonably bounded. To that end, we show the relation between m and a natural quantity defined on the original program $P(\boldsymbol{\theta})$ (i.e., before applying any $\frac{\partial}{\partial\theta}(\cdot)$ operator) called the *occurrence count* of the parameter θ.

Definition 7.1. *The "Occurrence Count for θ_j" in $P(\boldsymbol{\theta})$, denoted $\mathrm{OC}_j(P(\boldsymbol{\theta}))$, is defined as follows:*

1. *If $P(\boldsymbol{\theta}) \equiv$ **abort**$[\overline{v}]$|**skip**$[\overline{v}]|q := |0\rangle$ ($q \in \overline{v}$), then $\mathrm{OC}_j(P(\boldsymbol{\theta})) = 0$;*
2. *$P(\boldsymbol{\theta}) \equiv U(\boldsymbol{\theta})$: if $U(\boldsymbol{\theta})$ trivially uses θ_j, then $\mathrm{OC}_j(P(\boldsymbol{\theta})) = 0$; otherwise $\mathrm{OC}_j(P(\boldsymbol{\theta})) = 1$.*
3. *If $P(\boldsymbol{\theta}) \equiv U(\boldsymbol{\theta}) = P_1(\boldsymbol{\theta}); P_2(\boldsymbol{\theta}))$ then $\mathrm{OC}_j(P(\boldsymbol{\theta})) = \mathrm{OC}_j(P_1(\boldsymbol{\theta})) + \mathrm{OC}_j(P_2(\boldsymbol{\theta}))$.*
4. *If $P(\boldsymbol{\theta}) \equiv$ **case** $M[\overline{q}] = \overline{m \to P_m(\boldsymbol{\theta})}$ **end** then $\mathrm{OC}_j(P(\boldsymbol{\theta})) = \max_m \mathrm{OC}_j(P_m(\boldsymbol{\theta}))$.*
5. *If $P(\boldsymbol{\theta}) \equiv$ **while**$^{(T)}$ $M[\overline{q}] = 1$ **do** $P_1(\boldsymbol{\theta})$ **done** then $\mathrm{OC}_j(P(\boldsymbol{\theta})) = T \cdot \mathrm{OC}_j(P_1(\boldsymbol{\theta}))$.*

Intuition of the "Occurrence Count" definition is clear: it basically counts the non-trivial number of occurrence of θ_j in the program, treating **case** as if it is deterministic. To see why this is a reasonable quantity, consider the auto-differentiation in the classical case. For any non-trivial variable v (i.e., v has some dependence on the parameter θ), we will compute both v and $\frac{\partial}{\partial\theta}(v)$ and store them both as variables in the new program. Thus, the classical auto-differentiation essentially needs the number of non-trivial occurrences more space and related resources. As we argued in the introduction, we cannot directly mimic the classical case due to the non-cloning theorem. The extra space requirement in the classical setting turns into the requirement on the extra copies of the input state in the quantum setting. Indeed, we can bound m by the occurrence count.

Proposition 7.2. $|\#\frac{\partial}{\partial\theta_j}(P(\boldsymbol{\theta}))| \le \mathrm{OC}_j(P(\boldsymbol{\theta}))$.

Proof. Structural induction. For details, see Appendix E.1. □

8 Implementation and Case Study

We have built a compiler (written in OCaml) that implements our code transformation and compilation rules. We use it to train one VQC instance with controls and empirically verify its resource-efficiency on representative VQC instances. All codes and test programs are available in the supplementary materials. Complete details are in Appendix F. Experiments are performed on a MacBook Pro with a Dual-Core Intel Core i5 Processor clocked at 2.7 GHz, and 8GB of RAM.

8.1 Training VQC instances with controls

Consider a simple classification problem over 4-bit inputs $z = z_1z_2z_3z_4 \in \{0,1\}^4$ with true label given by $f(z) = \neg(z_1 \oplus z_4)$. We construct two 4-qubit VQCs P_1 (**no control**) and P_2 (**with control**) that consists of a single-qubit Pauli X,Y and Z rotation gate on each qubit and compare their performance.

For parameters $\Gamma = \{\gamma_1, \ldots, \gamma_{12}\}$ define the program

$$\begin{aligned} Q(\Gamma) \equiv\ & R_X(\gamma_1)[q_1]; R_X(\gamma_2)[q_2]; R_X(\gamma_3)[q_3]; R_X(\gamma_4)[q_4]; \\ & R_Y(\gamma_5)[q_1]; R_Y(\gamma_6)[q_2]; R_Y(\gamma_7)[q_3]; R_Y(\gamma_8)[q_4]; \\ & R_Z(\gamma_9)[q_1]; R_Z(\gamma_{10})[q_2]; R_Z(\gamma_{11})[q_3]; R_Z(\gamma_{12})[q_4], \end{aligned}$$

where q_1, q_2, q_3, q_4 refer to 4 qubit registers. Given parameters $\Theta = \{\theta_1, \ldots, \theta_{12}\}$, $\Phi = \{\phi_1, \ldots, \phi_{12}\}$, define

$$P_1(\Theta, \Phi) \equiv Q(\Theta); Q(\Phi). \tag{8.1}$$

Similarly, for parameters $\Theta = \{\theta_1, \ldots, \theta_{12}\}$,$\Phi = \{\phi_1, \ldots, \phi_{12}\}$, $\Psi = \{\psi_1, \ldots, \psi_{12}\}$, define

$$\begin{aligned} P_2(\Theta, \Phi, \Psi) \equiv Q(\Theta); \textbf{case } M[q_1] = 0 &\to Q(\Phi) \\ 1 &\to Q(\Psi). \end{aligned} \tag{8.2}$$

Note that P_1 and P_2 execute the same number of gates for each run. To use P_i to perform the classification or in the training, we first initialize q_1, q_2, q_3, q_4 to the classical feature vector $z = z_1z_2z_3z_4$ and then execute P_i. The predicted label y is given by measuring the 4^{th} qubit q_4 in 0/1 basis.

We will conduct a supervised learning by minimizing the following *loss* function given by the average negative log likelihood of correctly classifying the inputs. Thus

$$\text{loss} = -\frac{1}{16} \sum_{z \in \{0,1\}^4} \log\left(\Pr[y = f(z) = \neg(z_1 \oplus z_4)]\right). \tag{8.3}$$

Note that loss is a function on Θ, Φ (or $\Theta, \Phi.\Psi$). More importantly, for each z, $\Pr[y = \neg(z_1 \oplus z_4)]$ can be represented by the observable semantics of P_1(or P_2) with observable $|0\rangle\langle 0|$ or $|1\rangle\langle 1|$. Thus, the gradient of loss can be obtained by using the collection of $\frac{\partial}{\partial \alpha}(P_1)$ for $\alpha \in \Theta, \Phi$ (or $\frac{\partial}{\partial \alpha}(P_2)$ for $\alpha \in \Theta, \Phi, \Psi$). We classically simulate the above training procedure with gradient descent[7] (see Figure 6). After 1000 epochs with manually optimized hyperparameters, the loss for P_1 (**no control**) attains a minimum of 0.6931 in less than 100 epochs and subsequently plateaus. The loss for P_2 (**with control**) continues to decrease and attains a minimum of 0.0936. It demonstrates the advantage of both controls in quantum machine learning and our scheme to handle controls, whereas previous schemes (such as Pennylane due to its quantum-node design [7]) fail to do so.

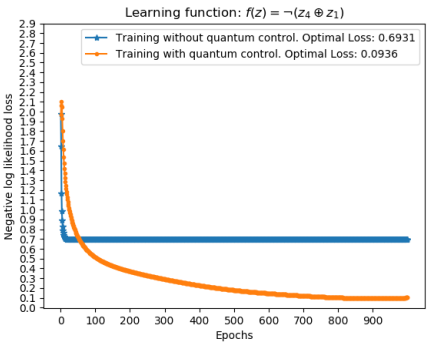

Figure 6. Training P_1 and P_2 to classify inputs according to the labelling function $f(z) = \neg(z_1 \oplus z_4)$.

8.2 Benchmark testing on representative VQCs

We also test our compiler on important VQC candidates such as quantum neural-networks (QNN) [16], quantum approximate optimization algorithms (QAOA) [15], and variational quantum eigensolver (VQE) [35]. We enrich these examples, by adding simple controls or 2-bounded loops and increasing the number of qubits to 18~40, to make them sufficiently sophisticated but yet realistic for near-term quantum applications. A selective output performance of our compiler is in Table 2, with full details in Appendix F. Our scheme is resource-efficient as $|\#\frac{\partial}{\partial\theta}(\cdot)|$ is always reasonably bounded.

$P(\boldsymbol{\theta})$	OC(·)	$\lvert\#\frac{\partial}{\partial\theta}(\cdot)\rvert$	#gates	#lines	#layers	#qb's
$\text{QNN}_{M,i}$	24	24	165	189	3	18
$\text{QNN}_{M,w}$	56	24	231	121	5	18
$\text{QNN}_{L,i}$	48	48	363	414	6	36
$\text{QNN}_{L,w}$	504	48	2079	244	33	36
$\text{VQE}_{M,i}$	15	15	224	241	3	12
$\text{VQE}_{M,w}$	35	15	224	112	5	12
$\text{VQE}_{L,i}$	40	40	576	628	5	40
$\text{VQE}_{L,w}$	248	40	1984	368	17	40
$\text{QAOA}_{M,i}$	18	18	120	142	3	18
$\text{QAOA}_{M,w}$	42	18	168	94	5	18
$\text{QAOA}_{L,i}$	36	36	264	315	6	36
$\text{QAOA}_{L,w}$	378	36	1512	190	33	36

Table 2. Output on selective examples. $\{M, L\}$ stands for "medium, large"; $\{i, w\}$ stands for including "if, while".

[7] We attempted to directly use Pennylane to train the system without control, but were not able to make it work since Pennylane does not implement the gradient of the log function.

Acknowledgments

SZ, SH, and XW were supported in part by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Quantum Testbed Pathfinder Program under Award Number DE-SC0019040. SC and XW were supported in part by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Quantum Algorithms Teams program. XW also received support from the National Science Foundation (grants CCF-1755800 and CCF-1816695).

References

[1] Ali Javadi Abhari, Arvin Faruque, Mohammad Javad Dousti, Lukas Svec, Oana Catu, Amlan Chakrabati, Chen-Fu Chiang, Seth Vanderwilt, John Black, Fred Chong, Margaret Martonosi, Martin Suchara, Ken Brown, Massoud Pedram, and Todd Brun. 2012. *Scaffold: Quantum Programming Language.* Technical Report TR-934-12. Princeton University.

[2] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C. Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando G. S. L. Brandao, David A. Buell, Brian Burkett, Yu Chen, Zijun Chen, Ben Chiaro, Roberto Collins, William Courtney, Andrew Dunsworth, Edward Farhi, Brooks Foxen, Austin Fowler, Craig Gidney, Marissa Giustina, Rob Graff, Keith Guerin, Steve Habegger, Matthew P. Harrigan, Michael J. Hartmann, Alan Ho, Markus Hoffmann, Trent Huang, Travis S. Humble, Sergei V. Isakov, Evan Jeffrey, Zhang Jiang, Dvir Kafri, Kostyantyn Kechedzhi, Julian Kelly, Paul V. Klimov, Sergey Knysh, Alexander Korotkov, Fedor Kostritsa, David Landhuis, Mike Lindmark, Erik Lucero, Dmitry Lyakh, Salvatore Mandrà, Jarrod R. McClean, Matthew McEwen, Anthony Megrant, Xiao Mi, Kristel Michielsen, Masoud Mohseni, Josh Mutus, Ofer Naaman, Matthew Neeley, Charles Neill, Murphy Yuezhen Niu, Eric Ostby, Andre Petukhov, John C. Platt, Chris Quintana, Eleanor G. Rieffel, Pedram Roushan, Nicholas C. Rubin, Daniel Sank, Kevin J. Satzinger, Vadim Smelyanskiy, Kevin J. Sung, Matthew D. Trevithick, Amit Vainsencher, Benjamin Villalonga, Theodore White, Z. Jamie Yao, Ping Yeh, Adam Zalcman, Hartmut Neven, and John M. Martinis. 2019. Quantum supremacy using a programmable superconducting processor. *Nature* 574, 7779 (2019), 505–510.

[3] Alexandru Baltag and Sonja Smets. 2011. Quantum Logic as a Dynamic Logic. *Synthese* 179, 2 (2011).

[4] Atılım Günes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. 2017. Automatic Differentiation in Machine Learning: A Survey. *J. Mach. Learn. Res.* 18, 1 (Jan. 2017), 5595–5637. http://dl.acm.org/citation.cfm?id=3122009.3242010

[5] Thomas Beck and Herbert Fischer. 1994. The if-problem in automatic differentiation. *J. Comput. Appl. Math.* 50, 1-3 (1994), 119–131.

[6] Marcello Benedetti, Erika Lloyd, and Stefan Sack. 2019. Parameterized quantum circuits as machine learning models. *arXiv e-prints*, Article arXiv:1906.07682 (Jun 2019), arXiv:1906.07682 pages. arXiv:quant-ph/1906.07682

[7] Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, and Nathan Killoran. 2018. PennyLane: Automatic differentiation of hybrid quantum-classical computations. *arXiv preprint arXiv:1811.04968* (2018).

[8] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. 2017. Quantum machine learning. *Nature* 549, 7671 (2017), 195. arXiv:arXiv:1611.09347

[9] Olivier Brunet and Philippe Jorrand. 2004. Dynamic Quantum Logic for Quantum Programs. *International Journal of Quantum Information* 2, 1 (2004).

[10] Rohit Chadha, Paulo Mateus, and Amílcar Sernadas. 2006. Reasoning About Imperative Quantum Programs. *Electronic Notes in Theoretical Computer Science* 158 (2006).

[11] George Corliss, Christèle Faure, Andreas Griewank, Lauren Hascoët, and Uwe Naumann (Eds.). 2002. *Automatic Differentiation of Algorithms: From Simulation to Optimization.* Springer-Verlag New York, Inc., New York, NY, USA.

[12] Thomas Ehrhard and Laurent Regnier. 2003. The differential lambda-calculus. *Theoretical Computer Science* 309, 1-3 (2003), 1–41.

[13] Conal M. Elliott. 2009. Beautiful Differentiation. In *Proceedings of the 14th ACM SIGPLAN International Conference on Functional Programming (ICFP '09).* ACM, New York, NY, USA, 191–202. https://doi.org/10.1145/1596550.1596579

[14] Conal M. Elliott. 2018. The Simple Essence of Automatic Differentiation. *Proc. ACM Program. Lang.* 2, ICFP, Article 70 (July 2018), 29 pages. https://doi.org/10.1145/3236765

[15] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. 2014. A Quantum Approximate Optimization Algorithm. (2014). arXiv:arXiv:1411.4028

[16] Edward Farhi and Hartmut Neven. 2018. Classification with Quantum Neural Networks on Near Term Processors. *arXiv e-prints*, Article arXiv:1802.06002 (Feb 2018), arXiv:1802.06002 pages. arXiv:quant-ph/1802.06002

[17] Yuan Feng, Runyao Duan, Zhengfeng Ji, and Mingsheng Ying. 2007. Proof Rules for the Correctness of Quantum Programs. *Theoretical Computer Science* 386, 1-2 (2007).

[18] Simon J. Gay. 2006. Quantum Programming Languages: Survey and Bibliography. *Mathematical Structures in Computer Science* 16, 4 (2006).

[19] Gian Giacomo Guerreschi and Mikhail Smelyanskiy. 2017. Practical optimization for hybrid quantum-classical algorithms. *arXiv e-prints*, Article arXiv:1701.01450 (Jan 2017), arXiv:1701.01450 pages. arXiv:quant-ph/1701.01450

[20] Martin Giles. 2019. IBM's new 53-qubit quantum computer is the most powerful machine you can use. https://www.technologyreview.com/f/614346/ibms-new-53-qubit-quantum-computer-is-the-most-powerful-machine-you-can-use/

[21] Pranav Gokhale. 2018. Variational Quantum Eigensolver Demo. *ISCA 2018* (2018).

[22] Jonathan Grattage. 2005. A Functional Quantum Programming Language. In *LICS*.

[23] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, AdriàPuigdomènech Badia, Karl Moritz Hermann, Yori Zwols, Georg Ostrovski, Adam Cain, Helen King, Christopher Summerfield, Phil Blunsom, Koray Kavukcuoglu, and Demis Hassabis. 2016. Hybrid computing using a neural network with dynamic external memory. *Nature* 538 (10 2016), 471.

[24] Alexander S. Green, Peter LeFanu Lumsdaine, Neil J. Ross, Peter Selinger, and Benoît Valiron. 2013. Quipper: A Scalable Quantum Programming Language. In *PLDI*.

[25] Edward Grefenstette, Karl Moritz Hermann, Mustafa Suleyman, and Phil Blunsom. 2015. Learning to Transduce with Unbounded Memory. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'15).* MIT Press, Cambridge, MA, USA, 1828–1836. http://dl.acm.org/citation.cfm?id=2969442.2969444

[26] Andreas Griewank. 2000. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

[27] Yoshihiko Kakutani. 2009. A Logic for Formal Verification of Quantum Programs. In *ASIAN*.

[28] Gershon Kedem. 1980. Automatic Differentiation of Computer Programs. *ACM Trans. Math. Softw.* 6, 2 (June 1980), 150–165. https://doi.org/10.1145/355887.355890

[29] Jin-Guo Liu and Lei Wang. 2018. Differentiable learning of quantum circuit Born machines. *Phys. Rev. A* 98 (Dec 2018), 062324. Issue 6. https://doi.org/10.1103/PhysRevA.98.062324

[30] Nikolaj Moll, Panagiotis Barkoutsos, Lev S. Bishop, Jerry M. Chow, Andrew Cross, Daniel J. Egger, Stefan Filipp, Andreas Fuhrer, Jay M. Gambetta, Marc Ganzhorn, Abhinav Kandala, Antonio Mezzacapo, Peter Muller, Walter Riess, Gian Salis, John Smolin, Ivano Tavernelli, and Kristan Temme. 2018. Quantum optimization using variational algorithms on near-term quantum devices. *Quantum Science and Technology* 3, 3 (2018), 030503. arXiv:arXiv:1710.01022

[31] Michael A. Nielsen and Isaac Chuang. 2000. *Quantum Computation and Quantum Information*. Cambridge University Press.

[32] Bernhard Ömer. 2003. *Structured Quantum Programming*. Ph.D. Dissertation. Vienna University of Technology.

[33] Jennifer Paykin, Robert Rand, and Steve Zdancewic. 2017. QWIRE: A Core Language for Quantum Circuits. In *POPL*.

[34] Barak A. Pearlmutter and Jeffrey Mark Siskind. 2008. Reverse-mode AD in a Functional Framework: Lambda the Ultimate Backpropagator. *ACM Trans. Program. Lang. Syst.* 30, 2, Article 7 (March 2008), 36 pages. https://doi.org/10.1145/1330017.1330018

[35] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, Alán Aspuru-Guzik, and Jeremy L. O'brien. 2014. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications* 5 (2014), 4213. arXiv:arXiv:1304.3061

[36] Gordon Plotkin. 2018. Some Principles of Differential Programming Languages. *POPL 2018* (2018).

[37] John Preskill. 2018. Quantum computing in the NISQ era and beyond. *Quantum* 2 (2018), 79. arXiv:arXiv:1801.00862

[38] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature* 323, 6088 (1986), 533–536.

[39] Amr Sabry. 2003. Modeling Quantum Computing in Haskell. In *The Haskell Workshop*.

[40] Jeff W. Sanders and Paolo Zuliani. 2000. Quantum Programming. In *MPC*.

[41] Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. 2018. Evaluating analytic gradients on quantum hardware. *arXiv preprint arXiv:1811.11184* (2018).

[42] Maria Schuld, Alex Bocharov, Krysta Svore, and Nathan Wiebe. 2018. Circuit-centric quantum classifiers. *arXiv e-prints*, Article arXiv:1804.00633 (Apr 2018), arXiv:1804.00633 pages. arXiv:quant-ph/1804.00633

[43] Peter Selinger. 2004. A Brief Survey of Quantum Programming Languages. In *FLOPS*.

[44] Peter Selinger. 2004. Towards a Quantum Programming Language. *Mathematical Structures in Computer Science* 14, 4 (2004).

[45] Bert Speelpenning. 1980. *Compiling Fast Partial Derivatives of Functions Given by Algorithms*. Ph.D. Dissertation. Champaign, IL, USA. AAI8017989.

[46] Krysta Svore, Alan Geller, Matthias Troyer, John Azariah, Christopher Granade, Bettina Heim, Vadym Kliuchnikov, Mariia Mykhailova, Andres Paz, and Martin Roetteler. 2018. Q#: Enabling Scalable Quantum Computing and Development with a High-level DSL. In *RWDSL*.

[47] Qingfeng Wang and Tauqir Abdullah. 2018. An Introduction to Quantum Optimization Approximation Algorithm.

[48] John Watrous. 2006. Introduction to Quantum Computation. https://cs.uwaterloo.ca/~watrous/LectureNotes/CPSC519.Winter2006/all.pdf. Course notes.

[49] Dave Wecker and Krysta Svore. 2014. LIQUi|⟩: A Software Design Architecture and Domain-Specific Language for Quantum Computing. *CoRR* abs/1402.4467 (2014). arXiv:1402.4467

[50] Robert Edwin Wengert. 1964. A Simple Automatic Derivative Evaluation Program. *Commun. ACM* 7, 8 (Aug. 1964), 463–464. https://doi.org/10.1145/355586.364791

[51] William K. Wootters and Wojciech H. Zurek. 1982. A single quantum cannot be cloned. *Nature* 299, 5886 (1982), 802–803.

[52] Mingsheng Ying. 2011. Floyd–Hoare Logic for Quantum Programs. *ACM Transactions on Programming Languages and Systems* 33, 6 (2011).

[53] Mingsheng Ying. 2016. *Foundations of Quantum Programming*. Morgan Kaufmann.

[54] Mingsheng Ying and Yangjia Li. 2018. Reasoning about Parallel Quantum Programs. *arXiv preprint arXiv:1810.11334* (2018).

[55] Mingsheng Ying, Shenggang Ying, and Xiaodi Wu. 2017. Invariants of Quantum Programs: Characterisations and Generation. In *POPL*.

[56] Daiwei Zhu, NM Linke, Marcello Benedetti, KA Landsman, NH Nguyen, CH Alderete, Alejandro Perdomo-Ortiz, Nathan Korda, A Garfoot, Charles Brecque, et al. 2018. Training of quantum circuits on a hybrid quantum computer. *arXiv preprint arXiv:1812.08862* (2018).

A Detailed Quantum Preliminary

This is a more detailed treatment of Section 2. For a further extended background, we recommend the notes by Watrous [48] and the textbook by Nielsen and Chuang [31].

A.1 Preliminaries

For any non-negative integer n, an n-dimensional Hilbert space $\mathcal{H}$ is essentially the space $\mathbb{C}^n$ of complex vectors. We use Dirac's notation, $|\psi\rangle$, to denote a complex vector in $\mathbb{C}^n$. The inner product of two vectors $|\psi\rangle$ and $|\phi\rangle$ is denoted by $\langle\psi|\phi\rangle$, which is the product of the Hermitian conjugate of $|\psi\rangle$, denoted by $\langle\psi|$, and vector $|\phi\rangle$. The norm of a vector $|\psi\rangle$ is denoted by $\||\psi\rangle\| = \sqrt{\langle\psi|\psi\rangle}$.

We define (linear) *operators* as linear mappings between Hilbert spaces. Operators between n-dimensional Hilbert spaces are represented by $n \times n$ matrices. For example, the identity operator $I_{\mathcal{H}}$ can be identified by the identity matrix on $\mathcal{H}$. The Hermitian conjugate of operator A is denoted by $A^\dagger$. Operator A is *Hermitian* if $A = A^\dagger$. The trace of an operator A is the sum of the entries on the main diagonal, i.e., $\mathrm{tr}(A) = \sum_i A_{ii}$. We write $\langle\psi|A|\psi\rangle$ to mean the inner product between $|\psi\rangle$ and $A|\psi\rangle$. A Hermitian operator A is *positive semidefinite* (resp., *positive definite*) if for all vectors $|\psi\rangle \in \mathcal{H}$, $\langle\psi|A|\psi\rangle \geq 0$ (resp., > 0). This gives rise to the *Löwner order* $\sqsubseteq$ among operators:

$A \sqsubseteq B$ if $B - A$ is positive semidefinite,
$A \sqsubset B$ if $B - A$ is positive definite.

A.2 Quantum States

The state space of a quantum system is a Hilbert space. The state space of a *qubit*, or quantum bit, is a 2-dimensional Hilbert space. One important orthonormal basis of a qubit system is the *computational* basis with $|0\rangle = (1,0)^\dagger$ and $|1\rangle = (0,1)^\dagger$, which encode the classical bits 0 and 1 respectively. Another important basis, called the $\pm$ basis, consists of $|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$ and $|-\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$. The state space of multiple qubits is the *tensor product* of single qubit state spaces. For example, classical 00 can be encoded by $|0\rangle \otimes |0\rangle$ (written $|0\rangle|0\rangle$ or even $|00\rangle$ for short) in the Hilbert space $\mathbb{C}^2 \otimes \mathbb{C}^2$. An important 2-qubit state is the EPR state $|\beta_{00}\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$. The Hilbert space for an m-qubit system is $(\mathbb{C}^2)^{\otimes m} \cong \mathbb{C}^{2^m}$.

A *pure* quantum state is represented by a unit vector, i.e., a vector $|\psi\rangle$ with $\||\psi\rangle\| = 1$. A *mixed* state can be represented by a classical distribution over an ensemble of pure states $\{(p_i, |\psi_i\rangle)\}_i$, i.e., the system is in state $|\psi_i\rangle$ with probability p_i. One can also use *density operators* to represent both pure and mixed quantum states. A density operator ρ for a mixed state representing the ensemble $\{(p_i, |\psi_i\rangle)\}_i$ is a positive semidefinite operator $\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|$, where $|\psi_i\rangle\langle\psi_i|$ is the outer-product of $|\psi_i\rangle$; in particular, a pure state $|\psi\rangle$ can be identified with the density operator $\rho = |\psi\rangle\langle\psi|$. Note that $\mathrm{tr}(\rho) = 1$ holds for all density operators. A positive semidefinite operator ρ on $\mathcal{H}$ is said to be a *partial* density operator if $\mathrm{tr}(\rho) \leq 1$. The set of partial density operators is denoted by $\mathcal{D}(\mathcal{H})$.

A.3 Quantum Operations

Operations on quantum systems can be characterized by unitary operators. Denoting the set of linear operators on $\mathcal{H}$ as $L(\mathcal{H})$, an operator $U \in L(\mathcal{H})$ is *unitary* if its Hermitian conjugate is its own inverse, i.e., $U^\dagger U = UU^\dagger = I_{\mathcal{H}}$. For a pure state $|\psi\rangle$, a unitary operator describes an *evolution* from $|\psi\rangle$ to $U|\psi\rangle$. For a density operator ρ, the corresponding evolution is $\rho \mapsto U\rho U^\dagger$. Common unitary operators include

$$H = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix},\ X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},\ Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},\ Y = -iXZ$$

The *Hadamard* operator H transforms between the computational and the $\pm$ basis. For example, $H|0\rangle = |+\rangle$ and $H|1\rangle = |-\rangle$. The *Pauli* X operator is a bit flip, i.e., $X|0\rangle = |1\rangle$ and $X|1\rangle = |0\rangle$. The *Pauli* Z operator is a phase flip, i.e., $Z|0\rangle = |0\rangle$ and $Z|1\rangle = -|1\rangle$. *Pauli* Y maps $|0\rangle$ to $i|1\rangle$ and $|1\rangle$ to $-i|0\rangle$. The *CNOT* gate C maps $|00\rangle \mapsto |00\rangle, |01\rangle \mapsto |01\rangle, |10\rangle \mapsto |11\rangle, |11\rangle \mapsto |10\rangle$. One may obtain the EPR state $|\beta_{00}\rangle$ via $|00\rangle \overset{H_1}{\mapsto} \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)|0\rangle \overset{C_{1,2}}{\mapsto} \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$.

More generally, the evolution of a quantum system can be characterized by an *admissible superoperator* $\mathcal{E}$, which is a *completely-positive* and *trace-non-increasing* linear map from $\mathcal{D}(\mathcal{H})$ to $\mathcal{D}(\mathcal{H}')$ for Hilbert spaces $\mathcal{H}, \mathcal{H}'$. A superoperator is positive if it maps from $\mathcal{D}(\mathcal{H})$ to $\mathcal{D}(\mathcal{H}')$ for Hilbert spaces $\mathcal{H}, \mathcal{H}'$. A superoperator $\mathcal{E}$ is k-positive if for any k-dimensional Hilbert space $\mathcal{A}$, the superoperator $\mathcal{E} \otimes I_{\mathcal{A}}$ is a positive map on $\mathcal{D}(\mathcal{H} \otimes \mathcal{A})$. A superoperator is said to be completely positive if it is k-positive for any positive integer k. A superoperator $\mathcal{E}$ is trace-non-increasing if for any initial state $\rho \in \mathcal{D}(\mathcal{H})$, the final state $\mathcal{E}(\rho) \in \mathcal{D}(\mathcal{H}')$ after applying $\mathcal{E}$ satisfies $\mathrm{tr}(\mathcal{E}(\rho)) \leq \mathrm{tr}(\rho)$.

For every superoperator $\mathcal{E} : \mathcal{D}(\mathcal{H}) \to \mathcal{D}(\mathcal{H}')$, there exists a set of Kraus operators $\{E_k\}_k$ such that $\mathcal{E}(\rho) = \sum_k E_k \rho E_k^\dagger$ for any input $\rho \in \mathcal{D}(\mathcal{H})$. Note that the set of Kraus operators is finite if the Hilbert space is finite-dimensional. The *Kraus form* of $\mathcal{E}$ is written as $\mathcal{E} = \sum_k E_k \circ E_k^\dagger$. A unitary evolution can be represented by the superoperator $\mathcal{E} = U \circ U^\dagger$. An identity operation refers to the superoperator $\mathcal{I}_{\mathcal{H}} = I_{\mathcal{H}} \circ I_{\mathcal{H}}$. The Schrödinger-Heisenberg *dual* of a superoperator $\mathcal{E} = \sum_k E_k \circ E_k^\dagger$, denoted by $\mathcal{E}^*$, is defined as follows: for every state $\rho \in \mathcal{D}(\mathcal{H})$ and any operator A, $\mathrm{tr}(A\mathcal{E}(\rho)) = \mathrm{tr}(\mathcal{E}^*(A)\rho)$. The Kraus form of $\mathcal{E}^*$ is $\sum_k E_k^\dagger \circ E_k$.

A.4 Quantum Measurements and Observables

The way to extract information about a quantum system is called a quantum *measurement*. A quantum measurement on a system over Hilbert space $\mathcal{H}$ can be described by a set of linear operators $\{M_m\}_m$ with $\sum_m M_m^\dagger M_m = I_{\mathcal{H}}$. If we perform a

measurement $\{M_m\}_m$ on a state ρ, the outcome m is observed with probability $p_m = \mathrm{tr}(M_m \rho M_m^\dagger)$ for each m. A major difference between classical and quantum computation is that a quantum measurement changes the state. In particular, after a measurement yielding outcome m, the state collapses to $M_m \rho M_m^\dagger / p_m$. For example, a measurement in the computational basis is described by $M = \{M_0 = |0\rangle\langle 0|, M_1 = |1\rangle\langle 1|\}$. If we perform the computational basis measurement M on state $\rho = |+\rangle\langle +|$, then with probability $\frac{1}{2}$ the outcome is 0 and ρ becomes $|0\rangle\langle 0|$. With probability $\frac{1}{2}$ the outcome is 1 and ρ becomes $|1\rangle\langle 1|$.

B More on the definition of parameterized quantum programs

B.1 Definition of qVar

Given $P(\boldsymbol{\theta})$ a T-bounded k-parameterized quantum **while**-program, let $qVar(P(\boldsymbol{\theta}))$, read *the set of quantum variables accesible to* P*, be recursively defined as follows* [53]:

1. If $P(\boldsymbol{\theta}) \equiv \mathbf{skip}[\overline{q}]$, $\mathbf{abort}[\overline{q}]$ or $U[\overline{q}]$, then $qVar(P(\boldsymbol{\theta})) = \overline{q}$.
2. If $P(\boldsymbol{\theta}) \equiv q := |0\rangle$, then $qVar(P(\boldsymbol{\theta})) = q$.
3. If $P(\boldsymbol{\theta}) \equiv P_1(\boldsymbol{\theta}); P_2(\boldsymbol{\theta})$, then $qVar(P(\boldsymbol{\theta})) = qVar(P_1(\boldsymbol{\theta})) \cup qVar(P_2(\boldsymbol{\theta}))$. When analyzing $P_1(\boldsymbol{\theta}); P_2(\boldsymbol{\theta})$, we identify $P_1(\boldsymbol{\theta})$ with $\mathcal{I} \otimes P_1(\boldsymbol{\theta})$, where $\mathcal{I} \equiv I_{qVar(P_2(\boldsymbol{\theta}))\setminus qVar(P_1(\boldsymbol{\theta}))} \circ I_{qVar(P_2(\boldsymbol{\theta}))\setminus qVar(P_1(\boldsymbol{\theta}))}$, the identity operation on the variables where P_1 originally has no access to.
4. If $P(\boldsymbol{\theta}) \equiv \mathbf{case}\, M[\overline{q}] = \overline{m \to P_m(\boldsymbol{\theta})}\, \mathbf{end}$, then $qVar(P(\boldsymbol{\theta})) = \overline{q} \cup \bigcup_m qVar(P_i(\boldsymbol{\theta}))$. We make the same identification for $P_i(\boldsymbol{\theta})$'s as the above.
5. If $P(\boldsymbol{\theta}) \equiv \mathbf{while}^{(T)}\ M[\overline{q}] = 1\ \mathbf{do}\ P_1(\boldsymbol{\theta})\ \mathbf{done}$, then $qVar(P(\boldsymbol{\theta})) = \overline{q} \cup qVar(P_1(\boldsymbol{\theta}))$. We make the same identification for $P_1(\boldsymbol{\theta})$ as the above.

One defines $qVar(P)$ for unparameterized P analogously.

C Detailed Proof from Section 4

C.1 Proof of Prop 4.2: Non-deterministic Compilation Rules Well-Defined

Proof. Structural induction.

1. (Atomic) and (Sequence $\downarrow$): as operational semantics of atomic operations are inherited from parameterized quantum while programs, they do not induce non-determinism. For these operations, $[\![\underline{P(\boldsymbol{\theta}^*)}]\!]\rho = \{|[\![\underline{P(\boldsymbol{\theta}^*)}]\!]\rho|\}$ as is the right hand side. (Sequence $\downarrow$) follows from the fact that $\downarrow; \underline{P(\boldsymbol{\theta})} \equiv \underline{P(\boldsymbol{\theta})}$.
2. (Sequence) If one of $\mathtt{Compile}(\underline{P_b(\boldsymbol{\theta}^*)})$ is $\{\mathbf{abort}\}$ the statement is immediately true, so we assume otherwise for below.
   a. $\subseteq$: Let $\rho' \in \{|\rho' \neq \mathbf{0} : \langle \underline{P_1(\boldsymbol{\theta}^*); P_2(\boldsymbol{\theta}^*)}, \rho\rangle \to^* \langle\downarrow, \rho'\rangle|\}$. By definition there exists $\rho_2 \neq \mathbf{0}$ s.t. $\langle \underline{P_1(\boldsymbol{\theta}^*)}; \underline{P_2(\boldsymbol{\theta}^*)}, \rho\rangle \to^* \langle P_2(\boldsymbol{\theta}^*), \rho_2\rangle$ and $\langle \underline{P_2(\boldsymbol{\theta}^*)}, \rho_2\rangle \to^* \langle\downarrow, \rho'\rangle$. By inductive hypothesis, $\rho' \in \coprod_{Q_j(\boldsymbol{\theta}) \in \mathtt{Compile}(\underline{P_2(\boldsymbol{\theta})})} \{|\rho' \neq \mathbf{0} : \langle Q_j(\boldsymbol{\theta}^*), \rho_2\rangle \to^* \langle\downarrow, \rho'\rangle|\}$.
   Since $\langle \underline{P_1(\boldsymbol{\theta}^*); P_2(\boldsymbol{\theta}^*)}, \rho\rangle \to^* \langle P_2(\boldsymbol{\theta}^*), \rho_2\rangle$, there must be some $\langle \underline{P_3(\boldsymbol{\theta}^*)}, \rho_3\rangle \to \langle\downarrow, \rho_2\rangle$ triggering the last transition according to our operational semantics. Inductively apply such an argument, knowing that all computation paths are finitely long,[8] we conclude that $\rho_2 \in \{|\rho_2' : \langle \underline{P_1(\boldsymbol{\theta})}, \rho\rangle \to^* \langle\downarrow, \rho_2'\rangle|\}$. By the inductive hypothesis we have $\rho_2 \in \coprod_{R_i(\boldsymbol{\theta}) \in \mathtt{Compile}(\underline{P_1(\boldsymbol{\theta})})} \{|\rho_2' \neq 0 : \langle R_i(\boldsymbol{\theta}^*), \rho\rangle \to^* \langle\downarrow, \rho_2'\rangle|\}$. Thus

$$\rho' \in \coprod_{R_i(\boldsymbol{\theta}) \in \mathtt{Compile}(\underline{P_1(\boldsymbol{\theta})}), Q_j(\boldsymbol{\theta}) \in \mathtt{Compile}(\underline{P_2(\boldsymbol{\theta})})} \{|\rho' \neq 0 : \langle R_i(\boldsymbol{\theta}^*), \rho\rangle \to^* \langle\downarrow, \rho_2\rangle \tag{C.1–C.3}$$
$$\bigwedge \langle Q_j(\boldsymbol{\theta}^*), \rho_2\rangle \to^* \langle\downarrow, \rho'\rangle|\}$$
$$= \coprod_{R_i(\boldsymbol{\theta}) \in \mathtt{Compile}(\underline{P_1(\boldsymbol{\theta})}), Q_j(\boldsymbol{\theta}) \in \mathtt{Compile}(\underline{P_2(\boldsymbol{\theta})})} \{|\rho' \neq 0 : \langle R_i(\boldsymbol{\theta}^*); Q_j(\boldsymbol{\theta}^*), \rho\rangle \to^* \langle\downarrow, \rho'\rangle|\} \tag{C.4–C.5}$$

   as desired.
   b. $\supseteq$: Let ρ' be a member of the multiset in right hand side ("RHS") of $C.5$ for some $R_i(\boldsymbol{\theta}) \in \mathtt{Compile}(\underline{P_1(\boldsymbol{\theta})})$, $Q_j(\boldsymbol{\theta}) \in \mathtt{Compile}(\underline{P_2(\boldsymbol{\theta})})$. Then by inductive hypothesis, $\langle \underline{P_2(\boldsymbol{\theta})}, [\![R_i(\boldsymbol{\theta}^*)]\!]\rho\rangle \to^* \langle\downarrow, \rho'\rangle$ ($\rho' \neq \mathbf{0}$), and $\langle \underline{P_1(\boldsymbol{\theta})}, \rho\rangle \to^* \langle\downarrow, [\![R_i(\boldsymbol{\theta}^*)]\!]\rho\rangle$ ($[\![R_i(\boldsymbol{\theta}^*)]\!]\rho \neq \mathbf{0}$). One may start with $\langle \underline{P_2(\boldsymbol{\theta}^*)}, [\![R_i(\boldsymbol{\theta}^*)]\!]\rho\rangle \to^* \langle\downarrow, \rho'\rangle$ then repeatedly apply the Sequential rule in the operational semantics, tracing back the path $\langle \underline{P_1(\boldsymbol{\theta})}, \rho\rangle \to^* \langle\downarrow, [\![R_i(\boldsymbol{\theta}^*)]\!]\rho\rangle$, and conclude that $\langle \underline{P_1(\boldsymbol{\theta}^*)}; \underline{P_2(\boldsymbol{\theta}^*)}, \rho\rangle \to^* \langle\downarrow, \rho'\rangle$ ($\rho' \neq 0$).
   The above said the two multisets have the same set of distinct elements, and furthermore each copy of $\rho' \in [\![\underline{P(\boldsymbol{\theta}^*)}]\!]\rho$ induces at least (by $\subseteq$, or by the deterministic nature of $R_i(\boldsymbol{\theta}^*); Q_j(\boldsymbol{\theta}^*)$) as well as at most (by $\supseteq$) one copy of the same element in RHS($C.5$). This says the two multisets are equal.
3. (Case) Applying the operational transition rule for 1 step one may observe that (writing "IH" the inductive

[8] We don't count "transitions" like $\langle P(\boldsymbol{\theta}^*), \rho\rangle \to \langle P(\boldsymbol{\theta}^*), \rho\rangle$ when analyzing computation paths, as nothing evolves in these "trivial transitions".

hypothesis)

$$
\begin{aligned}
& \{|\mathbf{0} \neq \rho' \in [\![\underline{\textbf{case}}\ M[\overline{q}] = \overline{m \to \underline{P_m(\boldsymbol{\theta})}}\ \textbf{end}]\!]\rho|\} \\
= & \coprod_{m^*} \{|\rho' \neq 0 : \langle P_m(\boldsymbol{\theta}^*), M_m \rho M_m^\dagger\rangle \to \langle \downarrow, \rho'\rangle|\} \\
\overset{IH}{=} & \coprod_{m^*} \coprod_{Q_{m^*,i_{m^*}}(\boldsymbol{\theta}) \in \texttt{Compile}(\underline{P_{m^*}(\boldsymbol{\theta})})} \qquad \text{(C.6)} \\
& \{|\rho' \neq 0 : \langle [\![Q_{m^*,i_{m^*}}(\boldsymbol{\theta}^*)]\!], M_{m^*}\rho M_{m^*}^\dagger\rangle \qquad \text{(C.7)} \\
& \qquad \to^* \langle \downarrow, \rho'\rangle|\}
\end{aligned}
$$

On the other hand we have

$$
\begin{aligned}
& \coprod_{Q(\boldsymbol{\theta}) \in \texttt{Compile}(\underline{\textbf{case}}\ M[\overline{q}] = \overline{m \to \underline{P_m(\boldsymbol{\theta})}}\ \textbf{end})} \\
& \{|\rho' \neq \mathbf{0} : \langle Q(\boldsymbol{\theta}^*), \rho\rangle \to^* \langle \downarrow, \rho'\rangle|\} \\
\overset{\text{CP,Case}}{=} & \coprod_{Q(\boldsymbol{\theta}) \in \texttt{FB}(\underline{\textbf{case}}\ M[\overline{q}] = \overline{m \to \underline{P_m(\boldsymbol{\theta})}}\ \textbf{end})} \\
& \{|\rho' \neq \mathbf{0} : \langle Q(\boldsymbol{\theta}^*), \rho\rangle \to^* \langle \downarrow, \rho'\rangle|\} \\
\overset{(**)}{=} & \coprod_{m^*} \coprod_{Q_{m^*,i_{m^*}}(\boldsymbol{\theta}) \in \texttt{Compile}(\underline{P_{m^*}(\boldsymbol{\theta})})} \{|\rho' \neq \mathbf{0} : \langle [\![Q_{m^*,i_{m^*}}(\boldsymbol{\theta}^*)]\!], \\
& M_{m^*}\rho M_{m^*}^\dagger\rangle \to^* \langle \downarrow, \rho'\rangle|\} \qquad \text{(C.8)}
\end{aligned}
$$

as desired. Note that the last step (**) is due to the behavior of the superoperators in $\texttt{FB}(\underline{\textbf{case}}\ M[\overline{q}] = \overline{m \to \underline{P_m(\boldsymbol{\theta})}}\ \textbf{end})$: when evolved by one transition, they either go to some *non-(essentially-aborting)* $\langle [\![Q_{m^*,i_{m^*}}(\boldsymbol{\theta}^*)]\!], M_{m^*}\rho M_{m^*}^\dagger\rangle$ for some $Q_{m^*,i_{m^*}}(\boldsymbol{\theta}) \in \texttt{Compile}(P_{m^*}(\boldsymbol{\theta}))$ for some m^*, or go to $\langle \textbf{abort}, M_{m^*}\rho M_{m^*}^\dagger\rangle$. Besides, each copy of final state of non-(essentially-aborting) $\langle [\![Q_{m^*,i_{m^*}}(\boldsymbol{\theta}^*)]\!], M_{m^*}\rho M_{m^*}^\dagger\rangle$ appears exactly once in the multiset, Due to our construction of $\texttt{FB}(\underline{\textbf{case}}\ M[\overline{q}] = \overline{m \to \underline{P_m(\boldsymbol{\theta})}}\ \textbf{end})$. On the other hand, computational paths starting with *essentially aborting* $\langle Q_{m^*,i_{m^*}}(\boldsymbol{\theta}^*), M_{m^*}\rho M_{m^*}^\dagger\rangle$ always terminate in $\langle \downarrow, \mathbf{0}\rangle$, therefore not counted on either side of (**).

4. (While$^{(T)}$) By definition, While$^{(T)}$ reduces to (Case) and (Sequence).
5. (Sum Components) If one of $\texttt{Compile}(\underline{P_b(\boldsymbol{\theta}^*)})$ is $\{\textbf{abort}\}$ the statement is immediately true, so we assume otherwise for below. Like in the **case** case, applying the operational transition rule for 1 step one may observe that (writing "IH" the inductive hypothesis)

$$
\begin{aligned}
& \mathbf{0} \neq \rho' \in [\![\underline{P_1(\boldsymbol{\theta}^*)} + \underline{P_2(\boldsymbol{\theta}^*)}]\!]\rho \qquad \text{(C.9)} \\
= & \coprod_{m^*=1,2} \{|\rho' \neq \mathbf{0} : \langle P_m(\boldsymbol{\theta}^*), \rho\rangle \to \langle \downarrow, \rho'\rangle|\} \qquad \text{(C.10)} \\
\overset{IH}{=} & \coprod_{m^*=1,2} \{|\rho' \neq \mathbf{0} : \langle [\![Q_{m^*,i_{m^*}}(\boldsymbol{\theta}^*)]\!], \rho\rangle \to^* \\
& \langle \downarrow, \rho'\rangle, Q_{m^*,i_{m^*}}(\boldsymbol{\theta}) \in \texttt{Compile}(\underline{P_{m^*}(\boldsymbol{\theta})})|\} \\
\overset{\text{CP,Sum Component}}{=} & \coprod_{Q(\boldsymbol{\theta}) \in \texttt{Compile}(\underline{P_1(\boldsymbol{\theta})} + \underline{P_2(\boldsymbol{\theta})})} \\
& \{|\rho' \neq 0 : \langle Q(\boldsymbol{\theta}^*), \rho\rangle \to^* \langle \downarrow, \rho'\rangle|\}, \qquad \text{(C.11)}
\end{aligned}
$$

as desired!

□

D Detailed Proofs from Section 6

Throughout, we use LHS, RHS to denote "left hand side" and "right hand side" resp., and "IH" to denote "Inductive Hypothesis". Wherever applicable, we adopt the overloading convention explained in the main text (See Eqn 6.6).

Before giving proofs, we record the full details for code transformation rule for while$^{(T)}$ for your interest. Let us denote $\frac{\partial}{\partial\theta}(\underline{\textbf{while}^{(T)}}\ M[\overline{q}] = 1\ \textbf{do}\ \underline{S_1(\boldsymbol{\theta})}\ \textbf{done}) \equiv \textbf{Seq}^T$, with:

$$
\begin{aligned}
\textbf{Seq}^{(1)} \equiv\ & \underline{\textbf{case}}\ M[\overline{q}] = 0 \to \quad \underline{\textbf{abort}}, \\
& 1 \to \quad \left(\tfrac{\partial}{\partial\theta_j}(\underline{P_1(\boldsymbol{\theta})}); \underline{\textbf{abort}}\right) + \\
& \left(\underline{P_1(\boldsymbol{\theta})}; \underline{\textbf{abort}}\right),
\end{aligned}
$$

and $\textbf{Seq}^{(T\geq 2)}$ recursively defined as:

$$
\begin{aligned}
\textbf{Seq}^{(T\geq 2)} \equiv\ & \underline{\textbf{case}}\ M[\overline{q}] = 0 \to \quad \underline{\textbf{abort}}, \\
& 1 \to \quad \left(\tfrac{\partial}{\partial\theta_j}(\underline{P_1(\boldsymbol{\theta})}); \textbf{while}^{(T-1)}\right) + \\
& \left(\underline{P_1(\boldsymbol{\theta})}; \textbf{Seq}^{(T-1)}\right)
\end{aligned}
$$

Note that $\textbf{Seq}^{(T-1)}$ essentially aborts. Let us now introduce some helper lemmas for the soundness proof:

D.1 Technical Lemmas

Lemma D.1. *$U(\theta) \in \{R_\sigma(\theta), R_{\sigma\otimes\sigma}(\theta)\}_{\sigma\in\{X,Y,Z\}}$. Let $\frac{d}{d\theta}U(\theta)$ denote the entry-wise derivative of $U(\theta)$. Then,*

$$\begin{aligned}
\frac{d}{d\theta}U(\theta) &= \frac{1}{2}U(\theta+\pi); && \text{(D.1)}\\
&\quad \frac{\partial}{\partial\theta}(\llbracket(O,\rho)\to\underline{U(\theta)}\rrbracket)\\
&= \mathrm{tr}(O\cdot U(\theta)\cdot\rho\cdot(\frac{d}{d\theta}U(\theta))^\dagger) +\\
&\quad \mathrm{tr}(O\cdot\frac{d}{d\theta}U(\theta)\cdot\rho\cdot U^\dagger(\theta)) && \text{(D.2)}\\
&\overset{D.1}{=} \frac{1}{2}\mathrm{tr}\Big(O\Big(U(\theta)\rho U^\dagger(\theta+\pi) +\\
&\quad U(\theta+\pi)\rho U^\dagger(\theta)\Big)\Big). && \text{(D.3)}
\end{aligned}$$

Proof. Let $U(\theta)\in\{R_\sigma(\theta),R_{\sigma\otimes\sigma}(\theta)\}_{\sigma\in\{X,Y,Z\}}$. Then $U(\theta) = \cos\frac{\theta}{2}I_\bullet - i\sin(\frac{\theta}{2})\boldsymbol{\sigma}$, where $\boldsymbol{\sigma}\in\{X,Y,Z,X\otimes X, Y\otimes Y, Z\otimes Z\}$. In the cases where $\boldsymbol{\sigma}$ is 1-qubit gate, $I_\bullet$ denotes identity on that one qubit; likewise for 2-qubit cases. Correctness of Eqns D.1, D.2 basically follows from straightforward computation.

1. Equation D.1:

$$\begin{aligned}
&\frac{d}{d\theta}(U(\theta))\\
&= \frac{1}{2}\Big((-\sin(\frac{\theta}{2}))I_\bullet - i\cos(\frac{\theta}{2})\boldsymbol{\sigma}\Big) && \text{(D.4)}\\
&= \frac{1}{2}\Big(\cos(\frac{\theta}{2}+\frac{\pi}{2})I_\bullet - i\sin(\frac{\theta}{2}+\frac{\pi}{2})\boldsymbol{\sigma}\Big) && \text{(D.5)}\\
&= \frac{1}{2}\Big(\cos(\frac{(\theta+\pi)}{2})I_\bullet - i\sin(\frac{(\theta+\pi)}{2})\boldsymbol{\sigma}\Big) && \text{(D.6)}\\
&= \frac{1}{2}U(\theta+\pi) && \text{(D.7)}
\end{aligned}$$

2. Equation D.2:

$$\begin{aligned}
&\frac{\partial}{\partial\theta}(\llbracket(O,\rho)\to\underline{U(\theta)}\rrbracket)\\
&= \frac{\partial}{\partial\theta}(\mathrm{tr}(OU(\theta)\rho U^\dagger(\theta))) && \text{(D.8)}\\
&= \frac{\partial}{\partial\theta}(\mathrm{tr}\Big(O\cdot(\cos\frac{\theta}{2}I_\bullet - i\sin(\frac{\theta}{2})\boldsymbol{\sigma})\cdot\rho\cdot\\
&\quad (\cos\frac{\theta}{2}I_\bullet + i\sin(\frac{\theta}{2})\boldsymbol{\sigma}^\dagger)\Big)) && \text{(D.9)}\\
&= \frac{\partial}{\partial\theta}\Big(\mathrm{tr}(\cos^2(\frac{\theta}{2})O\rho + i\cos(\frac{\theta}{2})\sin(\frac{\theta}{2})O\rho\boldsymbol{\sigma}^\dagger && \text{(D.10)}\\
&\quad -i\sin(\frac{\theta}{2})\cos(\frac{\theta}{2})O\boldsymbol{\sigma}\rho + \sin^2(\frac{\theta}{2})\boldsymbol{\sigma}\rho\boldsymbol{\sigma}^\dagger)\Big) && \text{(D.11)}\\
&= \mathrm{tr}(2\cos(\frac{\theta}{2})\cdot(\cos(\frac{\theta}{2}))'O\rho) && \text{(D.12)}\\
&\quad +\mathrm{tr}(i[\cos(\frac{\theta}{2})(\sin(\frac{\theta}{2}))' + (\cos(\frac{\theta}{2}))'\sin(\frac{\theta}{2})]O(\rho\boldsymbol{\sigma}^\dagger - \boldsymbol{\sigma}\rho))\\
&\quad +\mathrm{tr}(2\sin(\frac{\theta}{2})\cdot(\sin(\frac{\theta}{2}))'O\boldsymbol{\sigma}\rho\boldsymbol{\sigma}^\dagger), && \text{(D.13)}
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\mathrm{tr}(O\cdot U(\theta)\cdot\rho\cdot(\frac{d}{d\theta}U(\theta))^\dagger) +\\
&\mathrm{tr}(O\cdot\frac{d}{d\theta}U(\theta)\cdot\rho\cdot U^\dagger(\theta)) && \text{(D.14)}\\
&= \mathrm{tr}\Big(O\cdot(\cos\frac{\theta}{2}I_\bullet - i\sin(\frac{\theta}{2})\boldsymbol{\sigma})\cdot\rho\cdot\\
&\quad ((\cos\frac{\theta}{2})'I_\bullet + i(\sin(\frac{\theta}{2}))'\boldsymbol{\sigma}^\dagger)\Big) +\\
&\quad \mathrm{tr}\Big(O\cdot((\cos\frac{\theta}{2})'I_\bullet - i(\sin(\frac{\theta}{2}))'\boldsymbol{\sigma})\cdot\rho\cdot\\
&\quad (\cos\frac{\theta}{2}I_\bullet + i\sin(\frac{\theta}{2})\boldsymbol{\sigma}^\dagger)\Big) && \text{(D.15)}\\
&= \mathrm{tr}(2\cos(\frac{\theta}{2})\cdot(\cos(\frac{\theta}{2}))'O\rho) && \text{(D.16)}\\
&\quad +\mathrm{tr}(i[\cos(\frac{\theta}{2})(\sin(\frac{\theta}{2}))' + (\cos(\frac{\theta}{2}))'\sin(\frac{\theta}{2})]O(\rho\boldsymbol{\sigma}^\dagger - \boldsymbol{\sigma}\rho))\\
&\quad +\mathrm{tr}(2\sin(\frac{\theta}{2})\cdot(\sin(\frac{\theta}{2}))'O\boldsymbol{\sigma}\rho\boldsymbol{\sigma}^\dagger), && \text{(D.17)}
\end{aligned}$$

as desired.

□

Lemma D.2. *Let $P(\boldsymbol{\theta})\in \boldsymbol{q\text{-}while}^{(T)}_{\overline{v}}(\boldsymbol{\theta}), P'(\boldsymbol{\theta})\in \boldsymbol{q\text{-}while}^{(T)}_{\overline{v}\cup\{A\}}(\boldsymbol{\theta})$. Then for arbitrary $\boldsymbol{\theta}^*\in\mathbb{R}^k$, $\rho\in\mathcal{D}(\mathcal{H}_{\overline{v}})$, $O\in O_{\overline{v}}$,*

1. *$\llbracket(O,\rho)\to P(\boldsymbol{\theta}^*); P'(\boldsymbol{\theta}^*)\rrbracket = \llbracket(O,\llbracket P(\boldsymbol{\theta}^*)\rrbracket\rho)\to P'(\boldsymbol{\theta}^*)\rrbracket$.*
2. *$\llbracket(O,\rho)\to P'(\boldsymbol{\theta}^*); P(\boldsymbol{\theta}^*)\rrbracket = \llbracket(\llbracket P(\boldsymbol{\theta}^*)\rrbracket^*(O),\rho)\to P'(\boldsymbol{\theta}^*)\rrbracket$, with $\llbracket P(\boldsymbol{\theta}^*)\rrbracket^*$ the dual of $\llbracket P(\boldsymbol{\theta}^*)\rrbracket$.*

Proof. Observe that both $P(\boldsymbol{\theta}^*); P'(\boldsymbol{\theta}^*)$ and $P'(\boldsymbol{\theta}^*); P(\boldsymbol{\theta}^*)$ lives in $\mathbf{q\text{-}while}^{(T)}_{\overline{v}\cup\{A\}}(\boldsymbol{\theta})$, so we identify the "smaller" program $P(\boldsymbol{\theta}^*)$ with $\mathcal{I}_A\otimes P(\boldsymbol{\theta}^*)$, where $\mathcal{I}\equiv I_A\circ I_A$ denotes the identity operation on Ancilla.

1. Unfolding definition 5.2,

$$\begin{aligned}
&\llbracket(O,\rho)\to P(\boldsymbol{\theta}^*); P'(\boldsymbol{\theta}^*)\rrbracket\\
&= \mathrm{tr}((Z_A\otimes O)\llbracket P(\boldsymbol{\theta}^*); P'(\boldsymbol{\theta}^*)\rrbracket((|0\rangle_A\langle 0|)\otimes\rho))\\
&\overset{\text{Fig1}b,\ \text{Sequence}}{=} \mathrm{tr}((Z_A\otimes O)\llbracket P'(\boldsymbol{\theta}^*)\rrbracket\llbracket\mathcal{I}_A\otimes P(\boldsymbol{\theta}^*)\rrbracket((|0\rangle_A\langle 0|)\otimes\rho))\\
&= \mathrm{tr}((Z_A\otimes O)\llbracket P'(\boldsymbol{\theta}^*)\rrbracket((|0\rangle_A\langle 0|)\otimes(\llbracket P(\boldsymbol{\theta}^*)\rrbracket\rho)))\\
&= \llbracket(O,\llbracket P(\boldsymbol{\theta}^*)\rrbracket\rho)\to P'(\boldsymbol{\theta}^*)\rrbracket, \text{ as desired.}
\end{aligned}$$

2. Observe that:

$$\begin{aligned}
&\llbracket(O,\rho)\to P'(\boldsymbol{\theta}^*); P(\boldsymbol{\theta}^*)\rrbracket\\
&= \mathrm{tr}((Z_A\otimes O)\llbracket P'(\boldsymbol{\theta}^*); P(\boldsymbol{\theta}^*)\rrbracket\cdot\\
&\quad \Big((|0\rangle_A\langle 0|)\otimes\rho\Big)\Big)\\
&\overset{\text{Fig1}b,\ \text{Sequence}}{=} \mathrm{tr}((Z_A\otimes O)\llbracket\mathcal{I}_A\otimes P(\boldsymbol{\theta}^*)\rrbracket\llbracket P'(\boldsymbol{\theta}^*)\rrbracket\cdot\\
&\quad \Big((|0\rangle_A\langle 0|)\otimes\rho\Big)\Big)\\
&\overset{\zeta\equiv\llbracket P'(\boldsymbol{\theta}^*)\rrbracket((|0\rangle_A\langle 0|)\otimes\rho)}{=} \mathrm{tr}((Z_A\otimes O)\llbracket\mathcal{I}_A\otimes P(\boldsymbol{\theta}^*)\rrbracket\zeta), \quad \text{(D.18)}
\end{aligned}$$

while

$$[\![([\![P(\boldsymbol{\theta}^*)]\!]^*(O),\rho)\to P'(\boldsymbol{\theta}^*)]\!] \tag{D.19}$$
$$= \mathrm{tr}((Z_A\otimes[\![P(\boldsymbol{\theta}^*)]\!]^*(O))[\![P'(\boldsymbol{\theta}^*)]\!]((|0\rangle_A\langle 0|)\otimes\rho))$$
$$= \mathrm{tr}((Z_A\otimes[\![P(\boldsymbol{\theta}^*)]\!]^*(O))\zeta). \tag{D.20}$$

Now RHS (Eqn$D.18$) equals RHS(Eqn(D.20)) via duality, as $\mathcal{I}_A\otimes P(\boldsymbol{\theta}^*)$ does nothing on the ancilla, while $[\![P(\boldsymbol{\theta}^*)]\!]^*(O)\in O_{\overline{v}}$ and Z_A are compatible observables.

□

Lemma D.3. *Let* $\underline{S_m(\boldsymbol{\theta})}\in$ ***add-q-while***$^{(T)}_{\overline{v}}(\boldsymbol{\theta})$ *(*$\forall m\in[0,w]$*, where* $w\geq 1$*). Denote:*[9]

$$\texttt{Compile}(\underline{S_m(\boldsymbol{\theta})}) \equiv \{|Q_{m,0}(\boldsymbol{\theta}),\cdots,Q_{m,s_m}(\boldsymbol{\theta})|\}, \tag{D.21}$$
$$\texttt{Compile}(\frac{\partial}{\partial\theta}(\underline{S_m(\boldsymbol{\theta})})) \equiv \{|P_{m,0}(\boldsymbol{\theta}),\cdots,P_{m,t_m}(\boldsymbol{\theta})|\}. \tag{D.22}$$

Then for arbitrary $\rho\in\mathcal{D}(\mathcal{H}_{\overline{v}}), O\in O_{\overline{v}}$*, we have the following computational properties regarding the observable semantics:*

$$[\![(O,\rho)\to\frac{\partial}{\partial\theta}(\underline{S_0(\boldsymbol{\theta});S_1(\boldsymbol{\theta})})]\!]$$
$$= [\![(O,\rho)\to\frac{\partial}{\partial\theta}(\underline{S_0(\boldsymbol{\theta})});\underline{S_1(\boldsymbol{\theta})}]\!]$$
$$+[\![(O,\rho)\to\underline{S_0(\boldsymbol{\theta})};\frac{\partial}{\partial\theta}(\underline{S_1(\boldsymbol{\theta})})]\!], \tag{D.23}$$

$$[\![(O,\rho)\to\frac{\partial}{\partial\theta}(\underline{\textbf{case}}\ M[\overline{q}]=\overline{m\to\underline{S_m(\boldsymbol{\theta})}}\ \textbf{end})]\!]$$
$$= \sum_m\sum_{i_m}[\![(O,M_m\rho M_m^\dagger)\to P_{m,i_m}(\boldsymbol{\theta})]\!],$$
$$\textit{where } i_m \textit{ runs through } [0,t_m]; \tag{D.24}$$

$$[\![(O,\rho)\to\left(\frac{\partial}{\partial\theta}(\underline{S_0(\boldsymbol{\theta})+S_1(\boldsymbol{\theta})})\right)]\!]$$
$$= [\![(O,\rho)\to\frac{\partial}{\partial\theta}(\underline{S_0(\boldsymbol{\theta})})]\!]+ \tag{D.25}$$
$$[\![(O,\rho)\to\frac{\partial}{\partial\theta}(\underline{S_1(\boldsymbol{\theta})})]\!] \tag{D.26}$$

$$[\![(O,\rho)\to\underline{S_0(\boldsymbol{\theta})+S_1(\boldsymbol{\theta})}]\!]$$
$$= [\![(O,\rho)\to\underline{S_0(\boldsymbol{\theta})}]\!]+[\![(O,\rho)\to\underline{S_1(\boldsymbol{\theta})}]\!], \tag{D.27}$$

and

$$[\![(O,\rho)\to\textbf{case}\ M[\overline{q}]=\overline{m\to\underline{S_m(\boldsymbol{\theta})}}\ \textbf{end}]\!]$$
$$= \sum_m\sum_{j_m\in[0,s_m]}[\![(O,\mathcal{E}_m\rho)\to Q_{m,j_m}(\boldsymbol{\theta})]\!] \tag{D.28}$$

[9] As a reminder: we shall frequently refer to Notations in Eqns D.21 and D.22 for the rest of the section.

Proof. We first make some direct application of the code transformation (hereinafter "CT", Fig 4) and compilation (hereinafter "CP", Fig 3) rules; some results will immediately follow from these application.

$$\texttt{Compile}\left(\frac{\partial}{\partial\theta}(\underline{S_0(\boldsymbol{\theta});S_1(\boldsymbol{\theta})})\right)$$
$$\overset{\text{CT,Sequence}}{=} \texttt{Compile}((\underline{S_0(\boldsymbol{\theta})};\frac{\partial}{\partial\theta}(\underline{S_1(\boldsymbol{\theta})}))+(\frac{\partial}{\partial\theta}(\underline{S_0(\boldsymbol{\theta})});\underline{S_1(\boldsymbol{\theta})}))$$
$$\overset{\text{CP,Sum Components}}{=} \texttt{Compile}\left(\underline{S_0(\boldsymbol{\theta})};\frac{\partial}{\partial\theta}(\underline{S_1(\boldsymbol{\theta})})\right) \tag{D.29}$$
$$\bigsqcup\texttt{Compile}\left(\frac{\partial}{\partial\theta}(\underline{S_0(\boldsymbol{\theta})});\underline{S_1(\boldsymbol{\theta})}\right)$$
$$\overset{\text{CP,Sequence}}{=} \{|Q_{0,g}(\boldsymbol{\theta});P_{1,h}(\boldsymbol{\theta})|\}_{g\in[s_0],h\in[t_1]} \tag{D.30}$$
$$\bigsqcup\{|P_{0,i}(\boldsymbol{\theta});Q_{1,j}(\boldsymbol{\theta})|\}_{i\in[t_0],j\in[s_1]}; \tag{D.31}$$

, where $[s_0],[s_1],[t_0],[t_1]$ stand for e.g. $[0,s_0],[0,s_1],[0,t_0],[0,t_1]$ respectively; and

$$\texttt{Compile}\left(\frac{\partial}{\partial\theta}(\underline{\textbf{case}}\ M[\overline{q}]=\overline{m\to\underline{S_m(\boldsymbol{\theta})}}\ \textbf{end})\right)$$
$$\overset{\text{CT,Case}}{=} \texttt{Compile}\left(\underline{\textbf{case}}\ M[\overline{q}]=\overline{m\to\frac{\partial}{\partial\theta}(\underline{S_m(\boldsymbol{\theta})})}\ \textbf{end}\right)$$
$$\overset{\text{CP,Case}}{=} \texttt{FB}\left(\underline{\textbf{case}}\ M[\overline{q}]=\overline{m\to\frac{\partial}{\partial\theta}(\underline{S_m(\boldsymbol{\theta})})}\ \textbf{end}\right);$$

also,

$$\texttt{Compile}\left(\frac{\partial}{\partial\theta}(S_0(\boldsymbol{\theta})+S_1(\boldsymbol{\theta}))\right) \tag{D.32}$$
$$\overset{\text{CT,Sum Component}}{=} \texttt{Compile}\left(\frac{\partial}{\partial\theta}(S_0(\boldsymbol{\theta}))+\frac{\partial}{\partial\theta}(S_1(\boldsymbol{\theta}))\right) \tag{D.33}$$
$$\overset{\text{CP,Sum Component}}{=} \{|P_{0,0}(\boldsymbol{\theta}),\cdots,P_{0,t_0}|\} \tag{D.34}$$
$$\bigsqcup\{|P_{1,0}(\boldsymbol{\theta}),\cdots,P_{1,t_1}|\}. \tag{D.35}$$

As promised, Eqns D.23 and D.26 immediately follows from the results above together with the corresponding definitions. So is Eqn D.27, following the same lines of unfolding by CP-(ND component) laid out in Eqns D.33~ D.35 then a direct pattern matching by definition. To obtain D.24, observe

that

$$[\![(O,\rho) \to \frac{\partial}{\partial\theta}(\underline{\textbf{case}}\ M[\overline{q}] = \overline{m \to \underline{S_m(\boldsymbol{\theta})}}\ \textbf{end})]\!] \tag{D.36}$$

$$= \sum_{Q(\boldsymbol{\theta})\in \mathsf{FB}(\underline{\textbf{case}}\ M[\overline{q}]= \overline{m\to \frac{\partial}{\partial\theta}(\underline{S_m(\boldsymbol{\theta})})}\ \textbf{end})} [\![(O,\rho) \to Q(\boldsymbol{\theta})]\!] \tag{D.37}$$

$$= \sum_{Q(\boldsymbol{\theta})\in \mathsf{FB}(\underline{\textbf{case}}\ M[\overline{q}]= \overline{m\to \frac{\partial}{\partial\theta}(\underline{S_m(\boldsymbol{\theta})})}\ \textbf{end})} \mathrm{tr}((Z_A\otimes O)[\![Q(\boldsymbol{\theta})]\!](|0\rangle_A\langle 0|\otimes\rho)) \tag{D.38}$$

$$\overset{A\notin\overline{q}\subseteq\overline{v}}{=} \sum_{m^*}\sum_{i_{m^*}\in[0,t_{m^*}]} \mathrm{tr}\Big((Z_A\otimes O)\cdot [\![P_{m^*,i_{m^*}}(\boldsymbol{\theta})]\!]((|0\rangle_A\langle 0|)\otimes M_{m^*}\rho M_{m^*}^\dagger)\Big) \tag{D.39}$$

$$= \sum_{m^*}\sum_{i_{m^*}} [\![(O, M_{m^*}\rho M_{m^*}^\dagger) \to P_{m^*,i_{m^*}}(\boldsymbol{\theta})]\!] \text{ as desired.} \tag{D.40}$$

A few more words on Eqn D.40. Since A is disjoint from $\overline{v}\supseteq\overline{q}$, for each $Q(\boldsymbol{\theta})\in\mathsf{FB}(\underline{\textbf{case}}\ M[\overline{q}] = \overline{m\to\frac{\partial}{\partial\theta}(\underline{S_m(\boldsymbol{\theta})})}\ \textbf{end})$, evolving one step per the operational semantics will lead to the configuration $\langle [\![P_{m^*,i_{m^*}}(\boldsymbol{\theta})]\!], ((|0\rangle_A\langle 0|)\otimes M_{m^*}\rho M_{m^*}^\dagger)\rangle$ for some non-(essentially-aborting) $P_{m^*,i_{m^*}}(\boldsymbol{\theta})$. Besides, each such configuration appears exactly once affording one summand, due to construction of $\mathsf{FB}(\bullet)$. On the other hand, if $P_{m^*,i_{m^*}}(\boldsymbol{\theta})$ essentially aborts, the trace computed as a summand vanishes, making no contribution to the sum, thereby maintaining the equation.

Eqn D.28 is proved using the same lines of logic:

$$[\![(O,\rho) \to \textbf{case}\ M[\overline{q}] = \overline{m\to\underline{S_m(\boldsymbol{\theta})}}\ \textbf{end}]\!] \tag{D.41}$$

$$= \sum_{Q(\boldsymbol{\theta})\in \mathsf{FB}(\underline{\textbf{case}}\ M[\overline{q}]= \overline{m\to\underline{S_m(\boldsymbol{\theta})}}\ \textbf{end})} [\![(O,\rho) \to Q(\boldsymbol{\theta})]\!]$$

$$= \sum_{Q(\boldsymbol{\theta})\in \mathsf{FB}(\underline{\textbf{case}}\ M[\overline{q}]= \overline{m\to\underline{S_m(\boldsymbol{\theta})}}\ \textbf{end})} \mathrm{tr}(O[\![Q(\boldsymbol{\theta})]\!]\rho) \tag{D.42}$$

$$\overset{A\notin\overline{q}\subseteq\overline{v}}{=} \sum_{m^*}\sum_{j_{m^*}\in[0,s_{m^*}]} \mathrm{tr}(O[\![Q_{m^*,j_{m^*}}(\boldsymbol{\theta})]\!]M_{m^*}\rho M_{m^*}^\dagger)$$

$$= \sum_{m^*}\sum_{j_{m^*}\in[0,s_{m^*}]} [\![(O,\mathcal{E}_{m^*}\rho) \to Q_{m^*,j_{m^*}}(\boldsymbol{\theta})]\!]$$

as desired. □

Lemma D.4. *Keeping the notations in Lemma D.3, we have* $\forall\boldsymbol{\theta}^*\in\mathbb{R}^k, O\in\mathcal{O}_{\overline{v}}, \rho\in\mathcal{D}(\mathcal{H}_{\overline{v}})$,

$$\Big(\frac{\partial}{\partial\theta}([\![(O,\rho)\to\underline{S_0(\boldsymbol{\theta});S_1(\boldsymbol{\theta})}]\!])\Big)\Big|_{\boldsymbol{\theta}^*}$$

$$= \sum_{g\in[0,s_0]}\Big(\frac{\partial}{\partial\theta}([\![(O,[\![Q_{0,g}(\boldsymbol{\theta}^*)]\!]\rho)\to\underline{S_1(\boldsymbol{\theta})}]\!])\Big)\Big|_{\boldsymbol{\theta}^*} \tag{D.43}$$

$$+ \sum_{j\in[0,s_1]}\Big(\frac{\partial}{\partial\theta}([\![([\![Q_{1,j}(\boldsymbol{\theta}^*)]\!]^*(O),\rho)\to\underline{S_0(\boldsymbol{\theta})}]\!])\Big)\Big|_{\boldsymbol{\theta}^*} \tag{D.44}$$

Proof. Observe:

$$\overset{\text{CP,Sequence}}{=} \text{LHS}(D.43) = \sum_{g\in[0,s_0],j\in[0,s_1]}\Big(\frac{\partial}{\partial\theta}([\![(O,\rho)\to Q_{0,g}(\boldsymbol{\theta});Q_{1,j}(\boldsymbol{\theta})]\!])\Big)\Big|_{\boldsymbol{\theta}^*}$$

$$\overset{(***),\text{See Below}}{=} \sum_{g\in[0,s_0],j\in[0,s_1]}\Big(\frac{\partial}{\partial\theta}([\![(O,[\![Q_{0,g}(\boldsymbol{\theta}^*)]\!]\rho)\to Q_{1,j}(\boldsymbol{\theta})]\!])\Big)\Big|_{\boldsymbol{\theta}^*} \tag{D.45}$$

$$+ \sum_{g\in[0,s_0],j\in[0,s_1]}\Big(\frac{\partial}{\partial\theta}([\![([\![Q_{1,j}(\boldsymbol{\theta})]\!]^*(O),\rho)\to Q_{0,g}(\boldsymbol{\theta})]\!])\Big)\Big|_{\boldsymbol{\theta}^*} \tag{D.46}$$

$$= \sum_{g\in[0,s_0]}\Big(\frac{\partial}{\partial\theta}([\![(O,[\![Q_{0,g}(\boldsymbol{\theta}^*)]\!]\rho)\to\underline{S_1(\boldsymbol{\theta})}]\!])\Big)\Big|_{\boldsymbol{\theta}^*} \tag{D.47}$$

$$+ \sum_{j\in[0,s_1]}\Big(\frac{\partial}{\partial\theta}([\![([\![Q_{1,j}(\boldsymbol{\theta}^*)]\!]^*(O),\rho)\to\underline{S_0(\boldsymbol{\theta})}]\!])\Big)\Big|_{\boldsymbol{\theta}^*} \tag{D.48}$$

Where the last step is direct application of definition of input-space observable semantics. For the step (***) claiming RHS(D.45)= RHS(D.45)+ RHS(D.46), observe that it boils down to the following: for arbitrary $(\partial^*,j^*)\in[0,s_0]\times[0,s_1]$,

$$\Big(\frac{\partial}{\partial\theta}([\![(O,\rho)\to Q_{0,g^*}(\boldsymbol{\theta});Q_{1,j^*}(\boldsymbol{\theta})]\!])\Big)\Big|_{\boldsymbol{\theta}^*} \tag{D.49}$$

$$= \Big(\frac{\partial}{\partial\theta}([\![(O,[\![Q_{0,g^*}(\boldsymbol{\theta}^*)]\!]\rho)\to Q_{1,j^*}(\boldsymbol{\theta})]\!])\Big)\Big|_{\boldsymbol{\theta}^*} \tag{D.50}$$

$$+ \Big(\frac{\partial}{\partial\theta}([\![([\![Q_{1,j^*}(\boldsymbol{\theta})]\!]^*(O),\rho)\to Q_{0,g^*}(\boldsymbol{\theta})]\!])\Big)\Big|_{\boldsymbol{\theta}^*} \tag{D.51}$$

We argue the correctness of D.49~D.51 below. Simplifying notations, write $Q_{0,g^*}(\boldsymbol{\theta}), Q_{1,j^*}(\boldsymbol{\theta})$ as $Q_0(\boldsymbol{\theta}), Q_1(\boldsymbol{\theta})$ respectively, then let

$$Q_0(\boldsymbol{\theta}) \equiv \sum_x K_x(\boldsymbol{\theta}) \circ K_x^\dagger(\boldsymbol{\theta}), \tag{D.52}$$

$$Q_1(\boldsymbol{\theta}) \equiv \sum_y J_y(\boldsymbol{\theta}) \circ J_y^\dagger(\boldsymbol{\theta}) \tag{D.53}$$

denote the kraus operator decomposition of $Q_0(\boldsymbol{\theta}), Q_1(\boldsymbol{\theta})$, each with finitely many summands.[10]

Then,

$$\left(\frac{\partial}{\partial\theta}([\![(O,\rho) \to Q_g(\boldsymbol{\theta}); Q_j(\boldsymbol{\theta})]\!])\right)\Big|_{\boldsymbol{\theta}^*} \tag{D.56}$$

$$= \frac{\partial}{\partial\theta_j}\mathrm{tr}(O[\![Q_0(\boldsymbol{\theta}); Q_1(\boldsymbol{\theta})]\!]\rho)\Big|_{\boldsymbol{\theta}^*} \tag{D.57}$$

$$= \frac{\partial}{\partial\theta_j}\mathrm{tr}(O\sum_y J_y(\boldsymbol{\theta})(\sum_x K_x(\boldsymbol{\theta})\rho K_x^\dagger(\boldsymbol{\theta}))J_y^\dagger(\boldsymbol{\theta}))\Big|_{\boldsymbol{\theta}^*} \tag{D.58}$$

$$= \sum_{x,y}\frac{\partial}{\partial\theta_j}\mathrm{tr}(OJ_y(\boldsymbol{\theta})K_x(\boldsymbol{\theta})\rho K_x^\dagger(\boldsymbol{\theta})J_y^\dagger(\boldsymbol{\theta}))\Big|_{\boldsymbol{\theta}^*} \tag{D.59}$$

$$\overset{(a)}{=} \sum_y\Big[(\mathrm{tr}(O\frac{d}{d\theta_j}J_y(\boldsymbol{\theta})(\sum_x K_x(\boldsymbol{\theta}^*)\rho K_x^\dagger(\boldsymbol{\theta}^*))J_y^\dagger(\boldsymbol{\theta}^*))$$

$$+\mathrm{tr}(OJ_y(\boldsymbol{\theta}^*)(\sum_x K_x(\boldsymbol{\theta}^*)\rho K_x^\dagger(\boldsymbol{\theta}^*))\frac{d}{d\theta_j}J_y^\dagger(\boldsymbol{\theta}^*)))\Big]\Big|_{\boldsymbol{\theta}^*}$$

$$+\sum_y\Big[(\mathrm{tr}(OJ_y(\boldsymbol{\theta}^*)(\sum_x \frac{d}{d\theta_j}K_x(\boldsymbol{\theta})\rho K_x^\dagger(\boldsymbol{\theta}^*))J_y^\dagger(\boldsymbol{\theta}^*))$$

$$+\mathrm{tr}(OJ_y(\boldsymbol{\theta}^*)(\sum_x K_x(\boldsymbol{\theta}^*)\rho\frac{d}{d\theta_j}K_x^\dagger(\boldsymbol{\theta}))J_y^\dagger(\boldsymbol{\theta}^*)))\Big]\Big|_{\boldsymbol{\theta}^*} \tag{D.60}$$

$$\overset{(b)}{=} \frac{\partial}{\partial\theta_j}\mathrm{tr}(O[\![Q_1(\boldsymbol{\theta})]\!]([\![Q_0(\boldsymbol{\theta}^*)]\!]\rho))\Big|_{\boldsymbol{\theta}^*} \tag{D.61}$$

$$+\frac{\partial}{\partial\theta_j}\mathrm{tr}(O[\![Q_1(\boldsymbol{\theta}^*)]\!][\![Q_0(\boldsymbol{\theta})]\!]\rho)\Big|_{\boldsymbol{\theta}^*} \tag{D.62}$$

$$\overset{\text{S-H Dual}}{=} \left(\frac{\partial}{\partial\theta_j}([\![(O,[\![Q_0(\boldsymbol{\theta}^*)]\!]\rho) \to Q_1(\boldsymbol{\theta})]\!])\right)\Big|_{\boldsymbol{\theta}^*} \tag{D.63}$$

$$+\left(\frac{\partial}{\partial\theta_j}([\![([\![Q_1(\boldsymbol{\theta}^*)]\!]^*(O),\rho) \to Q_0(\boldsymbol{\theta})]\!])\right)\Big|_{\boldsymbol{\theta}^*} \tag{D.64}$$

where $\frac{d}{d\theta_j}\bullet$ again denotes the entry-wise derivative, and $(a), (b)$ is obtained by unfolding the definition of trace, noticing that all entries of parameterized unitaries are smooth. S-H dual here is short for Schrödinger-Heisenberg dual, as is everywhere else in this section.

□

D.2 Proof Details of Theorem 6.2: Soundness of Differentiation Logic

Proof. As stated in the main text, throughout we fix $O_A \equiv Z_A$. Unfolding by definition, it suffices to show

$$\forall O \in \mathcal{O}_{\overline{v}}, \rho \in \mathcal{D}(\mathcal{H}_{\overline{v}}), [\![((O,Z_A),\rho) \to \frac{\partial}{\partial\theta}(\underline{\mathsf{S}(\boldsymbol{\theta})})]\!]$$
$$= \frac{\partial}{\partial\theta}([\![(O,\rho) \to \underline{\mathsf{S}(\boldsymbol{\theta})}]\!]). \tag{D.65}$$

Let $O \in \mathcal{O}_{\overline{v}}, \rho \in \mathcal{D}(\mathcal{H}_{\overline{v}})$ be arbitrary. We consider each rule in Figure 5 possibly used as the final step of the derivation.

1. (Abort)~(Trivial Unitary): for $\underline{\mathsf{S}(\boldsymbol{\theta})} \equiv \underline{\mathbf{abort}[\overline{v}]}, \underline{\mathbf{skip}[\overline{v}]}$ or $\underline{q := |0\rangle}$, $[\![(O,\rho) \to \underline{\mathsf{S}(\boldsymbol{\theta})}]\!]$ is a constant, so $\frac{\partial}{\partial\theta}([\![(O,\rho) \to \underline{\mathsf{S}(\boldsymbol{\theta})}]\!]) = 0$. On the other hand, $\frac{\partial}{\partial\theta}(\underline{\mathsf{S}(\boldsymbol{\theta})}) \overset{\text{CT,trivialize}}{\equiv} \underline{\mathbf{abort}[\overline{v}\cup\{A\}]}$, meaning $[\![(O,\rho) \to \frac{\partial}{\partial\theta}(\underline{\mathsf{S}(\boldsymbol{\theta})})]\!] = \mathrm{tr}((Z_A\otimes\overline{O})\cdot\mathbf{0}) = 0$.
   In Trivial-Unitary, $[\![(O,\rho) \to \underline{U(\boldsymbol{\theta})}]\!]$ doesn't depend on θ_j since $\theta_j \notin \boldsymbol{\theta}$; hence $\frac{\partial}{\partial\theta_j}([\![(O,\rho) \to \underline{U(\boldsymbol{\theta})}]\!]) = 0$, as faithfully represented by the code transformation rule.
2. (Unitary): Assume $\underline{\mathsf{S}(\boldsymbol{\theta})} \equiv \underline{U(\theta)}$ with $U(\theta) \in \{R_\sigma(\theta), R_{\sigma\otimes\sigma}(\theta)\}_{\sigma\in\{X,Y,Z\}}$. Let $\boldsymbol{\sigma}$ range over $\{\sigma, \sigma\otimes\sigma\}_{\sigma\in X,Y,Z}$ (proof below works for all 6 cases where $\boldsymbol{\sigma}^2 = I_\bullet$). Let $\frac{d}{d\theta}U(\theta)$ denote the entry-wise derivative of $U(\theta)$, then

[10] This is doable because for each $\boldsymbol{\theta}^* \in \mathbb{R}^k$ we have

$$Q_0(\boldsymbol{\theta}^*) = \sum_{k=1} K_x(\boldsymbol{\theta}^*) \circ K_x^\dagger(\boldsymbol{\theta}^*), \tag{D.54}$$

$$Q_1(\boldsymbol{\theta}^*) = \sum_{q=1} J_y(\boldsymbol{\theta}^*) \circ J_y^\dagger(\boldsymbol{\theta}^*). \tag{D.55}$$

Since parameterized unitaries have a parameterized-matrix representation, the superoperator $\mathcal{U}(\boldsymbol{\theta}) = U(\boldsymbol{\theta}))\circ U^\dagger(\boldsymbol{\theta})$ for any parameterized unitaries also have a parameterized operator representation. In our parameterized quantum program syntax only unitaries are parameterized, thus the kraus operators $K_x(\boldsymbol{\theta})$, $J_y(\boldsymbol{\theta})$'s can be chosen uniformly. If one insists one may perform a standard structural induction to show the existence of such a parameterized Kraus operator decomposition for any $Q(\boldsymbol{\theta}) \in \mathbf{q\text{-}while}_{\overline{v}}^{(T)}(\boldsymbol{\theta})$. For example, assuming each $S_m(\boldsymbol{\theta})$ decomposes to $\sum_{t_{m,j}=t_{m,1}}^{t_{m,j_m}} K_{t_{m,j}}(\boldsymbol{\theta})\circ K_{t_{m,j}}^\dagger(\boldsymbol{\theta})$, then **case** $M[\overline{q}] = m \to S_m(\boldsymbol{\theta})$ decomposes into

$$\sum_m(\sum_{t_{m,j}} K_{t_{m,j}}(\boldsymbol{\theta})\circ K_{t_{m,j}}^\dagger(\boldsymbol{\theta}))\circ\mathcal{E}_m.$$

As in the unitary case, the linear Kraus operators are entry-wisesmooth, ensured by the entry-wise smoothness of matrix representations of parameterized unitaries.

recall from Lemma D.1 that

$$\begin{aligned}\frac{d}{d\theta}U(\theta) &= \frac{1}{2}U(\theta+\pi); && \text{(D.66)}\\ &\quad \frac{\partial}{\partial\theta}([\![(O,\rho)\to\underline{U(\theta)}]\!])\\ &= \mathrm{tr}(O\cdot U(\theta)\cdot\rho\cdot(\frac{d}{d\theta}U(\theta))^\dagger)\\ &\quad +\mathrm{tr}(O\cdot\frac{d}{d\theta}U(\theta)\cdot\rho\cdot U^\dagger(\theta)). && \text{(D.67)}\\ &\overset{D.66}{=} \frac{1}{2}\mathrm{tr}\Big(O\Big(U(\theta)\rho U^\dagger(\theta+\pi)\\ &\quad +U(\theta+\pi)\rho U^\dagger(\theta)\Big)\Big)\end{aligned}$$

Meanwhile,

$$\begin{aligned}&[\![(O,\rho)\to\frac{\partial}{\partial\theta}(\underline{U(\theta)})]\!]\\ \overset{\text{CT,1-qb,2-qb}}{=}\ &\mathrm{tr}((Z_A\otimes O)[\![R'_{\boldsymbol{\sigma}}(\theta)]\!]\cdot\\ &((|0\rangle_A\langle 0|)\otimes\rho)) && \text{(D.68)}\end{aligned}$$

Let us unfold the definition of $R'_{\boldsymbol{\sigma}}(\theta)$:

$$\begin{aligned}&\text{RHS}(D.68)\\ =\ &\frac{1}{2}\sum_{x,y\in\{0,1\}}\mathrm{tr}((Z_A\otimes O)[\![C-R_{\boldsymbol{\sigma}}(\theta);H_A]\!]\cdot\\ &((|x\rangle_A\langle y|)\otimes\rho))\\ =\ &\frac{1}{2}\mathrm{tr}\Big((Z_A\otimes O)[\![H_A]\!]\Big(|0\rangle_A\langle 0|\otimes U(\theta)\rho U^\dagger(\theta)\\ &+|0\rangle_A\langle 1|\otimes U(\theta)\rho U^\dagger(\theta+\pi) && \text{(D.69)}\\ &+|1\rangle_A\langle 0|\otimes U(\theta+\pi)\rho U^\dagger(\theta) && \text{(D.70)}\\ &+|1\rangle_A\langle 1|\otimes U(\theta+\pi)\rho U^\dagger(\theta+\pi)\Big)\Big) && \text{(D.71)}\\ =\ &\frac{1}{2}\mathrm{tr}\Big((Z_A\otimes O)\Big(\frac{\sum_{x,y}|x\rangle_A\langle y|}{2}\otimes U(\theta)\rho U^\dagger(\theta) && \text{(D.72)}\\ &+\frac{\sum_x|x\rangle_A\langle 0|-\sum_x|x\rangle_A\langle 1|}{2}U(\theta)\rho U^\dagger(\theta+\pi) && \text{(D.73)}\\ &+\frac{\sum_x|0\rangle_A\langle x|-\sum_x|1\rangle_A\langle x|}{2}U(\theta+\pi)\rho U^\dagger(\theta) && \text{(D.74)}\\ &+\frac{\sum_x|x\rangle_A\langle x|-(|1\rangle_A\langle 0|+|0\rangle_A\langle 1|)}{2}\cdot\\ &U(\theta+\pi)\rho U^\dagger(\theta+\pi)\Big)\Big) && \text{(D.75)}\end{aligned}$$

Let's regroup the terms of D.72 ∼ D.75, bearing in mind that when observing $Z_A\otimes O$ on the final state, the "off-diagonal" part is killed by Z_A.[11] Hence,

$$\begin{aligned}&\text{RHS}(D.72\sim D.75) && \text{(D.76)}\\ =\ &\frac{1}{4}\mathrm{tr}\Big((Z_A\otimes O)\Big(\Big(|1\rangle_A\langle 1|+|0\rangle_A\langle 0|\Big) && \text{(D.77)}\\ &\otimes\Big(U(\theta)\rho U^\dagger(\theta)+U(\theta+\pi)\rho U^\dagger(\theta+\pi)\Big) && \text{(D.78)}\\ &+\Big(|0\rangle_A\langle 0|-|1\rangle_A\langle 1|\Big) && \text{(D.79)}\\ &\otimes\Big(U(\theta)\rho U^\dagger(\theta+\pi)+U(\theta+\pi)\rho U^\dagger(\theta)\Big)\Big)\Big) && \text{(D.80)}\end{aligned}$$

Lastly, the behavior of $Z_A\otimes O$ ensures

$$\text{RHS}(D.77\sim D.80)=\frac{1}{2}\mathrm{tr}\Big(O\Big(U(\theta)\rho U^\dagger(\theta+\pi)+U(\theta+\pi)\rho U^\dagger(\pi)\Big)\Big),$$

as desired. From this point, we again adopt the notations from Lemma D.3 Eqns D.21, D.22.

3. (Sum Component)

$$\begin{aligned}&[\![(O,\rho)\to\Big(\frac{\partial}{\partial\theta}(\underline{S_0(\boldsymbol{\theta})+S_1(\boldsymbol{\theta})})\Big)]\!]\\ \overset{\text{Lem } D.3}{=}\ &[\![(O,\rho)\to\frac{\partial}{\partial\theta}(\underline{S_0(\boldsymbol{\theta})})]\!] && \text{(D.81)}\\ &+[\![(O,\rho)\to\frac{\partial}{\partial\theta}(\underline{S_1(\boldsymbol{\theta})})]\!] && \text{(D.82)}\\ \overset{\text{IH}}{=}\ &\frac{\partial}{\partial\theta}([\![(O,\rho)\to\underline{S_0(\boldsymbol{\theta})}]\!] && \text{(D.83)}\\ &+[\![(O,\rho)\to\underline{S_1(\boldsymbol{\theta})}]\!]) && \text{(D.84)}\\ \overset{\text{Lem } D.3}{=}\ &\frac{\partial}{\partial\theta}([\![(O,\rho)\to\underline{S_0(\boldsymbol{\theta})}+\underline{S_1(\boldsymbol{\theta})}]\!]), && \text{(D.85)}\end{aligned}$$

as desired.

4. (Sequence) Assume $\underline{S(\boldsymbol{\theta})}\equiv\underline{S_0(\boldsymbol{\theta})};\underline{S_1(\boldsymbol{\theta})}$; It suffices to show that $\forall\boldsymbol{\theta}^*$,

$$[\![(O,\rho)\to\frac{\partial}{\partial\theta}(\underline{S_0(\boldsymbol{\theta}^*)};\underline{S_1(\boldsymbol{\theta}^*)})]\!]=\Big(\frac{\partial}{\partial\theta}([\![(O,\rho)\to\underline{S_0(\boldsymbol{\theta})};\underline{S_1(\boldsymbol{\theta})}]\!])\Big|_{\boldsymbol{\theta}^*} \quad \text{(D.86)}$$

Let us first manipulate the equation using some computational properties developped from the helper lemmas:

$$\begin{aligned}&\text{LHS}(D.86) && \text{(D.87)}\\ \overset{\text{Lemma } D.3,\text{Eqn } D.23}{=}\ &[\![(O,\rho)\to\frac{\partial}{\partial\theta}(\underline{S_0(\boldsymbol{\theta}^*)});S_1(\boldsymbol{\theta}^*)]\!]\\ &+[\![(O,\rho)\to\underline{S_0(\boldsymbol{\theta}^*)};\frac{\partial}{\partial\theta}(S_1(\boldsymbol{\theta}^*))]\!]\end{aligned}$$

[11] Namely, the trace computation can be decomposed into a finite linear combination of the form $\mathrm{tr}((Z_A\otimes O)(|b\rangle_A\langle b'|\otimes\bullet))$ $(b,b'\in\{0,1\})$, and this term vanishes whenever $b\neq b'$.

This equals, by Lem D.3 Eqns D.30,D.31,

$$\sum_{g\in[0,s_0],h\in[0,t_1]} [\![(O,\rho)\to Q_{0,g}(\boldsymbol{\theta}^*);P_{1,h}(\boldsymbol{\theta}^*)]\!] \quad \text{(D.88)}$$

$$+\sum_{i\in[0,t_0],j\in[0,s_1]} [\![(O,\rho)\to P_{0,i}(\boldsymbol{\theta}^*);Q_{1,j}(\boldsymbol{\theta}^*)]\!] \quad \text{(D.89)}$$

Via Lem D.2, this translates to

$$\sum_{g\in[0,s_0],h\in[0,t_1]} [\![(O,[\![Q_{0,g}(\boldsymbol{\theta}^*)]\!]\rho)\to P_{1,h}(\boldsymbol{\theta}^*)]\!] \quad \text{(D.90)}$$

$$+\sum_{i\in[0,t_0],j\in[0,s_1]} [\![([\![Q_{1,j}(\boldsymbol{\theta}^*)]\!]^*(O),\rho)\to P_{0,i}(\boldsymbol{\theta}^*)]\!]$$

$$= \sum_{g\in s_0} [\![(O,[\![Q_{0,g}(\boldsymbol{\theta}^*)]\!]\rho)\to \frac{\partial}{\partial\theta}(\underline{S_1(\boldsymbol{\theta}^*)})]\!] \quad \text{(D.91)}$$

$$+\sum_{j\in s_1} [\![([\![Q_{1,j}(\boldsymbol{\theta}^*)]\!]^*(O),\rho)\to \frac{\partial}{\partial\theta}(\underline{S_0(\boldsymbol{\theta}^*)})]\!], \quad \text{(D.92)}$$

where the last step comes from definition of observable semantics with ancilla. Now apply the inductive hypothesis on $\underline{S_1(\boldsymbol{\theta})}$, a universal statement for all $(\rho',O')\in\mathcal{D}(\mathcal{H}_{\overline{v}})\times\overline{O_{\overline{v}}}$, to the instances $([\![Q_{0,g}(\boldsymbol{\theta}^*)]\!]\rho,O))$ $(\forall\partial\in[0,s_0])$, we have the first summand

$$\text{RHS}(D.91) = \sum_g \left(\frac{\partial}{\partial\theta}([\![(O,[\![Q_{0,g}(\boldsymbol{\theta}^*)]\!]\rho)\to\underline{S_1(\boldsymbol{\theta})}]\!])\right)\Big|_{\boldsymbol{\theta}^*} \quad \text{(D.93)}$$

Likewise, apply the inductive hypothesis on $\underline{S_0(\boldsymbol{\theta})}$ to the instances $(\rho,[\![Q_{1,j}(\boldsymbol{\theta}^*)]\!]^*O))$ $(\forall j\in[0,s_1])$, we have the second summand

$$\text{RHS}(D.92) = \sum_j \left(\frac{\partial}{\partial\theta}([\![([\![Q_{1,j}(\boldsymbol{\theta}^*)]\!]^*(O),\rho)\to\underline{S_0(\boldsymbol{\theta})}]\!])\right)\Big|_{\boldsymbol{\theta}^*} \quad \text{(D.94)}$$

Collecting everything, using again a techical lemma from above,

$$\text{RHS}(D.91)+\text{RHS}(D.92) = \text{RHS}(D.93)+\text{RHS}(D.94) \quad \text{(D.95)}$$

$$\overset{\text{Lem } D.4}{=} \text{RHS}(D.86),\ \text{as desired.} \quad \text{(D.96)}$$

5. (Case) This follows again from the computational lemmas above, and applications of all inductive hypothesis. For conciseness, let us denote by $\text{IH}(m)\mapsto(M_m\rho M_m^\dagger,O)$ the application of the m-th $(m\in[0,w])$ inductive hypothesis to the instance $(M_m\rho M_m^\dagger,O)\in\mathcal{D}(\mathcal{H}_{\overline{v}})\times O_{\overline{v}}$. Then,

$$[\![(O,\rho)\to\frac{\partial}{\partial\theta}(\underline{\textbf{case}}\ M[\overline{q}]=\overline{m\to\underline{S_m(\boldsymbol{\theta})}}\ \textbf{end})]\!]$$

$$\overset{\text{Lem } D.3}{=} \sum_m\sum_{i_m\in[0,t_m]} [\![(O,M_m\rho M_m^\dagger)\to P_{m,i_m}(\boldsymbol{\theta})]\!] \quad \text{(D.97)}$$

$$= \sum_m [\![(O,M_m\rho M_m^\dagger)\to\underline{\frac{\partial}{\partial\theta}(S_m(\boldsymbol{\theta}))}]\!] \quad \text{(D.98)}$$

Applying the inductive hypothesis $\text{IH}(m)\mapsto(M_m\rho M_m^\dagger,O)$, $\forall m$, we get that the above equals

$$\sum_m \frac{\partial}{\partial\theta}([\![(O,(M_m\rho M_m^\dagger)\rangle)\to\underline{S_m(\boldsymbol{\theta})}]\!])$$

$$= \frac{\partial}{\partial\theta}(\sum_m\sum_{j_m\in[0,s_m]}[\![(O,\mathcal{E}_m\rho\rangle)\to Q_{m,j_m}(\boldsymbol{\theta})]\!])$$

$$\overset{\text{Lem } D.3}{=} \frac{\partial}{\partial\theta}([\![(O,\rho)\to\textbf{case}\ M[\overline{q}]=\overline{m\to\underline{S_m(\boldsymbol{\theta})}}\ \textbf{end}]\!])$$

, as desired.

6. (While$^{(T)}$) This is verified by successive application of (Case) rule and (Sequence) rule T times.

□

E Detailed Proofs from Section 7

E.1 Proof for Proposition 7.2: Upper Bound of $|\#\frac{\partial}{\partial\theta_j}(P(\boldsymbol{\theta}))|$

Proof. 1. If $P(\boldsymbol{\theta})\equiv\textbf{abort}[\overline{v}]|\textbf{skip}[\overline{v}]|q:=|0\rangle$ $(q\in\overline{v})$, then $\text{OC}_j(P(\boldsymbol{\theta}))=0=|\#\frac{\partial}{\partial\theta_j}(P(\boldsymbol{\theta}))|$ by definition. Similarly, if $P(\boldsymbol{\theta})\equiv U(\boldsymbol{\theta})$ trivially used θ_j, no non-aborting programs exists in $\texttt{Compile}(\frac{\partial}{\partial\theta}(P(\boldsymbol{\theta})))$, yielding $|\#\frac{\partial}{\partial\theta_j}(P(\boldsymbol{\theta}))|=0$; otherwise, $|\#\frac{\partial}{\partial\theta_j}(P(\boldsymbol{\theta}))|$ consists of a single $R'_{\boldsymbol{\sigma}}(\boldsymbol{\theta})$, so both numbers are 1.

2. Assume $|\#\frac{\partial}{\partial\theta_j}(P_b(\boldsymbol{\theta}))|\le\text{OC}_j(P_b(\boldsymbol{\theta}))$ for each $b\in\{1,2\}$ and $P(\boldsymbol{\theta})\equiv P_1(\boldsymbol{\theta});P_2(\boldsymbol{\theta})$. Then

$$|\#\frac{\partial}{\partial\theta_j}(P(\boldsymbol{\theta}))|$$

$$\overset{(*)}{=} |\#(P_1(\boldsymbol{\theta});\frac{\partial}{\partial\theta_j}(P_2(\boldsymbol{\theta})))|$$

$$+|\#(\frac{\partial}{\partial\theta_j}(P_1(\boldsymbol{\theta}));P_2(\boldsymbol{\theta}))| \quad \text{(E.1)}$$

$$\le |\#(\frac{\partial}{\partial\theta_j}(P_2(\boldsymbol{\theta})))|$$

$$+|\#(\frac{\partial}{\partial\theta_j}(P_1(\boldsymbol{\theta})))| \quad \text{(E.2)}$$

$$\overset{\text{Inductive Hypothesis}}{\le} \text{OC}_j(P_1(\boldsymbol{\theta}))+\text{OC}_j(P_2(\boldsymbol{\theta})) \quad \text{(E.3)}$$

$$= \text{OC}_j(P(\boldsymbol{\theta})). \quad \text{(E.4)}$$

where step $(*)$ is via rules CT,CP-(ND-Component).

3. Assume $|\#\frac{\partial}{\partial\theta_j}(P_m(\boldsymbol{\theta}))|\le\text{OC}_j(P_m(\boldsymbol{\theta}))$ for each $m\in[0,w]$ and $P(\boldsymbol{\theta})\equiv\textbf{case}\ M[\overline{q}]=\overline{m\to P_m(\boldsymbol{\theta})}\ \textbf{end}$. Then by the CT and CP rules of **case**, $\frac{\partial}{\partial\theta_j}(P(\boldsymbol{\theta}))$ compiles to $\max_m|\#\frac{\partial}{\partial\theta_j}(P_m(\boldsymbol{\theta}))|$ programs. Assume $m^*\in[0,w]$ is

s.t. $|\#\frac{\partial}{\partial\theta_j}(P_{m^*}(\boldsymbol{\theta}))|$ attains that maximum. Then,

$$|\#\frac{\partial}{\partial\theta_j}(P(\boldsymbol{\theta}))| = |\#\frac{\partial}{\partial\theta_j}(P_{m^*}(\boldsymbol{\theta}))| \quad \text{(E.5)}$$

$$\overset{\text{Inductive Hypothesis}}{\leq} \mathrm{OC}_{m^*}(P_{m^*}(\boldsymbol{\theta})) \quad \text{(E.6)}$$

$$\leq \max_m \mathrm{OC}_m(P_m(\boldsymbol{\theta})) \quad \text{(E.7)}$$

$$= \mathrm{OC}_j(P(\boldsymbol{\theta})) \quad \text{(E.8)}$$

, as desired.

4. If $P(\boldsymbol{\theta}) \equiv \mathbf{while}^{(T)}\ M[\overline{q}] = 1\ \mathbf{do}\ P_1(\boldsymbol{\theta})\ \mathbf{done}$, it suffices to show $|\#\frac{\partial}{\partial\theta}(P(\boldsymbol{\theta}))| \leq T \cdot |\#\frac{\partial}{\partial\theta}(P_1(\boldsymbol{\theta}))|$. This is true for $T = 1$ as $\frac{\partial}{\partial\theta}(\mathbf{while}^{(1)})$ essentially aborts. By induction if this is true for T, then for $T+1$ we have

$$|\#\frac{\partial}{\partial\theta}(\mathbf{while}^{(T+1)}\ M[\overline{q}] = 1\ \mathbf{do}\ P_1(\boldsymbol{\theta})\ \mathbf{done})| \quad \text{(E.9)}$$

$$\leq |\#\frac{\partial}{\partial\theta}(P_1(\boldsymbol{\theta}); \mathbf{while}^{(T)}\ M[\overline{q}] = 1\ \mathbf{do}\ P_1(\boldsymbol{\theta})\ \mathbf{done})|$$

$$\overset{\text{Eqn } E.2}{\leq} |\#\frac{\partial}{\partial\theta}(P_1(\boldsymbol{\theta}))| + |\#\frac{\partial}{\partial\theta}(\mathbf{while}^{(T)}\ M[\overline{q}] = 1\ \mathbf{do}\ P_1(\boldsymbol{\theta})\ \mathbf{done})|$$

$$\leq (T+1)|\#\frac{\partial}{\partial\theta}(P_1(\boldsymbol{\theta}))|, \text{ as desired.} \quad \text{(E.10)}$$

□

F Evaluation Details

F.1 Training VQC instances with controls

For any parameter α, we exhibit the final results of $\texttt{Compile}(\frac{\partial P_1(\Theta,\Phi)}{\partial\alpha})$ and $\texttt{Compile}(\frac{\partial P_2(\Theta,\Phi,\Psi)}{\partial\alpha})$, with $P_1(\Theta,\Phi)$ and $P_2(\Theta,\Phi,\Psi)$ defined in the main text:

1. If $\alpha \in \Theta$: without loss of generality assume $\alpha = \theta_1$ (other situations are completely analogous). We denote:

$$Q'(\Theta) \equiv R'_X(\theta_1)|A, q_1\rangle; \cdots; R_Z(\theta_{12})|q_{12}\rangle$$

then

$$\begin{aligned}\texttt{Compile}(\frac{\partial P_1(\Theta,\Phi)}{\partial\alpha}) &\equiv \{|\, Q'(\Theta); Q(\Phi)\,|\},\\ \texttt{Compile}(\frac{\partial P_2(\Theta,\Phi,\Psi)}{\partial\alpha}) &\equiv \{|\, Q'(\Theta);\\ \mathbf{case}\ M[q_1] = 0 &\to\ Q(\Phi)\\ 1 &\to\ Q(\Psi)\,|\}.\end{aligned}$$

2. If $\alpha \in \Phi$: without loss of generality assume $\alpha = \phi_1$ (other situations are completely analogous). We denote:

$$Q'(\Phi) \equiv R'_X(\phi_1)|A, q_1\rangle; \cdots; R_Z(\phi_{12})|q_{12}\rangle$$

$$\begin{aligned}\texttt{Compile}(\frac{\partial P_1(\Theta,\Phi)}{\partial\alpha}) &\equiv \{|\, Q(\Theta); Q'(\Phi)\,|\},\\ \texttt{Compile}(\frac{\partial P_2(\Theta,\Phi,\Psi)}{\partial\alpha}) &\equiv \{|\, Q(\Theta);\\ \mathbf{case}\ M[q_1] = 0 &\to\ Q'(\Phi)\\ 1 &\to\ Q(\Psi)\,|\}.\end{aligned}$$

3. If $\alpha \in \Psi$: without loss of generality assume $\alpha = \psi_1$ (other situations are completely analogous). We denote:

$$Q'(\Psi) \equiv R'_X(\psi_1)|A, q_1\rangle; \cdots; R_Z(\psi_{12})|q_{12}\rangle$$

$$\begin{aligned}\texttt{Compile}(\frac{\partial P_1(\Theta,\Phi)}{\partial\alpha}) &\equiv \{|\, \mathbf{abort}[A, q_1, \cdots, q_{12}]\,|\},\\ \texttt{Compile}(\frac{\partial P_2(\Theta,\Phi,\Psi)}{\partial\alpha}) &\equiv \{|\, Q(\Theta);\\ \mathbf{case}\ M[q_1] = 0 &\to\ Q(\Phi)\\ 1 &\to\ Q'(\Psi)\,|\}.\end{aligned}$$

F.2 Benchmark testing on representative VQCs

In this section we introduce three examples from real-world quantum machine learning and quantum approximation algorithms, all of which are very promising candidates for implementation on near-term quantum devices. Without loss of generality, throughout this section θ denotes θ_1.

Our first case, $\texttt{QNN}_*$, starts from a slightly simplified case from an actual quantum neural-network that has been implemented on ion-trap quantum machines [56]. $\texttt{QAOA}_*$ is a quantum algorithm that produces approximate solutions for combinatorial optimization problems, which is regarded as one of the most promising candidates for demonstrating quantum supremacy [15, 47]. Quantum eigensolvers, crucial to quantum phase estimation algorithms, usually requires fully coherent evolution. In [35] Peruzzo et al. introduced an alternative approach that relaxes the fully coherent evolution requirement combined with a new approach to state preparation based on ansä and classical optimization [21, 30, 35], namely our third example $\texttt{VQE}_*$. The control flow creating multi-layer structures for the three families of examples are very similar, and these algorithms mainly differ from each other in their basic "rotation-entanglement" blocks. Therefore I will introduce the basic blocks for all three families, then use $\texttt{QNN}_*$ as an example to illustrate how the control flow (if, bounded while) works.

The basic "rotate-entangle" building block for QNN consists of a rotation stage and an entanglement stage – we consider these two stages together a single layer: in the rotation stage, one performs parameterized Z rotations followed by parameterized X rotations and then again parameterized Z rotations on the first 4 (small scale) or 6 qubits (medium or large scale); in the entanglement stage, one performs the parameterized $X \otimes X$ rotation on all pairs from the first 4 or 6 qubits. See Figure 7.

Basic block for VQE consists of three stages: the first stage is paramaterized X followed by parameterized Z; the second

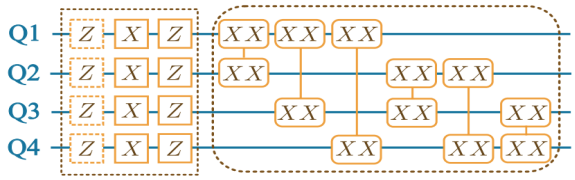

Figure 7. Circuit representation of the basic building block of QNN. Figure credit: [56]

stage uses H and CNOT to entangle; the third stage performs parameterized Z, X, Z in that order. Basic block for QAOA, on the other hand, entangles using H and CNOT in the first stage, and then performs parameterized X rotations on the second stage.

The more interesting part lies in building multiple layers using control flow, which we will explain using $\mathtt{QNN}_*$. The small scale if-controlled QNN (denoted as $\mathtt{QNN}_{S,i}$) is two-layered. Let B denote the basic rotate-entangle block for QNN explained as above; B', B'' be two similarly-structured rotate-entangle blocks using different parameter-qubit combinations. $\mathtt{QNN}_{S,i}$ performs B as the first layer, then measures on the first qubit, and performs B' or B'' as the second layer dependent on the measurement output. For medium and large scale if-controlled QNN (i.e. $\mathtt{QNN}_{M,i}$ and $\mathtt{QNN}_{L,i}$) we have larger rotate-entangle blocks, various parameter-qubit combinations (hence more of the B', B'''s involving different parameter-qubit combinations), and more layers of measurement-based control.

(Bounded) while-controled QNN works similarly: take $\mathtt{QNN}_{S,w}$ as an example, it performs the rotate-entangle block B as the first layer, then measure the first qubit; if it outputs 0 we halt; otherwise we perform some B', measure qubit $q1$ again, halts if outputs 0, performs B' the third time and aborts otherwise. Note that this is just a verbose way of saying "we wrapped B' with a 2-bounded while-loop"! And similarly, we build more layers on larger systems ($\mathtt{QNN}_{M,w}$, $\mathtt{QNN}_{L,w}$) using larger rotate-entangle blocks and more layers of bounded-while loops. One should note how bounded while is a succinct way to represent the circuits: as shown in Table 3, in $\mathtt{QNN}_{L,w}$ we managed to represent 2079 unitary gates in just 244 lines of code.

We auto-differentiated the three families of quamtum programs using our code transformation (hereinafter "CT") and code compilation (hereinafter "CP") rules. As shown in Table 3, the computation outputs the desired multiset of derivative programs, and the number of non-aborting programs ($|\#\frac{\partial}{\partial\theta}(P(\boldsymbol{\theta}))|$) agrees with the upper bound described in Proposition 7.2. It should be noted that Table 3 indicates our auto-differentiation scheme works well for variously sized input programs, be the size measured by code length, gate count, layer count or qubit count, to name a few. In particular, a 40-qubit system is very close to a real world quantum machine learning system developed by Google. For details of the code and the differentiation result, please see supplementary materials.

$P(\boldsymbol{\theta})$	OC(·)	$\lvert\#\frac{\partial}{\partial\theta}()\rvert$	#gates	#lines	#layers	#qb's
$\mathtt{QNN}_{S,b}$	1	1	20	24	1	4
$\mathtt{QNN}_{S,s}$	5	5	20	24	1	4
$\mathtt{QNN}_{S,i}$	10	10	60	67	2	4
$\mathtt{QNN}_{S,w}$	15	10	60	66	3	4
$\mathtt{QNN}_{M,i}$	24	24	165	189	3	18
$\mathtt{QNN}_{M,w}$	56	24	231	121	5	18
$\mathtt{QNN}_{L,i}$	48	48	363	414	6	36
$\mathtt{QNN}_{L,w}$	504	48	2079	244	33	36
$\mathtt{VQE}_{S,b}$	1	1	14	16	1	2
$\mathtt{VQE}_{S,s}$	2	2	14	16	1	2
$\mathtt{VQE}_{S,i}$	4	4	28	38	2	2
$\mathtt{VQE}_{S,w}$	6	4	42	32	3	2
$\mathtt{VQE}_{M,i}$	15	15	224	241	3	12
$\mathtt{VQE}_{M,w}$	35	15	224	112	5	12
$\mathtt{VQE}_{L,i}$	40	40	576	628	5	40
$\mathtt{VQE}_{L,w}$	248	40	1984	368	17	40
$\mathtt{QAOA}_{S,b}$	1	1	12	15	1	3
$\mathtt{QAOA}_{S,s}$	3	3	12	15	1	3
$\mathtt{QAOA}_{S,i}$	6	6	36	41	2	3
$\mathtt{QAOA}_{S,w}$	9	6	36	29	3	3
$\mathtt{QAOA}_{M,i}$	18	18	120	142	3	18
$\mathtt{QAOA}_{M,w}$	42	18	168	94	5	18
$\mathtt{QAOA}_{L,i}$	36	36	264	315	6	36
$\mathtt{QAOA}_{L,w}$	378	36	1512	190	33	36

Table 3. Output on Test Examples. Note that: (1) $\{S, M, L\}$ stands for "small, medium, large"; $\{b, s, i, w\}$ stands for "basic, shared, if, while"; #lines is the number of lines for input programs. (2) gate count and layer count for T-bounded while is done by multiplying the corresponding count for loop body with T. (3) for $*_{*,w}$ we have $|\#\frac{\partial}{\partial\theta}(P(\boldsymbol{\theta}))| < \mathrm{OC}(P(\boldsymbol{\theta}))$ because differentiating the unrolled bounded while generates essentially aborting programs which are optimized out of the final multiset.