



Optimized numerical gradient and Hessian estimation for variational quantum algorithmsY. S. Teo *Department of Physics and Astronomy, Seoul National University, 08826 Seoul, South Korea* (Received 1 July 2022; revised 26 September 2022; accepted 29 March 2023; published 17 April 2023)

Sampling noisy intermediate-scale quantum devices is a fundamental step that converts coherent quantum-circuit outputs to measurement data for running variational quantum algorithms that utilize gradient and Hessian methods in cost-function optimization tasks. This step, however, introduces estimation errors in the resulting gradient or Hessian computations. To minimize these errors, we discuss tunable numerical estimators, which are the finite difference (including their generalized versions) and scaled parameter-shift estimators [introduced in *Phys. Rev. A* **103**, 012405 (2021)], and propose operational circuit-averaged methods to optimize them. We show that these optimized numerical estimators offer estimation errors that drop exponentially with the number of circuit qubits for a given sampling-copy number, revealing a direct compatibility with the barren-plateau phenomenon. In particular, there exists a critical sampling-copy number below which an optimized difference estimator gives a smaller average estimation error in contrast to the standard (analytical) parameter-shift estimator, which exactly computes gradient and Hessian components. Moreover, this critical number grows exponentially with the circuit-qubit number. Finally, by forsaking analyticity, we demonstrate that the scaled parameter-shift estimators beat the standard unscaled ones in estimation accuracy under any situation, with comparable performances to those of the difference estimators within significant copy-number ranges, and are the best ones if larger copy numbers are affordable.

DOI: [10.1103/PhysRevA.107.042421](https://doi.org/10.1103/PhysRevA.107.042421)**I. INTRODUCTION**

With the inception of quantum information theory [1], quantum computers and devices [2–4] that function according to the laws of quantum mechanics have been envisioned to be the new-age tools for performing computations and other information processing tasks. The subsequent identification of universal gate sets [5–9] motivated many innovative proposals for quantum-computation and cryptographic algorithms [10–16]. Despite the theoretical progress, there exist practical challenges that hinder the actual implementation of truly operational quantum devices. These include maintaining the fidelities of qubit sources, unitary gates and measurements [17–19], and coping with the large gate complexity needed to construct general-purpose quantum circuits [20].

The state of the art in quantum computing technologies revolves around devices that manipulate less than 1000 qubits using noisy unitary gates and measurements—the noisy intermediate-scale quantum (NISQ) devices [21]. These devices motivated the development of several kinds of NISQ algorithms [22–30], of which the class of variational quantum algorithms (VQAs) [31–35] that perform computations in a hybrid manner using both classical and NISQ devices, most commonly discussed in the context of variational quantum eigensolvers designed for quantum-chemistry [36–38] and combinatorial problems [39,40], are of relatively broad interest. In the field of quantum machine learning, VQAs running on circuits that also possess classical-data encodings have also been extensively studied. These include algorithms for classification tasks, nonlinear activation-function implementations, and multivariate function learning tasks [41–47].

Cost-function optimization with VQAs typically requires the statistical sampling of NISQ devices to estimate the gradient and Hessian of the quantum-circuit model function, which are necessary in, for example, steepest gradient-descent [48–51] and quantum natural gradient-descent methods [52–56]. Sampling NISQ devices inherently comes with errors originating from statistical fluctuation in the quantum-circuit measurements, which is especially relevant to NISQ devices as currently achievable noise levels forbid arbitrarily large error-mitigated measurement-data collection within reasonable algorithm run times. While the severity of this problem has indeed been raised [34,57] and asymptotic error bounds for sampling the Fisher information in quantum natural gradient methods were derived [58], more precise error expressions in estimating multiparameter gradients and Hessians on NISQ devices are necessary for developing novel methods that are statistically optimized for VQA executions.

In this paper, we examine the estimation accuracies of known methods used to estimate circuit-function gradients and Hessians, namely the (generalized) finite-difference strategy and (scaled) parameter-shift rule [59–61]. All of these methods are numerical except for the unscaled parameter-shift rule, which is analytical: the former approximates gradients and Hessians with a nonzero approximation error, and the latter exactly computes them. We present operational analytical expressions for the averaged estimation errors associated with these two kinds of strategies. All expressions are circuit averaged in contrast to those reported in Ref. [61], for instance, which permits the introduction of operationally tunable numerical estimators possessing parameters that can be optimally tuned to minimize NISQ estimation errors. These optimal estimators are designed for a broad class of

hardware-efficient quantum circuits that approximate two-design unitary operators, such as the multilayered *ansatz* comprising single-qubit and controlled-NOT (CNOT) gates, for which these estimators minimize estimation errors in the initial stages of cost minimization.

A key observation is that for a given sampling-copy number, the minimized average estimation errors of all numerical estimators scale commensurately with the average gradient and Hessian-component magnitudes, which in turn drop exponentially with the number of circuit qubits. Without increasing measurement resources, these desirable scaling behaviors prevent the optimally tuned estimators from effectively making random guesses about the estimated components even in the presence of the barren-plateau phenomenon [62–65]. Owing to this characteristic, we show that these optimal numerical estimators can outperform those produced by the analytical parameter-shift rule, which does not possess such a characteristic. One striking consequence is that all regimes of sampling-copy numbers in which the optimal (generalized) finite-difference strategies beat the analytical strategy grow exponentially with the circuit-qubit number.

Last, but not the least, we demonstrate that when one forgoes analyticity and, instead, employ the scaled parameter-shift rule [61], which is yet another numerical strategy, we find that its estimation accuracy is comparable to those of the numerical difference strategies in orders of magnitude for a certain range of sampling-copy numbers. Beyond this range, the scaled parameter-shift rule exhibits the most favorable estimation accuracy. This further confirms that numerical estimation schemes are better suited for improving gradient and Hessian estimation accuracies as one scales up NISQ devices.

II. BACKGROUND: VARIATIONAL QUANTUM ALGORITHMS

An especially important and widely studied computation task is function minimization. In various interdisciplinary applications that are related to quantum mechanics and quantum information, the cost function $C = C[\langle H_j \rangle]$ to be minimized is a (real) functional of expectation values of an (Hermitian) operator set $\{H_j\}$. The expectation $\langle H_j \rangle = \text{tr}[\rho H_j]$ is itself a function of a variable state ρ that is optimized in order to attain the minimum value of C . An immediate problem with the minimization task is the difficulty in evaluating expectation values of operators describing large physical systems, such as multiqubit systems considered in Fig. 1(a), using a classical computer.

In the NISQ era when fully quantum algorithms are out of reach, VQAs are the next viable class of quantum-classical hybrid algorithms for efficient cost-function minimization [see Fig. 1(b)]. It makes use of a quantum device, a unitary circuit for instance, to efficiently collect sampled data of expectation values. These data are then transferred to a classical computer that performs an iterative update on the current quantum-circuit parameters using a prechosen optimization scheme, which are then used to tune the quantum device for another round of sampling. The action of quantum-circuit tuning is also widely termed quantum-circuit training, borrowing terminologies from machine learning. The entire iterative VQA terminates after the cost function C is mini-

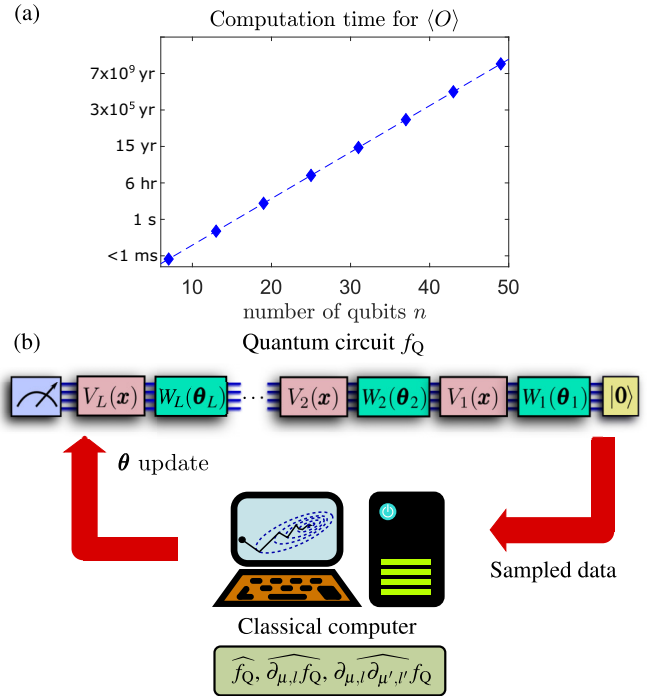


FIG. 1. (a) Classical computation of $\langle O \rangle$ on an AMD Ryzen 9 5900HX CPU. For every n , the computation time is averaged over 1000 expectation-value calculations from randomly chosen n -qubit Hermitian operators O and pure states. Owing to memory limitations, a combination of actual computation-time data for observables up to $n = 14$ followed by a fitted extrapolation to $n = 50$ shows that the average computation time with randomly generated pure states grows exponentially with n . (b) Schematic of a VQA for minimizing a cost function $C = C[f_Q]$, which is generally a functional of another parametrized function $f_Q = f_Q(\theta; \mathbf{x})$. Data collected from sampling a NISQ circuit that models f_Q using a series of parametrized training $[W_l(\theta_l)]$ and classically encoded $[V_l(\mathbf{x})]$ modules are fed to a classical machine, where the model function f_Q and its gradient $\partial_{\mu,l} f_Q$ (and Hessian $\partial_{\mu,l} \partial_{\mu',l'} f_Q$ if necessary) are estimated for carrying out a prechosen optimization scheme in an iterative fashion. Here, the pair (μ, l) labels the μ th trainable circuit parameter $\theta_{\mu l}$ located in the l th trainable module W_l . In each step, the updated circuit parameters using the estimated quantities are used to tune the training modules $W_l(\theta_l)$ in the NISQ circuit for subsequent sampling and classical optimization.

mized. An arbitrary quantum circuit of the NISQ device used to run the VQA is a sequence of training $[W_l(\theta_l)]$ and classically encoded $[V_l(\mathbf{x})]$ unitary operators, where θ_l are trainable parameters and \mathbf{x} are nontrainable ones.

The unitary operator $U_{\theta, \mathbf{x}} \equiv \prod_{l=L}^1 V_l(\mathbf{x}) W_l(\theta_l)$ of a finite depth L represents the most general quantum-circuit model for all VQA applications; here, θ and \mathbf{x} are shorthand for the respective complete sets of parameters. If the fixed initial pure product state $|\mathbf{0}\rangle\langle\mathbf{0}| = (|0\rangle\langle 0|)^{\otimes n}$ is prepared, the relevant cost function $C = C[f_Q(\theta; \mathbf{x})]$ is then defined in terms of the circuit model function

$$f_Q(\theta; \mathbf{x}) = \langle \mathbf{0} | U_{\theta, \mathbf{x}}^\dagger O U_{\theta, \mathbf{x}} | \mathbf{0} \rangle \quad (1)$$

and measurement observable O . As specific examples of VQAs, in variational quantum eigensolver (VQE) prob-

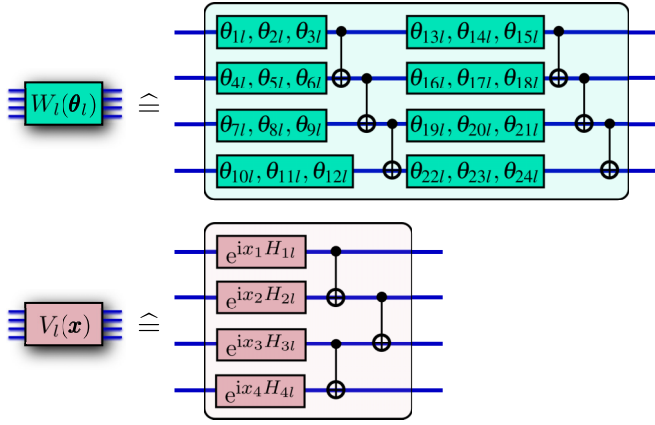


FIG. 2. Examples of hardware-efficient setups for trainable unitary operators W_l (with two repeated units) and those of fixed encodings V_l , shown here for a four-qubit system with $\theta_l = (\theta_{1l} \theta_{2l} \dots \theta_{24l})^\top$ and $\mathbf{x} = (x_1 \dots x_4)^\top$. Each single-qubit green block in W_l represents a rotation operator defined by an angle triple ϕ . For instance, $R(\phi = (\theta_{1l}, \theta_{2l}, \theta_{3l})) = R_Z(\theta_{1l})R_Y(\theta_{2l})R_Z(\theta_{3l})$. The operator V_l may also take similar structures, where x is encoded into entanglement-free unitary operators characterized by their respective generators H_{kl} along with CNOT gates.

lems [36–38] and quantum approximate optimization algorithms (QAOA) [39,40], only trainable operators $W_l(\theta_l)$ are used to minimize the linear cost function $C = \langle \mathbf{0} | U_\theta^\dagger H U_\theta | \mathbf{0} \rangle$, where $O = H$ is a Hamilton operator that either describes the dynamics of a physical system, or corresponds to a combinatorial problem. In quantum machine learning tasks [41–47], a VQA is employed, for instance, to train the quantum circuit defined by $U_{\theta;x}$ to learn a particular multivariate function mapping $f(x)$ for different classical-data encoding parameters \mathbf{x} . Given m of these parameters $\{\mathbf{x}_j\}_{j=1}^m$, the quality of the learning procedure is defined by more sophisticated cost functions such as the mean-squared error $C = \sum_{j=1}^m [f_Q(\theta; \mathbf{x}_j) - f(\mathbf{x}_j)]^2 / m$.

Gradient-based (and Hessian-based) optimization routines in VQAs would then rely on the computation accuracies of f_Q , $\partial_{\mu,l} f_Q$ and $\partial_{\mu,l} \partial_{\mu',l'} f_Q$, with all arguments dropped from hereon for notational simplicity unless otherwise necessary. The estimations of gradient $\partial_{\mu,l} f_Q$ and Hessian components $\partial_{\mu,l} \partial_{\mu',l'} f_Q$ require the specifications of actual physical *ansatz* structures that make up the W_l and V_l operators. A very common type of parametrized quantum-circuit multiqubit *Ansätze* encode training parameters θ on single-qubit rotation unitary operators that are easy to manipulate. They consist of alternating layers of single-qubit rotation and CNOT gates [1,66–68], examples of which are illustrated in Fig. 2. In general, we may further decompose any single-qubit rotation gate $R(\phi)$ defined by the trivariate angular parameter $\phi = (\phi_1 \phi_2 \phi_3)^\top$, $R(\phi) = R_Z(\phi_1)R_Y(\phi_2)R_Z(\phi_3)$, into a sequence of rotations about the Y and Z axes, so that each basic Pauli rotation gate $R_j(\phi_j) = e^{-i\phi_j \sigma_j / 2}$ is attributed to the relevant single-qubit Pauli operator σ_j of the j th axis.

To set the stage for analyzing gradient and Hessian estimations, we will consider the following hardware-efficient scenario:

(1) We focus on Pauli-encoded parametrized quantum circuits (PEPQCs). Such a circuit comprises an arbitrary chain of trainable (W_l) and nontrainable (V_l) encoding modules as shown in Fig. 1, where all trainable parameters θ in W_l are encoded onto Pauli rotation operators that constitute single-qubit rotation gates, and CNOT gates will be used to generate entanglement between all qubits (as in Fig. 2). These circuits are widely employed in common VQAs (such as quantum eigensolvers and quantum machine learning) and other quantum tasks [66–71].

(2) The structures of V_l that house the nontrainable parameters \mathbf{x} can be arbitrary.

(3) With no loss of generality, the measurement observable is some traceless Hermitian operator $O = \sum_k h_k O_k / \|\mathbf{h}\|$ written in terms of the multiqubit Pauli basis operators O_k . Any Hermitian operator is then O displaced by a multiple of the identity up to normalization. We will assume that each O_k is sampled independently for the same set of PEPQC parameters.

We note that the set of W_l s, of circuit depths polynomial in the number of qubits n , consisting of randomized single-qubit rotation and CNOT gates, are also approximately two-design circuits [72]; that is, the operator moments $\langle W_l \rangle$ and $\langle W_l^{\otimes 2} \rangle$ are approximately those from the Haar measure [73] over the $U(2^n)$ group.

III. FIGURE OF MERIT: MEAN-SQUARED ERROR

Throughout the paper, separate notations for averages of different kinds are adopted to avoid confusion when several kinds appear at the same time. Averages over (unitary) operator spaces are denoted by the angled parentheses $\langle \dots \rangle$, which, for instance, could mean averages according to the Haar measure. Numerical or vectorial averages over NISQ-sampling distributions will be denoted by $\mathbb{E}[\dots]$. Those over the nontrainable parameters \mathbf{x} are denoted by \dots .

To explicitly quantify errors for estimating the model function f_Q , its gradient ($\partial_{\mu,l} f_Q$) and Hessian components ($\partial_{\mu,l} \partial_{\mu',l'} f_Q$) in any gradient- or Hessian-based methods, we will examine the mean-squared error (MSE)

$$\mathcal{D}(Y) = \overline{\langle (\widehat{Y} - Y)^2 \rangle}, \quad (2)$$

where \widehat{Y} is some generic estimator of the true component Y , which can refer to either the quantum-circuit function, its gradient or Hessian component, obtained from sampling the NISQ device. As we will soon realize, estimators of the latter two can be computed from direct sampling of quantum-circuit functions [defined in Eq. (1)] evaluated at various translated circuit parameters.

For VQAs running on $(d = 2^n)$ -dimensional circuits, measurements of the circuit observable $O = \sum_k h_k O_k / \|\mathbf{h}\|$ are performed independently with respect to the individual Pauli components O_k (see, for instance, Ref. [36]). Using the spectral decomposition of the Pauli observable $O_k = \sum_{l=0}^{d-1} |o_{kl}\rangle o_{kl} \langle o_{kl}|$, sampling the individual function components

$$f_{Q,k}(\theta; \mathbf{x}) = \sum_{l=0}^{d-1} o_{kl} |\langle o_{kl} | U_{\theta;x} | \mathbf{0} \rangle|^2 \equiv \sum_{l=0}^{d-1} o_{kl} p_{kl}(\theta; \mathbf{x}) \quad (3)$$

of $f_Q = \sum_k h_k f_{Q,k} / \|\mathbf{h}\|$ is equivalent to sampling the circuit probabilities $\sum_{l=0}^{d-1} p_{kl}(\boldsymbol{\theta}; \mathbf{x}) = 1$, where the eigenvalues $O_{kl} = \pm 1$. Since $\widehat{Y} - Y = \sum_k h_k (\widehat{Y}_k - Y_k) / \|\mathbf{h}\|$, it can be deduced easily, as a result of independence sampling with the O_{kl} s, that $\mathcal{D}(Y) = \sum_k h_k^2 \mathcal{D}(Y_k) / \|\mathbf{h}\|^2$ if the estimators \widehat{Y}_k s are unbiased ($\mathbb{E}[\widehat{Y}_k] = Y_k$).

A usual physical circumstance when sampling a circuit function defined by the measurement observable O_k is when detectors measuring the probabilities $p_{kl}(\boldsymbol{\theta}; \mathbf{x})$ register clicks at different detection sites l one at a time in a randomized sequence, where each click is statistically independent from the rest, so that a total of N clicks are registered. The resulting sampling distribution is multinomial, where the relative frequencies $v_{kl}(\boldsymbol{\theta}; \mathbf{x}) = n_{kl}(\boldsymbol{\theta}; \mathbf{x}) / N \rightarrow p_{kl}(\boldsymbol{\theta}; \mathbf{x})$ corresponding to the number of clicks $n_{kl}(\boldsymbol{\theta}; \mathbf{x})$ registered at the l th detector, with $N = \sum_{l=0}^{d-1} n_{kl}(\boldsymbol{\theta}; \mathbf{x})$. Thus, the average accuracy in estimating f_Q using the unbiased estimator

$$\widehat{f}_Q = \sum_k \frac{h_k}{\|\mathbf{h}\|} \widehat{f}_{Q,k} = \sum_k \frac{h_k}{\|\mathbf{h}\|} \sum_{l=0}^{d-1} O_{kl} v_{kl} \quad (4)$$

is quantified by the MSE.

We caution the reader that the MSEs involve circuit averages of gradient and Hessian components that are generally nontrivial to evaluate. As the present discussion hinges on the two-design approximative PEPQC *ansatz*, explicit MSE expressions are only available when trainable circuits are sufficiently deep for two-design modules to exist. It could very well be that a gradient operation separates a two-design module into two subcircuits, each of which may or may not be deep enough to be a two-design. In most cases, only MSE upper bounds are calculable. While explicit details on how circuit averaging is carried out are supplied in Appendix A 3, we simply state that throughout the main text, unless otherwise stated, all analytical MSE expressions will refer to the so-called two-design sandwich (TDS) condition (Case I in Fig. 7), where every gradient operation is sandwiched between two two-design trainable modules. Upper-bound expressions for all other cases are obtained from Table I in Appendix B, and also Appendix C.

IV. RESULTS: SAMPLING ERRORS IN GRADIENT AND HESSIAN ESTIMATIONS

A. Finite (generalized) difference gradient and Hessian methods

The well-known finite-difference (FD) numerical strategy [75,76] approximates the gradient components $\partial_{\mu,l} f_{Q,k}$ for each observable basis operator O_k according to

$$\begin{aligned} [\partial_{\text{FD}}]_{\mu,l}^\epsilon f_{Q,k} &= \frac{f_{Q,k}(\theta_{\mu l} + \epsilon/2; \mathbf{x}) - f_{Q,k}(\theta_{\mu l} - \epsilon/2; \mathbf{x})}{\epsilon} \\ &= \text{sinc}(\epsilon/2) \partial_{\mu l} f_{Q,k} \end{aligned} \quad (5)$$

by sampling the two quantum-circuit functions $f_{Q,k}(\theta_{\mu l} + \epsilon/2; \mathbf{x})$ and $f_{Q,k}(\theta_{\mu l} - \epsilon/2; \mathbf{x})$, each with trainable parameters displaced by equal magnitudes of $\epsilon/2$ for some $\epsilon > 0$. The second equality originates from the usage of Eq. (A7). This particular form approximates $\partial_{\mu,l} f_{Q,k}$ up to $O(\epsilon^2)$ since $\text{sinc}(\epsilon/2) \cong 1 - \epsilon^2/24$ for a small ϵ . As a consequence of Eq. (A7) as well, the Hessian approximator for $\partial_{\mu,l} \partial_{\mu',l'} f_{Q,k}$, defined by applying a second operation $[\partial_{\text{FD}}]_{\mu',l'}^\epsilon$

on $[\partial_{\text{FD}}]_{\mu,l}^\epsilon f_{Q,k}$ in (5), reads

$$\begin{aligned} &[\partial_{\text{FD}}]_{\mu,l}^\epsilon [\partial_{\text{FD}}]_{\mu',l'}^\epsilon f_{Q,k} \\ &= [\text{sinc}(\epsilon/2)]^2 \partial_{\mu,l} \partial_{\mu',l'} f_{Q,k} \\ &= \frac{f_Q(\theta_{\mu l} + \frac{\epsilon}{2}, \theta_{\mu' l'} + \frac{\epsilon}{2}; \mathbf{x}) - f_Q(\theta_{\mu l} + \frac{\epsilon}{2}, \theta_{\mu' l'} - \frac{\epsilon}{2}; \mathbf{x})}{\epsilon^2} \\ &\quad - \frac{f_Q(\theta_{\mu l} - \frac{\epsilon}{2}, \theta_{\mu' l'} + \frac{\epsilon}{2}; \mathbf{x}) - f_Q(\theta_{\mu l} - \frac{\epsilon}{2}, \theta_{\mu' l'} - \frac{\epsilon}{2}; \mathbf{x})}{\epsilon^2}. \end{aligned} \quad (6)$$

The corresponding function estimators that enter $[\partial_{\text{FD}}]_{\mu,l}^\epsilon f_{Q,k}$ and $[\partial_{\text{FD}}]_{\mu,l}^\epsilon [\partial_{\text{FD}}]_{\mu',l'}^\epsilon f_{Q,k}$ take the form stated in Eq. (4). Thus, three independent function expectation values are measured for estimating each diagonal ($\mu = \mu'$ and $l = l'$) Hessian approximator component, and four independent function expectation values are measured for each off-diagonal ($\mu \neq \mu'$ and/or $l \neq l'$) component.

The nonzero- ϵ gradient and Hessian approximators of the FD strategy, collectively denoted by $\widehat{Y}_{k,\epsilon} = [\partial_{\text{FD}}]_{\mu,l}^\epsilon f_{Q,k}$, $[\partial_{\text{FD}}]_{\mu,l}^\epsilon [\partial_{\text{FD}}]_{\mu',l'}^\epsilon f_{Q,k}$, incur errors from both finite-copy sampling and nonzero- ϵ approximation. The corresponding MSE is hence decomposable into these two error types:

$$\begin{aligned} \mathcal{D}(Y) &= \overline{\langle \mathbb{E}[(\widehat{Y}_{k,\epsilon} - Y_k)^2] \rangle} \\ &= \underbrace{\overline{\langle \mathbb{E}[(\widehat{Y}_{k,\epsilon} - Y_{k,\epsilon})^2] \rangle}}_{\text{finite-copy error}} + \underbrace{\overline{\langle (Y_{k,\epsilon} - Y_k)^2 \rangle}}_{\text{approx. error}} \\ &\equiv \Delta_{\text{copy}}^2(Y_{k,\epsilon}) + \Delta_\epsilon^2(Y_k), \end{aligned} \quad (7)$$

where $\Delta_{\epsilon=0}^2(Y_k) = 0$, since $Y_{k,\epsilon=0} = Y_k = \partial_{\mu,l} f_{Q,k}$, $\partial_{\mu,l} \partial_{\mu',l'} f_{Q,k}$ according to the definitions in (5) and (6). The finite-copy error Δ_{copy}^2 for the gradient and Hessian FD estimators can be directly calculated using (C3) and (C4) as it only involves linear combinations of squared circuit functions. The nonzero- ϵ approximation error Δ_ϵ^2 , on the other hand, requires the evaluation of $\langle (\partial_{\mu,l} f_{Q,k})^2 \rangle$ and $\langle (\partial_{\mu,l} \partial_{\mu',l'} f_{Q,k})^2 \rangle$ over random circuit parameters.

With that said, by defining the total number of copies N_T distributed equally to all sampled quantum-circuit functions for one FD approximator per circuit-observable basis operator (that is, $N_T = 2N$ for $[\partial_{\text{FD}}]_{\mu,l}^\epsilon f_{Q,k}$, and $N_T = 3N$ and $4N$, respectively, for the diagonal and off-diagonal $[\partial_{\text{FD}}]_{\mu,l}^\epsilon [\partial_{\text{FD}}]_{\mu',l'}^\epsilon f_{Q,k}$ components), we list the MSE formulas for estimating gradient and Hessian components using the FD strategy that is applicable to any PEPQC under the TDS condition:

$$\begin{aligned} \mathcal{D}_{\text{FD}}(\partial f_Q) &= \underbrace{\frac{4d}{N_T(d+1)\epsilon^2}}_{\text{finite-copy error}} + \underbrace{\frac{d^2[1 - \text{sinc}(\epsilon/2)]^2}{2(d+1)(d^2-1)}}_{\text{approx. error}}, \\ \mathcal{D}_{\text{FD}}(\partial \partial f_Q) &= \frac{18d}{N_T(d+1)\epsilon^4} + \frac{d^2\{1 - [\text{sinc}(\epsilon/2)]^2\}^2}{2(d+1)(d^2-1)}, \\ &\quad (\text{diagonal Hessian components}) \\ \mathcal{D}_{\text{FD}}(\partial \partial' f_Q) &= \frac{16d}{N_T(d+1)\epsilon^4} + \frac{d^4\{1 - [\text{sinc}(\epsilon/2)]^2\}^2}{4(d+1)(d^2-1)^2}. \\ &\quad (\text{off-diagonal Hessian components}) \end{aligned} \quad (8)$$

For optimal estimation of each kind of components using either (5) or (6), the value of ϵ is therefore chosen such that it minimizes the corresponding operational \mathcal{D}_{FD} in (8).

We may also consider a type of generalized difference (GD) estimation strategy [77] where the corresponding nonzero- ϵ gradient and Hessian approximators,

$$[\partial_{\text{GD}}]_{\mu,l}^{J,\epsilon} f_{Q,k} \equiv \sum_{j=1}^J c_j [\partial_{\text{FD}}]_{\mu,l}^{j\epsilon} f_{Q,k}, \quad (9)$$

$$[\partial \partial'_{\text{GD}}]_{\mu,l;\mu',l'}^{J,\epsilon} f_{Q,k} \equiv \sum_{j=1}^J c_j [\partial_{\text{FD}}]_{\mu,l}^{j\epsilon} [\partial_{\text{FD}}]_{\mu',l'}^{j\epsilon} f_{Q,k}, \quad (10)$$

are weighted sums of FD approximators of integer multiples of the step size ϵ , which become the true gradient and Hessian components when $\epsilon = 0$ under the normalization constraint $\sum_{j=1}^J c_j = \mathbf{c}^\top \mathbf{I} = 1$ for the coefficient column \mathbf{c} , where \mathbf{I} is the J -dimensional column of ones. This strategy estimates each GDgrad, diagonal and off-diagonal GDHess component by, respectively, sampling $2J$, $2J + 1$, and $4J$ function expectation values, consistent with the number of measurements needed when $J = 1$. The forms of their MSE expressions under the TDS condition are rather technical and are instead given in Appendix A 4.

B. Parameter-shift rule

Both the FD and GD numerical strategies can in general be applied to approximate gradients and Hessians of any function that is not restricted to those originating from PEPQCs. A common criticism against these strategies is that the FD and GD approximators require extremely small ϵ to achieve good approximation qualities. For PEPQCs, using the identity in Eq. (A7), it is indeed easy to verify that, for any $s \geq 0$,

$$\partial_{\mu,l} f_Q = [\partial_{\text{PS}}]_{\mu,l} f_Q = \frac{f_Q(\theta_{\mu l} + s; \mathbf{x}) - f_Q(\theta_{\mu l} - s; \mathbf{x})}{2 \sin s}, \quad (11)$$

$$\begin{aligned} \partial_{\mu,l} \partial_{\mu',l'} f_Q &= [\partial_{\text{PS}}]_{\mu,l} [\partial_{\text{PS}}]_{\mu',l'} f_Q \\ &= \frac{f_Q(\theta_{\mu l} + s, \theta_{\mu' l'} + s; \mathbf{x}) - f_Q(\theta_{\mu l} + s, \theta_{\mu' l'} - s; \mathbf{x})}{4(\sin s)^2} \\ &\quad - \frac{f_Q(\theta_{\mu l} - s, \theta_{\mu' l'} + s; \mathbf{x}) - f_Q(\theta_{\mu l} - s, \theta_{\mu' l'} - s; \mathbf{x})}{4(\sin s)^2}. \end{aligned} \quad (12)$$

That there exist exact gradient and Hessian expressions for PEPQCs *via* simple training-parameter translations prompted the term parameter-shift rule (PS) [59–61]. The corresponding MSEs are therefore just finite-copy errors given by

$$\begin{aligned} \mathcal{D}_{\text{PS}}(\partial f_Q) &= \frac{d}{N_T(d+1)(\sin s)^2} \geq \frac{d}{N_T(d+1)}, \\ \mathcal{D}_{\text{PS}}(\partial \partial f_Q) &= \frac{9d}{8N_T(d+1)(\sin s)^4} \geq \frac{9d}{8N_T(d+1)}, \\ &\quad \text{(diagonal Hessian components)} \\ \mathcal{D}_{\text{PS}}(\partial \partial' f_Q) &= \frac{d}{N_T(d+1)(\sin s)^4} \geq \frac{d}{N_T(d+1)}. \\ &\quad \text{(off-diagonal Hessian components)} \end{aligned} \quad (13)$$

These MSEs are minimized when $s = \pi/2$, which is the standard shift value that we will consider for these analytical estimators.

The analytical PS strategy is now widely accepted as the go-to approach for estimating gradient and Hessian components. An attractive feature is the absence of approximation errors ($\Delta_\epsilon^2 = 0$), unlike FD or GD methods. Because of this, it is a belief that FD, for instance, which requires small ϵ values, would necessitate a large N_T in order to achieve comparable estimation errors. On the contrary, in Secs. V and VI, we show that there exist exponentially growing sampling regimes where optimally tuned FD and GD strategies can achieve very small estimation errors and outperform even the standard PS method.

C. Scaled parameter-shift rule

The standard PS, along with the entire class of parameter-shift rules, form the analytical strategy that exactly computes gradients and Hessians for PEPQCs and some other types of quantum-circuit *ansatz*. For any particular shift value s , there is no other free parameter characterizing the PS estimators.

In Ref. [61], the scaled parameter-shift (SPS) estimators were introduced. These estimators are essentially scaled versions of the PS estimators, where $[\partial_{\text{SPS}}]_{\mu,l} f_Q = \lambda [\partial_{\text{PS}}]_{\mu,l} f_Q$ and $[\partial_{\text{SPS}}]_{\mu,l} [\partial_{\text{SPS}}]_{\mu',l'} f_Q = \lambda [\partial_{\text{PS}}]_{\mu,l} [\partial_{\text{PS}}]_{\mu',l'} f_Q$ are characterized by an additional prefactor λ that ranges from zero to one. For arbitrary shift values of s and prefactor magnitudes, one can similarly arrive at the following TDS accuracy expressions:

$$\begin{aligned} \mathcal{D}_{\text{SPS}}(\partial f_Q) &= \frac{\overbrace{d\lambda^2}^{\text{finite-copy error}}}{N_T(d+1)(\sin s)^2} + \frac{\overbrace{d^2(1-\lambda)^2}^{\text{approx. error}}}{2(d+1)(d^2-1)}, \\ \mathcal{D}_{\text{SPS}}(\partial \partial f_Q) &= \frac{9d\lambda^2}{8N_T(d+1)(\sin s)^4} + \frac{d^2(1-\lambda)^2}{2(d+1)(d^2-1)}, \\ &\quad \text{(diagonal Hessian components)} \\ \mathcal{D}_{\text{SPS}}(\partial \partial' f_Q) &= \frac{d\lambda^2}{N_T(d+1)(\sin s)^4} + \frac{d^4(1-\lambda)^2}{4(d+1)(d^2-1)^2}. \\ &\quad \text{(off-diagonal Hessian components)} \end{aligned} \quad (14)$$

It is clear that $s = \pi/2$ will optimize all SPS MSEs. We will see that λ can be easily optimized to further enhance estimation accuracies. Notice that the introduction of these prefactors immediately results in the loss of analyticity, since these SPS estimators, for all $\lambda < 1$, no longer exactly compute the correct gradient and Hessian components, and therefore carry approximation errors just like the FD and GD estimators. The expressions for general cases are found in Appendix C 5.

V. RESULTS: OPTIMALLY TUNED NUMERICAL ESTIMATORS

A. Optimal FD estimators

From the results in (8), (A10), (A11), and (13), the first visual observation is that these formulas are functions of

only N_T , the number of qubits n of the circuit, and ϵ . The second observation is that the MSEs are oscillatory functions of ϵ by virtue of the unitary encoding. The third observation has to do with the choice of ϵ . If one picks very small ϵ values, then the finite-copy error Δ_{copy}^2 dominates as $O(1/\text{poly } \epsilon)$. If one, instead, picks larger ϵ values, then the approximation error Δ_ϵ^2 eventually catches up. The optimal $\epsilon = \epsilon_{\text{opt}} = \epsilon_{\text{opt}}(d, N_T)$ minimizes the combination $\text{MSE} = \Delta_{\text{copy}}^2 + \Delta_\epsilon^2$.

Computing ϵ_{opt} through minimizing \mathcal{D}_{FD} over ϵ requires the solutions of transcendental equations that do not generally admit analytical forms. Numerical optimization methods are therefore in order. Nevertheless, if $[1 - \text{sinc}(\epsilon/2)]^2 \approx \epsilon^4/576$ and $\{1 - [\text{sinc}(\epsilon/2)]^2\}^2 \approx \epsilon^4/144$ for small right-hand sides, we may obtain analytical approximations of both ϵ_{opt} and $\mathcal{D}_{\text{FD,opt}}$ for optimal FD gradient and Hessian estimation by approximating Δ_ϵ^2 to the smallest order $O(\epsilon^4)$. Minimizing the resulting leading-order expansions of all \mathcal{D}_{FD} therefore gives

$$\begin{aligned}\epsilon_{\text{opt}} &\cong \left(2304 \frac{d^2 - 1}{N_T d}\right)^{\frac{1}{6}} & (\partial f_Q \text{ estimation}); \\ \epsilon_{\text{opt}} &\cong \left(5184 \frac{d^2 - 1}{N_T d}\right)^{\frac{1}{8}} & (\partial \partial f_Q \text{ estimation}); \\ \epsilon_{\text{opt}} &\cong \left[9216 \frac{(d^2 - 1)^2}{N_T d^3}\right]^{\frac{1}{8}} & (\partial \partial' f_Q \text{ estimation}).\end{aligned}\quad (15)$$

The corresponding approximately optimal MSEs are

$$\begin{aligned}\mathcal{D}_{\text{FD,opt}}(\partial f_Q) &\cong \left(\frac{3}{32}\right)^{1/3} \frac{d^{4/3}}{(d+1)(d^2-1)^{1/3} N_T^{2/3}}, \\ \mathcal{D}_{\text{FD,opt}}(\partial \partial f_Q) &\cong \frac{d^{3/2}}{2(d+1)(d^2-1)^{1/2} N_T^{1/2}}, \\ \mathcal{D}_{\text{FD,opt}}(\partial \partial' f_Q) &\cong \frac{d^{5/2}}{3(d+1)(d^2-1) N_T^{1/2}}.\end{aligned}\quad (16)$$

We note that the formulas in (15) and (16) become accurate so long as N_T is sufficiently larger than d [that is, $N_T \gg O(2^n)$]. These formulas can therefore give us a fairly satisfactory description of the estimation errors.

Perhaps the most striking feature of all $\mathcal{D}_{\text{FD,opt}}$ s (be it approximated in the $N_T \gg d$ regime or not) is the fact that they decrease exponentially with the number of qubits n . This is desirable since it is shown in Appendices B 2–B 4 (summarized in Table I) that $\langle (\partial f_Q)^2 \rangle$, $\langle (\partial \partial f_Q)^2 \rangle$ and $\langle (\partial \partial' f_Q)^2 \rangle$ are all at most $O(1/d)$, which are manifestations of the so-called barren-plateau phenomenon [62–65]. Hence, this feature confirms that optimally tuned FD estimators are natural for gradient and Hessian estimations.

Ordinarily, things become more difficult to estimate as n increases, so this feature is interestingly counterintuitive. To understand this behavior, we first note that the average f_Q landscape rapidly flattens with increasing number of qubits (owing to the barren-plateau phenomenon). For very large n , the barren-plateau phenomenon is hence the bottleneck of VQAs. Next, using Table I, the optimized FD and PS

gradient estimators, for instance, can be shown to have the following average squared magnitudes in the limit of large d :

$$\text{opt. FD} : \overline{\langle \mathbb{E}[\widehat{\partial f_Q}^2] \rangle} \rightarrow [\text{sinc}(\epsilon_{\text{opt}}/2)]^2 \frac{1}{2d} + \frac{4}{N_T \epsilon_{\text{opt}}^2}, \quad (17)$$

$$\text{PS} : \overline{\langle \mathbb{E}[\widehat{\partial f_Q}^2] \rangle} \rightarrow \frac{1}{2d} + \frac{1}{N_T} \rightarrow \frac{1}{N_T}. \quad (18)$$

Since from numerical evidence, ϵ_{opt} grows exponentially for sufficiently large n (refer to the later Sec. VII for a concise summary discussion), the average FD estimator squared magnitude drops exponentially in n commensurately with the true components, so that these estimators themselves also approach zero for sufficiently large n . As a result, the difference between the estimator and true component converges to zero as n increases. On the other hand, in the large- d limit, (18) implies that the average PS estimator squared magnitude, and thus the corresponding MSE_{PS} is constant for the same N_T .

As a reminder, ϵ_{opt} s obtained from minimizing the MSEs in (8) apply when the TDS condition holds (see Sec. IV A). More generally, similar analyses for FD gradient and Hessian estimation that are not under the TDS condition are presented in Appendix C 3. We will take the TDS-based formulas in (8) as representatives useful for obtaining optimal FD estimators.

We additionally note that the $N_T^{-2/3}$ and $N_T^{-1/2}$ scaling behaviors in the first and second equations of (16) were also reported in Ref. [61]. The expressions presented there are large- N forms that are not averaged over quantum circuits, and therefore make no reference to average behaviors in d or n .

B. Optimal GD estimators

Optimizing each GD estimator requires the optimization of both \mathbf{c} and ϵ . For any ϵ , it is easy to carry out the minimization of the MSEs in (A10) over normalized \mathbf{c} , as every GD MSE takes the form $\mathcal{D}_{\text{GD}} = \mathbf{c}^\top \mathbf{M} \mathbf{c}$. Upon the standard usage of a Lagrange multiplier (see Appendix C 4), it is straightforward to arrive at

$$\min_{\mathbf{c} | \mathbf{c}^\top \mathbf{I} = 1} \mathcal{D}_{\text{GD}} = (\mathbf{I}^\top \mathbf{M}^{-1} \mathbf{I})^{-1}, \quad (19)$$

where $\mathbf{c}_{\text{opt}} = \mathbf{M}^{-1} \mathbf{I} / (\mathbf{I}^\top \mathbf{M}^{-1} \mathbf{I})$ and \mathbf{M} is any of the relevant matrices in (A11). It is clear that the right-hand side of Eq. (19) reduces to the MSE expressions for FD when $J = 1$.

For $J > 1$, there appears to be neither exact nor even approximate analytical forms for ϵ_{opt} , although numerical minimization of \mathcal{D}_{GD} s over ϵ to obtain the optimal $\mathcal{D}_{\text{GD,opt}}$ is efficient so long as J is not too large. Despite the lack of analytical formulas, it is still possible to acquire upper bounds of $\mathcal{D}_{\text{GD,opt}}$ as N_T and d grows. For this purpose, we first make use of the Cauchy-Schwarz inequality $(\mathbf{I}^\top \mathbf{I})^2 \leq \mathbf{I}^\top \mathbf{M} \mathbf{I} \mathbf{I}^\top \mathbf{M}^{-1} \mathbf{I}$ to arrive at the bound

$$\begin{aligned}\mathcal{D}_{\text{GD,opt}} &= \min_{\epsilon > 0} \min_{\mathbf{c} | \mathbf{c}^\top \mathbf{I} = 1} \mathcal{D}_{\text{GD}} \\ &= \min_{\epsilon > 0} (\mathbf{I}^\top \mathbf{M}^{-1} \mathbf{I})^{-1} \leq \frac{1}{J^2} \min_{\epsilon > 0} \mathbf{I}^\top \mathbf{M} \mathbf{I}.\end{aligned}\quad (20)$$

The inequality is saturated when $J = 1$. Next, under the TDS condition and $N_T \gg d$ (or small ϵ_{opt}), we can obtain the fol-

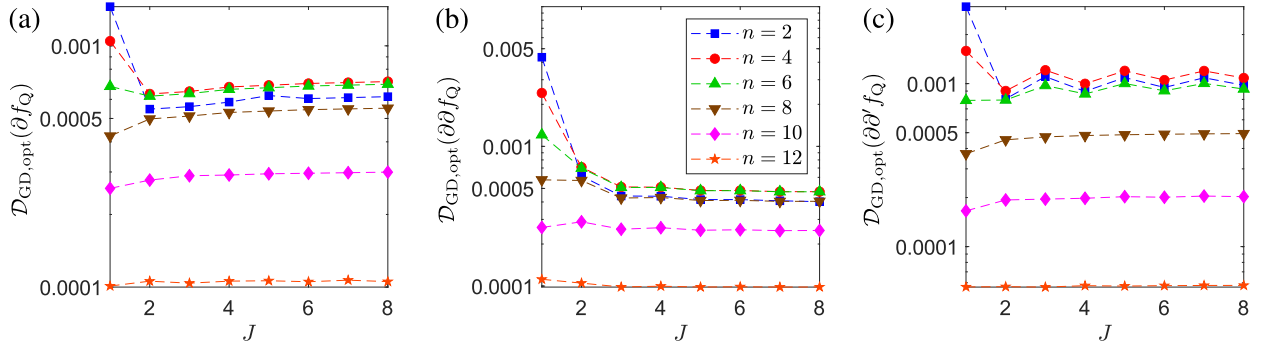


FIG. 3. Graphs of $\mathcal{D}_{\text{GD,opt}}$ (after analytically minimizing over ϵ in accordance with (19), followed by numerical minimization over ϵ) for (a) gradient and (b), (c) Hessian components under the TDS condition defined in Sec. IV A, presented for $1 \leq J \leq 8$, even number of qubits n , and $N_T = 2000$. It turns out that for estimating gradient and off-diagonal Hessian components, $J = 2$ appears to be optimal for $n \leq 6$, beyond which $J = 1$ gives the smallest $\mathcal{D}_{\text{GD,opt}}$. For estimating diagonal Hessian components, $J = 3$ seems to be the optimal choice, although the differences in estimation errors quickly vanish for large J and n values.

lowing explicit upper bounds:

$$\begin{aligned} \mathcal{D}_{\text{GD,opt}}(\partial f_Q) &\lesssim \frac{1}{4} \left[\frac{d^4(J+1)^2(2J+1)^2}{6N_T^2 J^2 (d+1)^3 (d^2-1)} \right]^{1/3} \mathbf{H}_{J,2}^{2/3}, \\ \mathcal{D}_{\text{GD,opt}}(\partial\partial f_Q) &\lesssim \frac{J+1}{36J} \left(\frac{2dJ+d}{d+1} \right)^{3/2} \sqrt{\frac{2\mathbf{H}_{J,2}^2 + \mathbf{H}_{J,4}}{N_T(d-1)}}, \\ \mathcal{D}_{\text{GD,opt}}(\partial\partial' f_Q) &\lesssim \frac{(J+1)(2J+1)}{18(d+1)(d^2-1)} \left(\frac{d^5 \mathbf{H}_{J,4}}{N_T J} \right)^{1/2}, \end{aligned} \quad (21)$$

where $\mathbf{H}_{j,k} = \sum_{m=1}^j m^{-k}$ is the generalized harmonic number. It is obvious that all right-hand sides reduce to the expressions in (8) when $J = 1$. Based on these crude upper bounds, one can already deduce the fundamental trend: $\mathcal{D}_{\text{GD,opt}}$ decreases with increasing d and N_T , which comes as no surprise on hindsight since the GD strategy is a generalization of the FD strategy, and naturally preserves the exponential-decay-in- n characteristic.

For any d and N_T , resorting to numerical optimization for deciding on the optimal value of J is inevitable. For a prechosen N_T , number of qubits n employed by the quantum circuit and the type of components (gradient or Hessian) to be estimated, the optimal value of J is the one that minimizes the numerical minimum $\mathcal{D}_{\text{GD,opt}}$. As examples, Fig. 3 illustrates the graphs of the respective $\mathcal{D}_{\text{GD,opt}}$ s for $N_T = 2000$ using various n values. In these cases, there exist optimal values of J below a certain critical n . Beyond this critical value, $J = 1$ is sufficient as larger J values generally do not significantly vary $\mathcal{D}_{\text{GD,opt}}$.

C. Optimal SPS estimators

The minimization of (14) with respect to λ , unlike the FD and GD strategies, can be carried out exactly for any d and N_T since this simply amounts to the minimization of quadratic functions in λ on the right-hand sides. Hence, one arrives at

the optimal scaling prefactors

$$\begin{aligned} \lambda_{\text{opt}} &= \frac{dN_T}{2d^2 + dN_T - 2} \quad (\partial f_Q \text{ estimation}), \\ \lambda_{\text{opt}} &= \frac{4dN_T}{9d^2 + 4dN_T - 9} \quad (\partial\partial f_Q \text{ estimation}), \\ \lambda_{\text{opt}} &= \frac{d^3 N_T}{4(d^2 - 1)^2 + d^3 N_T} \quad (\partial\partial' f_Q \text{ estimation}), \end{aligned} \quad (22)$$

and realizes that $\lambda_{\text{opt}} \rightarrow 1$ as $d \ll N_T \rightarrow \infty$, and $\lambda_{\text{opt}} \rightarrow 0$ as $N_T \ll d \rightarrow \infty$. The former limit for any fixed d is obvious, since in the large-data limit, scaling prefactors are not really necessary and all SPS estimation performances approach to those of PS. On the other hand, for a fixed N_T , sampling from a circuit that has a large number of qubits would pay off with very small prefactors.

Consequently, the corresponding optimized MSE expressions for these optimized SPS estimators are given by

$$\begin{aligned} \mathcal{D}_{\text{SPS,opt}}(\partial f_Q) &= \frac{d^2}{(d+1)(2d^2 + dN_T - 2)}, \\ \mathcal{D}_{\text{SPS,opt}}(\partial\partial f_Q) &= \frac{9d^2}{2(d+1)(9d^2 + 4dN_T - 9)}, \\ &\quad \text{(diagonal Hessian components)} \\ \mathcal{D}_{\text{SPS,opt}}(\partial\partial' f_Q) &= \frac{d^4}{(d+1)[4(d^2 - 1)^2 + d^3 N_T]}. \\ &\quad \text{(off-diagonal Hessian components)} \end{aligned} \quad (23)$$

For a fixed d , these optimized SPS MSEs converge to the PS MSEs in (13), as they should.

VI. PERFORMANCE

A. Applications in quantum supervised learning and quantum eigensolver problems

We will first demonstrate the performance of optimal FD and GD estimators in typical applications of VQAs. As a benchmark, we compare the estimation errors of these optimal estimators with those from PS estimators, using the

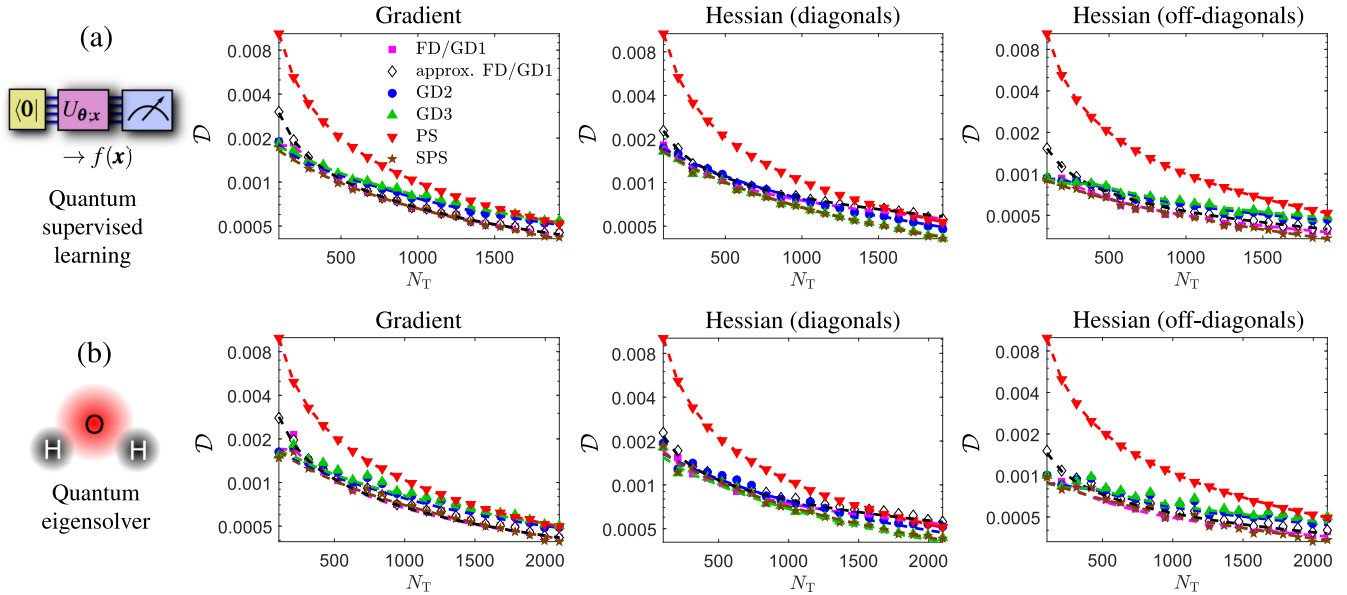


FIG. 4. Averaged performance plots of gradient- and Hessian-component estimation, generated *via* Monte Carlo simulations, in the first optimization step of VQAs pertaining to (a) quantum supervised learning that learns a (multivariate) function $f(x)$, and (b) a quantum eigensolver problem ($V_l = 1$ for all $1 \leq l \leq L$) that searches for the ground-state energy of a water molecule having restricted excitation levels, with each plot marker averaged over 500 random PEPQCs ($L = 5$ and four-repeated units of single-qubit and CNOT gates for each W_l) and 500 numerical experiments per PEPQC. All PEPQCs possess the Hilbert-space dimension $d = 2^8$, which are used to estimate $\partial_{50,1} f_Q$ (Case II), $(\partial_{50,1})^2 f_Q$ (Case II), and $\partial_{50,1} \partial_{49,2} f_Q$ (Case V). All FD (\equiv GD1 or $J = 1$), GD2 ($J = 2$), and GD3 ($J = 3$) estimators are optimized according to Sec. V [“approx.” refers to the approximated optimization from (15)], where the dashed curves represent the corresponding TDS analytical expressions in (8), (A10), (A11), and (13). As the cases (see also Fig. 7) considered here are different from the TDS case (Case I), these TDS curves act as guiding lines, which also show that the actual optimal MSEs do not deviate very far from them. For estimating gradient and off-diagonal Hessian components, both SPS and FD are comparable in performance, whereas it is GD3 that matches with SPS in estimating diagonal Hessian components.

MSE as figure of merit. The first important example of a VQA is quantum machine learning, where a PEPQC that defines the quantum model $f_Q(x) = \langle \mathbf{0} | U_{\theta,x}^\dagger O U_{\theta,x} | \mathbf{0} \rangle$ is trained to learn or express a general multivariate function $f(x)$ [$|f(x)| \leq 1$ for all x with no loss of generality] by minimizing an appropriate cost function. In this case, it is sufficient to assign the observable circuit O as the multiqubit Pauli operator $Y \otimes 1^{n-1}$.

The second widely studied example concerns quantum-eigensolver tasks that search for minimum eigenvalues of operators. Here, the observable O is one such operator of interest that is typically Hermitian (for example, a Hamilton operator describing the dynamics of an electronic system in a molecule), and hence, decomposable as a linear combination of multiqubit Pauli operators O_k . For this second application, we will consider a simplified (trace-subtracted) Hamilton operator O that describes the electrons in a water molecule with restricted excitation levels. Using the Jordan-Wigner transformation, one may write O as a weighted sum of 96 eight-qubit Pauli operators (see Appendix D).

Figure 4 showcases the estimation-error performances of gradient and Hessian estimation in some of the physical cases listed in Fig. 7 for the two aforementioned examples of VQA applications. As a benchmark, the results indeed confirm that optimal FD and GD strategies outperform the PS strategy for N_T below certain critical values that would depend on d and the types of components. We also remark that although the analytical curves in Fig. 4 are strictly meant for components

under Case I in Fig. 7 that is equivalent to the TDS condition, we observe, through these and other numerical evidence not shown here, that the analytical results in (8) and (A10) supplying those curves can also approximate the estimation errors for other cases well. Another important sanity verification from the figure is that the optimal FD approximators defined by the respective ϵ_{optS} prescribed in (15), along with their MSEs in (16), quickly converge to the exact optimal curves with increasing N_T .

The estimation accuracy greatly improves when analyticity is forsaken in SPS, where the optimization of the respective scaling prefactors with respect to the averaged MSEs offer comparable estimation performances with the optimal FD and GD strategies. As N_T increases, SPS eventually becomes the most efficient strategy.

B. Benefits of optimized numerical estimators for scalable NISQ devices

Relative to the standard PS strategy, the scaled version, SPS, is statistically more efficient in estimating gradient and Hessian components. This is immediately clear from either a direct comparison of (23) with (13) under the TDS condition, or the simple arguments in Appendix C 5 for all other cases: for any $d \geq 2$ and $N_T > 0$, $\mathcal{D}_{\text{SPS}}(\cdot) < \mathcal{D}_{\text{PS}}(\cdot)$.

That the optimized FD and GD strategies could give smaller estimation errors than PS for a significantly large regime of N_T can be understood by noting that the PS estima-

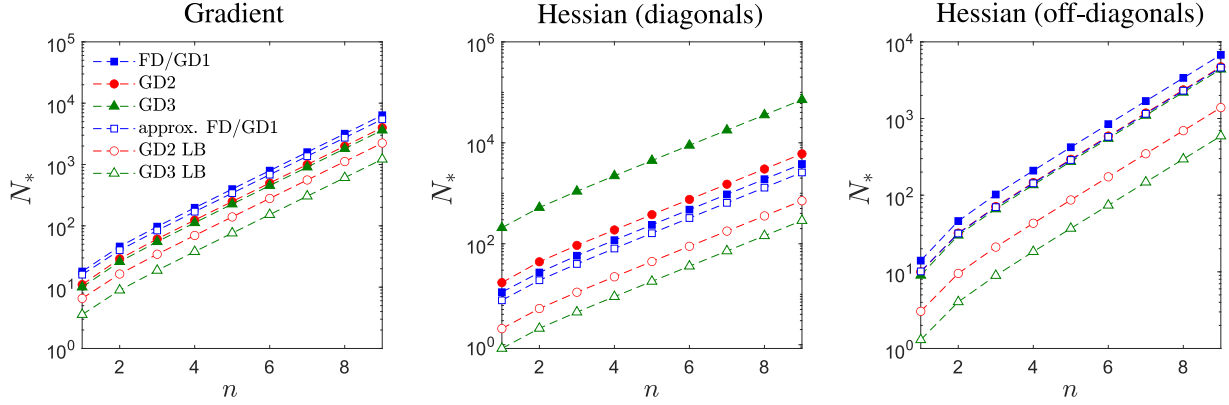


FIG. 5. Exponential increase of N_* in gradient and Hessian estimation schemes under the TDS condition. The N_* values corresponding to the approximated FD strategy [(24)], and those given by the cruder lower bounds of both the GD2 [$J = 2$, GD2 LB] and GD3 [$J = 3$, GD3 LB] strategies in (25), all serve as lower bounds to the actual N_* values computed respectively using the correct ϵ_{opt} s obtained from numerical optimization of MSEs in (8) and (A10) over ϵ . For each estimation scheme, a larger N_* is indicative of a greater sampling efficiency over the PS strategy.

tors defined in Eqs. (11) and (12) correspond to an effective $\epsilon \cong \pi$ that is generally not optimal for minimizing the estimation error—the combined errors Δ_{copy}^2 and Δ_{ϵ}^2 . When N_T is greater than a certain critical value N_* (defined as the value of N_T for which $\mathcal{D}_{\text{FD/GD,opt}} = \mathcal{D}_{\text{PS}}$), the contribution of Δ_{copy}^2 is small enough to be dominated by the nonzero- ϵ error Δ_{ϵ}^2 , such that the advantage of an ϵ -error-free PS estimator manifests itself with a smaller MSE relative to those of the FD or GD approximators.

To further support the usefulness of optimal FD and GD schemes, we answer the important question: How does N_* scale with the number of qubits n ? Basic intuition suggests that since the FD and GD estimation errors decreases with n according to Sec. V A, while the PS ones do not [recall (13)], the critical value $N_T = N_*$ required for the PS strategy to start outperforming the former schemes would also grow with n . Indeed, based on the approximately minimized MSE's in (16) for the FD strategy under the TDS condition and the regime $N_* \gg d$, we find that

$$\begin{aligned} N_* &\cong \frac{32(d^2 - 1)}{3d} \quad (\partial f_Q \text{ estimation}); \\ N_* &\cong \frac{81(d^2 - 1)}{16d} \quad (\partial \partial f_Q \text{ estimation}); \\ N_* &\cong \frac{9(d^2 - 1)^2}{d^3} \quad (\partial \partial' f_Q \text{ estimation}). \end{aligned} \quad (24)$$

For the GD strategy under the TDS condition, the upper bounds derived and stated in (21) conveniently permit us to write down approximate and loose lower bounds of N_* :

$$\begin{aligned} N_* &\gtrsim \frac{384J^2(d^2 - 1)}{d(J+1)^2(2J+1)^2 H_{J,2}^2} \quad (\partial f_Q \text{ estimation}); \\ N_* &\gtrsim \frac{6561J^2(d^2 - 1)}{4d(J+1)^2(2J+1)^3(2H_{J,2}^2 + H_{J,4})} \quad (\partial \partial f_Q \text{ est.}); \\ N_* &\gtrsim \frac{6561J(d^2 - 1)^2}{16d^3(J+1)^2(2J+1)^2 H_{J,4}} \quad (\partial \partial' f_Q \text{ estimation}). \end{aligned} \quad (25)$$

While Eqs. (24) and (25) strictly hold only when $N_* \gg d$, they analytically show, at least in this regime, that $N_* \gtrsim O(2^n)$. The plots in Fig. 5 clearly shows an exponential increase in N_* with respect to n regardless of whether ϵ_{opt} is found with numerical MSE optimization or large N_T approximation. In particular, for the optimized FD strategy, the differences between the exact numerically obtained N_* s and those from (24) are very small. In general, N_* is a useful measure for the sampling efficiency of a particular optimized scheme in question. A larger N_* implies that the optimized numerical scheme gives a smaller estimation error for a larger range of $1 \leq N_T \leq N_*$ in contrast to the analytical PS strategy. A scheme that exhibits an exponentially growing N_* with respect to n is therefore a much more statistically favorable one over PS for scalable VQAs. Figure 5 illustrates the exponential growth in N_* with respect to n for estimations performed under the TDS condition. More general arguments in Appendices B and C, which are applicable to all cases in Fig. 7, technically guarantee an exponentially growing N_* with n for all FD and GD estimation schemes. With this, optimally tuned FD and GD strategies may be regarded as prime candidates for scalable VQAs, especially on NISQ platforms where estimating large- n circuit-model expectation values with large numbers of sampling copies is practically infeasible.

VII. IMPORTANT REMARKS AND POTENTIAL PITFALL

The results in this paper show that numerical estimators possessing free parameters (ϵ for FD and GD and λ for SPS) can be optimized to yield more accurate gradient and Hessian estimation than analytical estimators (PS) that do not possess such a freedom. The optimization refers to the minimization of the relevant circuit-averaged MSE—an estimation-error quantifier for the circuit function, gradient and Hessian—of a given circuit *ansatz* and sampling-copy number N_T with respect to the free parameter. We recall that the average is performed over not just the click data per training circuit, but also over all possible training-circuit parameters according to the *ansatz* structure.

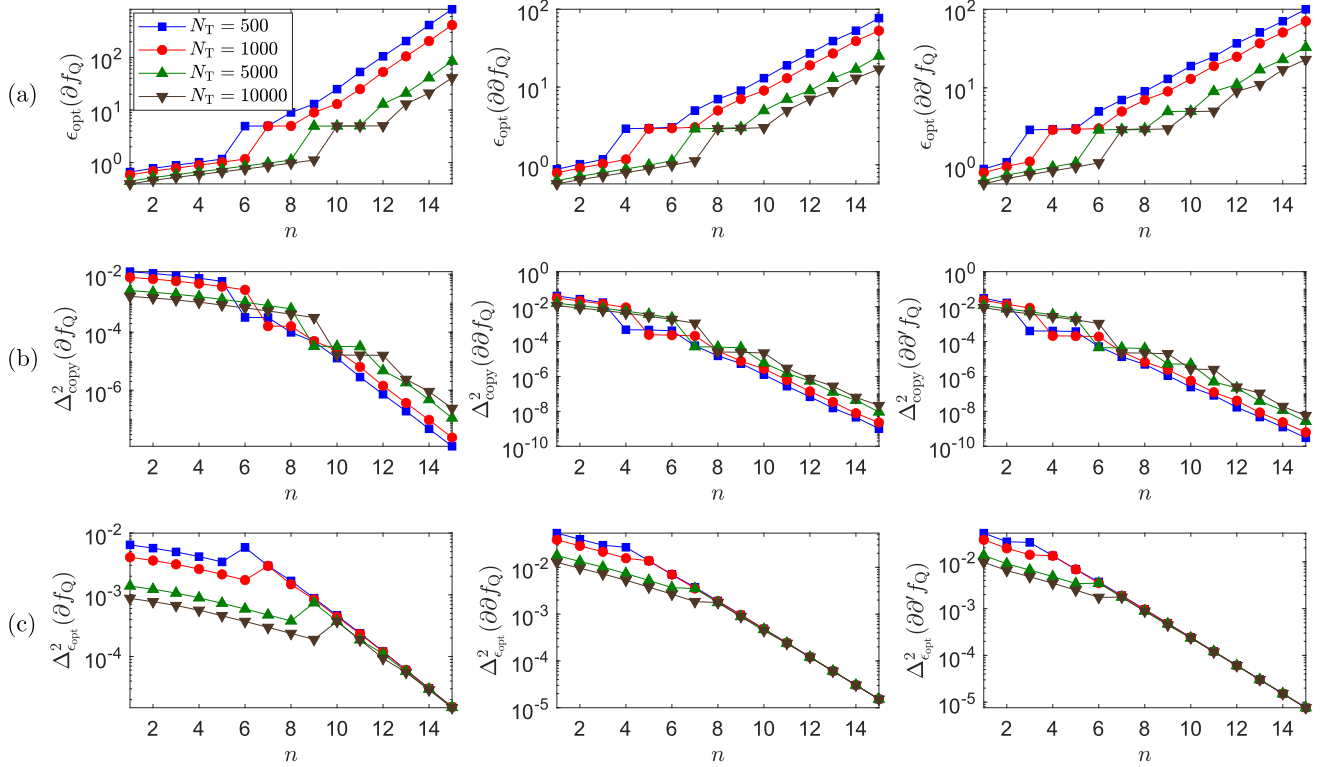


FIG. 6. Graphs of (a) ϵ_{opt} , (b) finite-copy error Δ_{copy}^2 , and (c) nonzero- ϵ_{opt} approximation error $\Delta_{\epsilon_{\text{opt}}}^2$ for all three types of components estimated with the FD strategy under various qubit numbers n and N_T . All ϵ_{opt} s are obtained by numerically minimizing the respective TDS MSE expressions in (8) with no approximations. Numerical evidence shows that ϵ_{opt} ultimately grows exponentially in n . At the same time, we see that both Δ_{copy}^2 and $\Delta_{\epsilon_{\text{opt}}}^2$ also asymptotically drop exponentially with n . Note that the exponential drop in $\Delta_{\epsilon_{\text{opt}}}^2$ is solely due to n . As a special case, when n is small so that $N_T \gg d$, ϵ_{opt} is clearly small. Otherwise, the magnitude of ϵ_{opt} can be large by virtue of the behaviors of Δ_{copy}^2 and $\Delta_{\epsilon_{\text{opt}}}^2$.

For optimized FD and GD schemes, we reiterate that ϵ_{opt} is consequently dependent on the circuit *ansatz* and N_T . For the PEPQC *ansatz* that leads to two-design approximating modules considered in this work, the resulting MSE expressions, as shown for instance in Eq. (8) and (A10), are nonlinear functions of ϵ , so that numerical methods for their minimization are the only resort. Otherwise, analytical approximations of ϵ_{opt} such as those in (15) for the FD strategy may be employed when $N_T \gg d$.

Moreover, rather unintuitively at first glance, we find that ϵ_{opt} is typically not small, regardless of whether numerical optimization or analytical approximations are invoked. In fact, as d increases, numerical experience shows that ϵ_{opt} grows roughly exponentially with n . One can already witness this behavior approximately from (15). This is, apparently, at odds with the usual narrative that ϵ_{opt} should be as small as possible, preferably $\epsilon_{\text{opt}} \rightarrow 0$, in order for the approximation of gradient and Hessian components to be as accurate as possible. While such a narrative is surely correct when no sampling is required to estimate the components, in which case one should just use the PS scheme and not even be bothered with FD, GD, or any other numerical scheme, matters greatly differ when sampling is required, as in the case of VQAs. For FD and GD, it is obvious (see also the start of Sec. V A) that ϵ_{opt} is neither minuscule nor astronomical, but somewhere

in between. However, for a fixed N_T , as d or n increases, ϵ_{opt} tends to larger values because the nonzero- ϵ_{opt} approximation error $\Delta_{\epsilon_{\text{opt}}}^2$ approaches zero while the finite-copy error Δ_{copy}^2 is asymptotically constant. The value of ϵ_{opt} becomes tiny only when N_T is astronomical, in which case the usual narrative applies. Figure 6 visually demonstrates all these remarks for the FD strategy as an example.

If one, for instance, inspects the magnitudes of the respective gradient estimators for FD and PS, one arrives at the large- d limit formulas in (17) and (18) for any fixed N_T . On the other hand, the barren-plateau phenomenon for the PEPQC *ansatz* implies that $\langle (\partial f_Q)^2 \rangle$ drops exponentially in n (see Tab. I). Therefore, given a fixed N_T , the presence of $\epsilon_{\text{opt}} = \epsilon_{\text{opt}}(d, N_T)$ in step-size-dependent strategies such as FD and GD can introduce a faster diminishing average gradient-estimator magnitude that is more compatible with the barren-plateau phenomenon, especially when d is large. More specifically, these strategies do so not by choosing some *ad hoc* ϵ_{opt} that is large, of course, but by minimizing either the TDS MSE, or MSE upper bounds in other non-TDS cases, provided that the circuit *ansatz* is known beforehand—the PEPQC *ansatz* in our case. The PS strategy lacks this additional parameter degree of freedom, and therefore depends only on a sufficiently large N_T to surpass the performances of FD and GD on average.

We forewarn that while optimized numerical estimators can boost estimation accuracies, this does not necessarily mean that these estimators will improve circuit trainability under the influence of the barren-plateau phenomenon. These are clearly two separate problems, the latter of which is not addressed by this paper. The analysis of the MSE throughout the paper reveals the statistical quality in estimating gradients and Hessians. The appearance of a free parameter that characterizes a numerical estimator permits further estimation-accuracy improvement with proper optimization of this parameter. For an extremely large number of qubits n , the landscape of $f_Q(\theta; \mathbf{x})$ from a randomized two-design circuit, for instance, is almost flat, so that the problem of distinguishing function magnitudes in different θ search directions still persists, even with accurate statistical estimations.

An absurd hypothetical scenario would be $n \rightarrow \infty$, where the initialized $\langle f_Q^2 \rangle \cong 0 \cong f_Q$, while the optimized estimation error of any numerical estimator (FD, GD, or SPS) is nearly zero [notice that from (13), analytical PS estimators still give $O(1/N_T)$ estimation errors that are not necessarily small in this scenario]. Such an infinitely-large quantum circuit is not trainable, so low estimation errors surely does not translate to better trainability. To see why, we note that trainability pertains to function-minimization efficiency, and so one should be strict about picking the right descent direction in every function-value update. For a very large n , such that the true gradient has a tiny magnitude, a very small gradient-estimation MSE could still lead to many wrong update directions with even slight statistical fluctuations, so that cost-function minimization can still be very slow on average. Thus, one should, instead, find ways to, as an example, reduce

$$\mathcal{F}_{\theta_0} = \left\langle \frac{\max\{\text{Var}[\widehat{f}_Q(\theta \pm \theta_0; \mathbf{x})]\}}{|\widehat{f}_Q(\theta + \theta_0; \mathbf{x}) - \widehat{f}_Q(\theta - \theta_0; \mathbf{x})|^2} \right\rangle \quad (26)$$

for some chosen displacement θ_0 when speaking of trainability. This quantifies the worst-case average relative spread (variance) of $\widehat{f}_Q(\theta \pm \theta_0; \mathbf{x})$, which if too large, results in the failure of distinguishing between $\widehat{f}_Q(\theta + \theta_0; \mathbf{x})$ and $\widehat{f}_Q(\theta - \theta_0; \mathbf{x})$ in the direction set by θ_0 through sampling. For quantum circuits of large d that exhibit two-design properties, the numerator approaches $O(1/N)$ for N sampling copies, and the denominator is at most $O(1/\text{poly } d)$, so that $\mathcal{F}_{\theta_0} \gtrsim O(\text{poly } d/N)$. For large circuits, N must thus at least be exponentially large in n for trainability [78].

Efforts in ameliorating the effects of barren plateaus would therefore require deeper understanding in both the circuit *ansatz* [78,79], and θ -initialization and optimization strategies [80], for instance. Improving model trainability in the presence of barren plateaus is a pertinent task without a doubt, but is beyond the scope of this paper. With that said, it cannot be overemphasized that both accurate estimation and trainability are equally important in quantum computation especially in the NISQ era where every bit of noise counts. The methodology for optimizing estimators may be extended to other circuit *Ansätze* and circuit-parameter initialization procedures beyond the two-design approximating *Ansätze* and randomized initialization considered in this paper.

VIII. CONCLUSION

Executing variational algorithms on modern NISQ devices typically necessitate the computation of circuit-function gradient and Hessian components through direct variational-circuit-function sampling. A thorough understanding of the inherent estimation errors is vital to ensure the reliability of NISQ computation. In this work, we provide detailed analyses on the estimation errors for various gradient and Hessian computation methods that are relevant not only to gradient and Hessian-assisted optimization approaches, but also nongradient-based routines, which require the estimation of circuit-function differences.

Armed with these fundamental results that apply to very general variational quantum-computation settings, we propose optimally tuned gradient and Hessian numerical estimators that offer significantly reduced average estimation errors on any NISQ device that can only supply a finite number of sampling copies within a given operation time duration. These optimized numerical estimators work especially well in improving the gradient and Hessian computation accuracies during the initial stages of cost-function optimization, where training parameters are first randomly initialized before the optimization procedure such that all polynomially deep training modules possessing a hardware-efficient *ansatz* behave closely as quantum two-designs. The simulation results suggest that such optimally tuned estimators are still extremely advantageous in estimation-error minimization for training modules that are shallow.

Moreover, these numerical estimators are compatible with the barren-plateau phenomenon; that is, given a fixed number of sampling copies, the average estimation errors based on these optimized estimators scale with the corresponding root-mean-squares of circuit-function gradient and Hessian components, both of which drop exponentially with the number of qubits employed. This desirable feature prevents gradient and Hessian computation from turning into random guesses for a fixed number of sampling copies as the circuit size increases.

For the same number of sampling copies, this is in contrast to the analytical unscaled parameter-shift rule, which estimates gradients and Hessians with errors that are asymptotically independent of the circuit-qubit number. We showed that this consequently requires an exponentially increasing number of sampling copies with the qubit number in order for the analytical estimators to overtake the corresponding optimally tuned ones in sampling performance. Hence, while the absence of approximation errors with this analytical rule is a commonly sought-after characteristic, optimally tuned numerical estimators (including the scaled parameter-shift estimators) still present a much more feasible estimation strategy on practical NISQ devices with finite sampling-copy numbers.

An obvious next step to take towards practical applications would be a performance analysis of known sampling strategies in the presence of realistic noisy environments that perforate typical quantum-computing architectures, such as photon loss and depolarization. Knowledge about how (potentially biased) noise models influence sampling computation is pertinent for proposing possibly noise-model-agnostic

optimized strategies to improve estimation qualities. Another interesting area of discussion begins with recognizing the error-mitigating effects in using the unscaled parameter-shift rule, the reason of which is due to a large fixed step size to define gradient and Hessian components, which could overlook slight noise perturbations in the components and render this analytical rule robust against noise. In contrast, conventional wisdom often suggests that finite-difference methods employ much smaller step sizes leading to estimators that are relatively less robust to noise. However, when the knowledge about the circuit *ansatz* is accounted for, the resulting optimized finite-difference strategies correspond to step-size magnitudes that could be comparable with those of the unscaled parameter-shift rule. Hence, the study of possible error-mitigative power for optimized numerical schemes will certainly be a part of the immediate future agenda.

ACKNOWLEDGMENTS

The author thanks S. Shin for fruitful discussions. This work is supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (Grants No. NRF-2020R1A2C1008609, No. NRF-2020K2A9A1A06102946, No. NRF-2019R1A6A1A10073437, and No. NRF-2022M3E4A1076099) via the Institute of Applied Physics at Seoul National University, and the Brain Korea 21 FOUR Project grant funded by the Korean Ministry of Education.

APPENDIX A: BASIC PROPERTIES OF PEPQCS

1. Haar-measure integration for quantum two-designs

An n -qubit serial-circuit model that contains trainable unitary modules W_l that each has a $O(\text{poly}(n, 2))$ circuit depth, so that a randomized W_l for a broad class of circuit *ansätze* (including circuits consisting of single-qubit and CNOT gates) may be approximated as a two-design [72]. PEPQCs form a subclass of such two-design approximating circuits.

In view of this, the following integration result

$$\begin{aligned} & \int (dU)_{\text{Haar}} U_{j_1'k_1'}^* U_{j_2'k_2'}^* U_{j_1k_1} U_{j_2k_2} \\ &= \frac{\delta_{j_1, j_1'} \delta_{j_2, j_2'} \delta_{k_1, k_1'} \delta_{k_2, k_2'} + \delta_{j_1, j_2'} \delta_{j_2, j_1'} \delta_{k_1, k_2'} \delta_{k_2, k_1'}}{d^2 - 1} \\ & \quad - \frac{\delta_{j_1, j_1'} \delta_{j_2, j_2'} \delta_{k_1, k_2'} \delta_{k_2, k_1'} + \delta_{j_1, j_2'} \delta_{j_2, j_1'} \delta_{k_1, k_1'} \delta_{k_2, k_2'}}{d(d^2 - 1)} \end{aligned} \quad (\text{A1})$$

in terms of the computational matrix elements $U_{jk} = \langle j|U|k\rangle$ of a d -dimensional random unitary operator U distributed according to the Haar measure $(dU)_{\text{Haar}}$ and the basic identity

$$\langle U O U^\dagger \rangle_{\text{Haar}} = \int (dU)_{\text{Haar}} U O U^\dagger = \frac{1}{d} \text{tr}\{O\}, \quad (\text{A2})$$

are relevant [73]. By tracking all indices, it is possible to derive another useful integral identity

$$\begin{aligned} \langle U^{\otimes 2} O U^{\dagger \otimes 2} \rangle_{\text{Haar}} &= \int (dU)_{\text{Haar}} U^{\otimes 2} O U^{\dagger \otimes 2} \\ &= \left[\frac{\text{tr}\{O\}}{d^2 - 1} - \frac{\text{tr}\{O\tau\}}{d(d^2 - 1)} \right] 1 \end{aligned}$$

$$+ \left[\frac{\text{tr}\{O\tau\}}{d^2 - 1} - \frac{\text{tr}\{O\}}{d(d^2 - 1)} \right] \tau, \quad (\text{A3})$$

where τ is the swap operator that carries the simple trace property $\text{tr}\{O_1 \otimes O_2 \tau\} = \text{tr}\{O_1 O_2\} = \text{tr}\{U^{\otimes 2} O_1 \otimes O_2 U^{\dagger \otimes 2} \tau\}$ for any two observables O_1 and O_2 , and unitary operator U . If one observable is traceless,

$$\langle U^{\otimes 2} O_1 \otimes O_2 U^{\dagger \otimes 2} \rangle_{\text{Haar}} = \frac{d\tau - 1}{d(d^2 - 1)} \text{tr}\{O_1 O_2\}. \quad (\text{A4})$$

2. Training-parameter translation in $f_{Q,k}$

If we denote $A = \prod_{l'=L}^{l'+1} V_{l'} W_{l'}$ and $B = \prod_{l'=l-1}^l V_{l'} W_{l'}$, then

$$\begin{aligned} \partial_{\mu,l} f_{Q,k} &= \frac{i}{2} \langle \mathbf{0} | B^\dagger W_l^{(2)\dagger} \sigma_{\mu l} W_l^{(1)\dagger} V_l^\dagger A^\dagger O_k A V_l W_l B | \mathbf{0} \rangle \\ &+ \text{c.c.}, \\ (\partial_{\mu,l})^2 f_{Q,k} &= \frac{1}{2} \langle \mathbf{0} | B^\dagger W_l^{(2)\dagger} \sigma_{\mu l} W_l^{(1)\dagger} V_l^\dagger A^\dagger O_k A V_l \\ &\quad \times W_l^{(1)} \sigma_{\mu l} W_l^{(2)} B | \mathbf{0} \rangle - \frac{1}{2} f_{Q,k}, \end{aligned} \quad (\text{A5})$$

where the argument \mathbf{x} is hereby unstated for notational simplicity unless otherwise necessary. From the unique property $\sigma_{\mu l}^2 = 1$ of (multiqubit) Pauli operators employed in PEPQCs, all higher-order derivatives are simply $\partial_{\mu,l} f_{Q,k}$ and $(\partial_{\mu,l})^2 f_{Q,k}$ multiplied by simple phase factors:

$$\begin{aligned} (\partial_{\mu,l})^{2k+1} f_{Q,k} &= (-1)^k \partial_{\mu,l} f_{Q,k}, \\ (\partial_{\mu,l})^{2k} f_{Q,k} &= (-1)^{k+1} (\partial_{\mu,l})^2 f_{Q,k}. \end{aligned} \quad (\text{A6})$$

From (A6), the Taylor series of $f_{Q,k}(\theta_{\mu l} + \theta_0; \mathbf{x})$ can be reduced to a finite linear combination of the zeroth-, first-, and second-order derivatives inasmuch as

$$\begin{aligned} f_{Q,k}(\theta_{\mu l} + \theta_0; \mathbf{x}) &= f_{Q,k}(\theta_{\mu l}; \mathbf{x}) + \sin \theta_0 \partial_{\mu,l} f_{Q,k}(\theta_{\mu l}; \mathbf{x}) \\ &+ (1 - \cos \theta_0) (\partial_{\mu,l})^2 f_{Q,k}(\theta_{\mu l}; \mathbf{x}). \end{aligned} \quad (\text{A7})$$

3. Conditions in gradient and Hessian averaging

Given a trainable module W_l , the gradient operation $\partial_{\mu,l} W_l = W_l^{(1)} \sigma_{\mu l} W_l^{(2)}$ can introduce a single-qubit Pauli operator $\sigma_{\mu l}$ associated to the parameter $\theta_{\mu l}$ that divides W_l into subcircuits of unitary operators $W_l^{(1)}$ and $W_l^{(2)}$ that may or may not be two-designs depending on whether these subcircuits are themselves sufficiently deep. Hence, strictly speaking, the details of the circuit averaging procedure would depend on the location of the gradient operations taken. We will explicitly state the premise in analyzing gradient and Hessian estimation methods:

(1) In Fig. 7, we list down all the physical cases in which there exists at least an approximate two-design module that is free from gradient operations. The reasonable assumption that the entire circuit should be at least deep enough for the above requirement to hold will allow us to subsequently analyze sampling errors.

(2) With such an extent of generality, an exact expression of the MSE for either a gradient or Hessian component is obtained when there exist at least two two-design-approximable training modules sandwiching every training

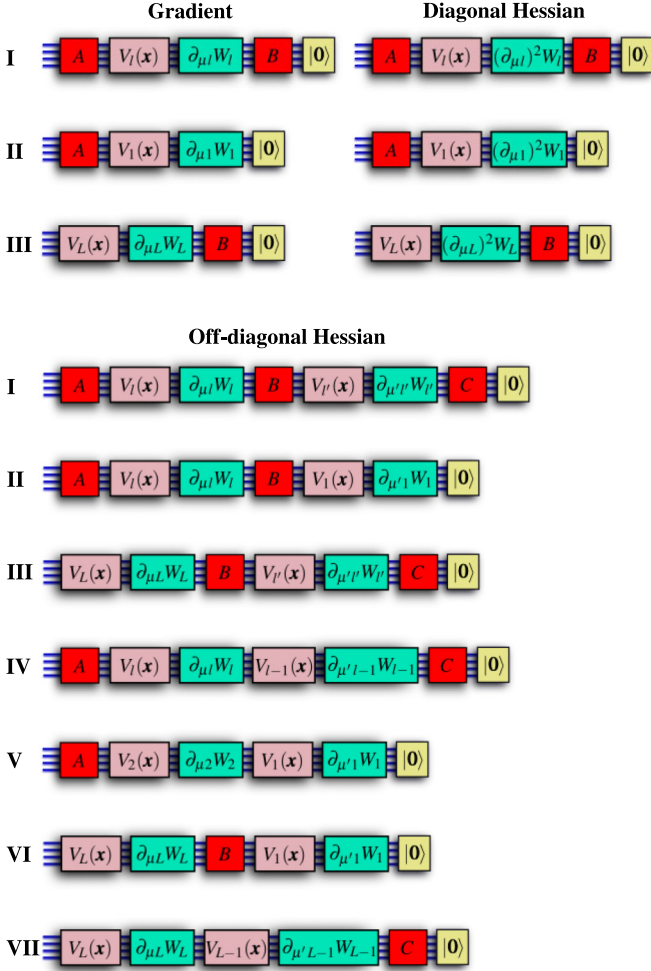


FIG. 7. A list of practically occurring cases in VQAs. All (red) modules A , B , and C contain two-design trainable and arbitrary nontrainable circuits. In all cases, it is reasonable to require that the entire trainable portion of the quantum circuit should be sufficiently deep, such that there exists at least one two-design module that is free from gradient operations at any estimation instance. The TDS condition, thus, corresponds to Case I for every component type.

module on which a derivative operation is performed for that component—the two-design sandwiching (TDS) condition (or Case I in Fig. 7).

(3) Whenever the TDS condition does not apply for any of the derivative operation in a component, an upper bound of the corresponding MSE is derived.

4. General difference gradient and Hessian method

Equations (9) and (10) are, respectively, equivalent to

$$[\partial_{\text{GD}}]_{\mu,l}^{j,\epsilon} f_{Q,k} \equiv \sum_{j=1}^J c_j \text{sinc}(j\epsilon/2) \partial_{\mu,l} f_{Q,k}, \quad (\text{A8})$$

$$[\partial \partial'_{\text{GD}}]_{\mu,l;\mu',l'}^{j,\epsilon} f_{Q,k} \equiv \sum_{j=1}^J c_j [\text{sinc}(j\epsilon/2)]^2 \partial_{\mu,l} \partial_{\mu',l'} f_{Q,k}. \quad (\text{A9})$$

Similar arguments for the TDS condition in Sec. IV A leads to the following quadratic forms for PEPQCs:

$$\mathcal{D}_{\text{GD}}(\cdot) = \mathbf{c}^\top (\mathbf{M} \cdot) \mathbf{c}, \quad \mathbf{M} \cdot > \mathbf{0},$$

$$\mathbf{M}_{\partial f_Q} = \frac{4Jd}{N_T(d+1)\epsilon^2} \mathbf{A}_{\partial f_Q} + \frac{d^2 \mathbf{v}_1 \mathbf{v}_1^\top}{2(d+1)(d^2-1)},$$

$$\mathbf{M}_{\partial \partial f_Q} = \frac{(2J+1)d}{N_T(d+1)\epsilon^4} \mathbf{A}_{\partial \partial f_Q} + \frac{d^2 \mathbf{v}_2 \mathbf{v}_2^\top}{2(d+1)(d^2-1)},$$

(diagonal Hessian components)

$$\mathbf{M}_{\partial \partial' f_Q} = \frac{16Jd}{N_T(d+1)\epsilon^4} \mathbf{A}_{\partial \partial' f_Q} + \frac{d^4 \mathbf{v}_2 \mathbf{v}_2^\top}{4(d+1)(d^2-1)^2},$$

(off-diagonal Hessian components) (A10)

where

$$\mathbf{v}_1 = \mathbf{I} - \begin{pmatrix} \text{sinc}(\epsilon/2) \\ \text{sinc}(\epsilon) \\ \vdots \\ \text{sinc}(J\epsilon/2) \end{pmatrix}, \quad \mathbf{v}_2 = \mathbf{I} - \begin{pmatrix} [\text{sinc}(\epsilon/2)]^2 \\ [\text{sinc}(\epsilon)]^2 \\ \vdots \\ [\text{sinc}(J\epsilon/2)]^2 \end{pmatrix},$$

$$\mathbf{A}_{\partial f_Q} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \frac{1}{2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{J^2} \end{pmatrix}, \quad \mathbf{A}_{\partial \partial' f_Q} = \mathbf{A}_{\partial f_Q}^2,$$

$$\mathbf{A}_{\partial \partial f_Q} = 2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \frac{1}{2^4} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{J^4} \end{pmatrix} + \begin{pmatrix} 1 \\ \frac{1}{2^2} \\ \vdots \\ \frac{1}{J^2} \end{pmatrix} 4 \begin{pmatrix} 1 & \frac{1}{2^2} & \frac{1}{J^2} \end{pmatrix}. \quad (\text{A11})$$

Here, N_T is still the total number of sampling copies distributed equally to all sampled quantum-circuit functions for one GD approximator per circuit-observable basis operator. Just like the optimal FD strategy, the optimal GD strategy in estimating gradient and Hessian components by invoking either (A8) or (A9) would entail the choices of both ϵ and normalized c that minimize the relevant operational \mathcal{D}_{GD} listed in (A10).

APPENDIX B: QUANTUM-CIRCUIT AVERAGES $\langle \cdot \rangle$

1. Inner-product average of translated $f_{Q,k}$ s

As a warmup for the upcoming expedition, we calculate the inner product $\langle f_{Q,k}(\theta_{\mu l} + \theta_0; \mathbf{x}) f_{Q,k}(\theta_{\mu l} + \theta'_0; \mathbf{x}) \rangle$ for arbitrary translations θ_0 and θ'_0 on the same randomized parameter $\theta_{\mu l}$. From Eq. (A7),

$$\begin{aligned} & \langle f_{Q,k}(\theta_{\mu l} + \theta_0; \mathbf{x}) f_{Q,k}(\theta_{\mu l} + \theta'_0; \mathbf{x}) \rangle \\ &= \langle f_{Q,k}^2 \rangle + (2 - \cos \theta_0 - \cos \theta'_0) \langle f_{Q,k}(\partial_{\mu,l})^2 f_{Q,k} \rangle \\ & \quad + \sin \theta_0 \sin \theta'_0 \langle (\partial_{\mu,l} f_{Q,k})^2 \rangle \\ & \quad + (1 - \cos \theta_0)(1 - \cos \theta'_0) \langle [(\partial_{\mu,l})^2 f_{Q,k}]^2 \rangle, \end{aligned} \quad (\text{B1})$$

where we will show that $\langle \partial_{\mu,l} f_{Q,k}(\partial_{\mu,l})^2 f_{Q,k} \rangle$ and $\langle f_{Q,k} \partial_{\mu,l} f_{Q,k} \rangle$ are zero for all three cases shown in Fig. 7.

$$\langle f_{Q,k} \partial_{\mu,l} f_{Q,k} \rangle \text{ for Cases I and II}$$

Upon taking the average over W_L using Eq. (A4), we have

$$\begin{aligned} & \langle f_{Q,k} \partial_{\mu,l} f_{Q,k} \rangle \\ &= \frac{i}{2(d+1)} \langle \mathbf{0} | B^{\dagger \otimes 2} W_l^{(2) \dagger \otimes 2} \mathbf{1} \otimes \sigma_{\mu} W_l^{(2) \otimes 2} B^{\otimes 2} | \mathbf{0} \rangle + \text{c.c.} \end{aligned} \quad (\text{B2})$$

Since $W_l^{(2) \dagger \otimes 2} \mathbf{1} \otimes \sigma_{\mu} W_l^{(2) \otimes 2}$ is yet another Pauli operator, these resulting expectation values are real regardless of whether $l = 1$ or not (that is, whether $B = 1$ correspondingly or not), so that $\langle f_{Q,k} \partial_{\mu,l} f_{Q,k} \rangle_{I,II} = 0$ for both cases.

$\langle f_{Q,k} \partial_{\mu,l} f_{Q,k} \rangle$ for Case III

For this case, we define the operators

$$\sigma' = W_L^{(1) \dagger} V_L^{\dagger} O_k V_L W_L^{(1)}, \quad (\text{B3})$$

$$Q_{III} = W_L^{(2) \dagger \otimes 2} \sigma' \otimes \sigma_{\mu L} \sigma' W_L^{(2) \otimes 2}. \quad (\text{B4})$$

From the realization that σ' in Eq. (B3) is a Pauli operator, the following two trace properties

$$\text{tr}\{V_{L-1}^{\dagger \otimes 2} Q_{III} V_{L-1}^{\otimes 2}\} = \text{tr}\{\sigma' \otimes \sigma_{\mu L} \sigma'\} = 0,$$

$$\text{tr}\{V_{L-1}^{\dagger \otimes 2} Q_{III} V_{L-1}^{\otimes 2} \tau\} = \text{tr}\{\sigma' \sigma_{\mu L} \sigma'\} = \text{tr}\{\sigma_{\mu L}\} = 0 \quad (\text{B5})$$

become apparent, giving us $\langle f_{Q,k} \partial_{\mu,l} f_{Q,k} \rangle_{III} = 0$.

By repeating the above calculations, we also find that $\langle (\partial_{\mu,l} f_{Q,k}) [(\partial_{\mu,l})^2 f_{Q,k}] \rangle_{I,II,III} = 0$. These results are consistent with the property that average inner products of odd combined derivative order is always zero, another inherent trait from a Pauli-type observable O_k .

We are now left with $\langle f_{Q,k} (\partial_{\mu,l})^2 f_{Q,k} \rangle$, $\langle |\partial_{\mu,l} f_{Q,k}|^2 \rangle$ and $\langle |(\partial_{\mu,l})^2 f_{Q,k}|^2 \rangle$. First, the average of

$$\begin{aligned} & f_{Q,k} (\partial_{\mu,l})^2 f_{Q,k} \\ &= \frac{1}{2} \langle \mathbf{0} | B^{\dagger \otimes 2} W_l^{(2) \dagger \otimes 2} \mathbf{1} \otimes \sigma_{\mu l} W_l^{(1) \dagger \otimes 2} V_l^{\dagger \otimes 2} \\ & \quad \times A^{\dagger \otimes 2} O_k^{\otimes 2} A^{\otimes 2} V_l^{\otimes 2} W_l^{(1) \otimes 2} \mathbf{1} \otimes \sigma_{\mu l} W_l^{(2) \otimes 2} B^{\otimes 2} | \mathbf{0} \rangle - \frac{f_{Q,k}^2}{2} \end{aligned} \quad (\text{B6})$$

involves $\langle f_{Q,k}^2 \rangle = 1/(d+1)$ according to Eq. (A4).

$\langle f_{Q,k} (\partial_{\mu,l})^2 f_{Q,k} \rangle$ for Case I

With (A4),

$$\begin{aligned} & \langle f_{Q,k} (\partial_{\mu,l})^2 f_{Q,k} \rangle_I \\ &= \frac{d}{2(d^2-1)} \langle \langle \mathbf{0} | B^{\dagger} W_l^{(2) \dagger} \sigma_{\mu l} W_l^{(2)} B | \mathbf{0} \rangle^2 \rangle - \frac{d}{2(d^2-1)}, \end{aligned} \quad (\text{B7})$$

$$\text{or } \langle f_{Q,k} (\partial_{\mu,l})^2 f_{Q,k} \rangle_I = -\frac{d^2}{2(d+1)(d^2-1)}. \quad (\text{B8})$$

$\langle f_{Q,k} (\partial_{\mu,l})^2 f_{Q,k} \rangle$ for Case II

We note that

$$\gamma_{II} = \langle \langle \mathbf{0} | W_l^{(2) \dagger} \sigma_{\mu l} W_l^{(2)} | \mathbf{0} \rangle^2 \rangle \leq 1, \quad (\text{B9})$$

so that

$$\langle f_{Q,k} (\partial_{\mu,l})^2 f_{Q,k} \rangle_{II} = -\frac{d(1-\gamma_{II})}{2(d^2-1)}. \quad (\text{B10})$$

$\langle f_{Q,k} (\partial_{\mu,l})^2 f_{Q,k} \rangle$ for Case III

For this case, consider the operator

$$Q'_{III} = W_L^{(2) \dagger \otimes 2} \mathbf{1} \otimes \sigma_{\mu L} \sigma'^{\otimes 2} \mathbf{1} \otimes \sigma_{\mu L} W_L^{(2) \otimes 2}. \quad (\text{B11})$$

Its trace properties include $\text{tr}\{Q'_{III}\} = 0$,

$$\text{tr}\{V_{L-1}^{\dagger \otimes 2} Q'_{III} V_{L-1}^{\otimes 2} \tau\} = \text{tr}\{\sigma' \sigma_{\mu L} \sigma' \sigma_{\mu L}\} = \gamma_{III}(\mathbf{x}). \quad (\text{B12})$$

With these,

$$\langle f_{Q,k} (\partial_{\mu,l})^2 f_{Q,k} \rangle_{III} = -\frac{d-\gamma_{III}(\mathbf{x})}{2d(d+1)}. \quad (\text{B13})$$

We catalog the calculations of $\langle (\partial_{\mu,l} f_{Q,k})^2 \rangle$ and $\langle |(\partial_{\mu,l})^2 f_{Q,k}|^2 \rangle$ in the following sections, and simply list the final answers:

$$\begin{aligned} & \langle f_{Q,k} (\theta_{\mu l} + \theta_0; \mathbf{x}) f_{Q,k} (\theta_{\mu l} + \theta'_0; \mathbf{x}) \rangle \\ &= \begin{cases} \frac{d^2 \left[\cos\left(\frac{\theta_0 - \theta'_0}{2}\right) \right]^2 - 1}{(d+1)(d^2-1)} & \text{for Case I,} \\ \frac{d(1-\gamma_{II}) \left[\cos\left(\frac{\theta_0 - \theta'_0}{2}\right) \right]^2 + d\gamma_{II} - 1}{d^2-1} & \text{for Case II,} \\ \frac{[d-\gamma_{III}(\mathbf{x})] \left[\cos\left(\frac{\theta_0 - \theta'_0}{2}\right) \right]^2 + \gamma_{III}(\mathbf{x})}{d(d+1)} & \text{for Case III.} \end{cases} \end{aligned} \quad (\text{B14})$$

By taking $\theta_0 = \theta'_0$, we recover the special case $\langle f_{Q,k} (\theta_{\mu l} + \theta_0; \mathbf{x})^2 \rangle = 1/(d+1)$.

2. Averages of gradient components $\partial f_{Q,k}$

We derive results concerning $\langle \partial_{\mu,l} f_{Q,k} \rangle$ and $\langle (\partial_{\mu,l} f_{Q,k})^2 \rangle$ for all the three cases in Fig. 7, beginning with the former.

$\langle \partial_{\mu,l} f_{Q,k} \rangle$ for Cases I and II

Again, as $\langle W_L^{\dagger} V_L^{\dagger} O_k V_L W_L \rangle_{\text{Haar}} = 0$ for a Pauli O_k ,

$$\langle \partial_{\mu,l} f_{Q,k} \rangle_{I,II} = 0. \quad (\text{B15})$$

$\langle \partial_{\mu,l} f_{Q,k} \rangle$ for Case III

We inspect the operator

$$Q''_{III} = W_L^{(2) \dagger} \sigma_{\mu L} W_L^{(1) \dagger} V_L^{\dagger} O_k V_L W_L^{(1)} W_L^{(2)}. \quad (\text{B16})$$

Note that $\text{tr}\{\sigma_{\mu L} \sigma'\}$ is clearly real, so that

$$\langle \partial_{\mu,L} f_{Q,k} \rangle_{III} = \frac{i}{2d} \langle \text{tr}\{\sigma_{\mu l} \sigma'\} \rangle - \frac{i}{2d} \langle \text{tr}\{\sigma_{\mu l} \sigma'\} \rangle = 0. \quad (\text{B17})$$

Now, for the latter:

$\langle |(\partial_{\mu,l} f_{Q,k})|^2 \rangle$ for Case I

Averaging over W_L yields

$$\begin{aligned} & \langle |\partial_{\mu,l} f_{Q,k}|^2 \rangle \\ &= -\frac{d}{4(d^2-1)} \langle \langle \mathbf{0} | B^{\dagger \otimes 2} W_l^{(2) \dagger \otimes 2} \sigma_{\mu l}^{\otimes 2} W_l^{(2) \otimes 2} B^{\otimes 2} | \mathbf{0} \rangle \rangle \\ &+ \frac{d}{4(d^2-1)} + \text{c.c.}, \end{aligned} \quad (\text{B18})$$

or

$$\langle |\partial_{\mu,l} f_{Q,k}|^2 \rangle_{\text{I}} = \frac{d^2}{2(d+1)(d^2-1)}. \quad (\text{B19})$$

$\langle |\partial_{\mu,l} f_{Q,k}|^2 \rangle$ for Case II

From (B9), we simply obtain

$$\langle |\partial_{\mu,l} f_{Q,k}|^2 \rangle_{\text{II}} = \frac{d(1-\gamma_{\text{II}})}{2(d^2-1)} \leq \frac{d}{2(d^2-1)}. \quad (\text{B20})$$

$\langle |\partial_{\mu,l} f_{Q,k}|^2 \rangle$ for Case III

For this case, properties of the operators

$$\begin{aligned} Q_{\text{III(a)}}''' &= W_L^{(2) \dagger \otimes 2} \sigma_{\mu L}^{\otimes 2} \sigma'^{\otimes 2} W_L^{(2) \otimes 2}, \\ Q_{\text{III(b)}}''' &= W_L^{(2) \dagger \otimes 2} \mathbf{1} \otimes \sigma_{\mu L} \sigma'^{\otimes 2} \sigma_{\mu L} \otimes 1 W_L^{(2) \otimes 2} \end{aligned} \quad (\text{B21})$$

are necessary, where we recall σ' from Eq. (B3). To start off,

$$\text{tr}\{Q_{\text{III(a)}}'''\} = \text{tr}\{\sigma_{\mu L} \sigma'\}^2 = \text{tr}\{Q_{\text{III(b)}}'''\}. \quad (\text{B22})$$

For the trace properties with the swap operator, they are

$$\begin{aligned} \text{tr}\{V_{L-1}^{\dagger \otimes 2} Q_{\text{III(a)}}''' V_{L-1}^{\otimes 2} \tau\} &= \text{tr}\{\sigma_{\mu L} \sigma' \sigma_{\mu L} \sigma'\} = \gamma_{\text{III}}(x), \\ \text{tr}\{V_{L-1}^{\dagger \otimes 2} Q_{\text{III(b)}}''' V_{L-1}^{\otimes 2} \tau\} &= \text{tr}\{\sigma' \sigma_{\mu L} \sigma_{\mu L} \sigma'\} = d. \end{aligned} \quad (\text{B23})$$

These are critical in evaluating the average over W_{L-1} by invoking Eq. (A4):

$$\begin{aligned} \gamma_{\text{III}}(\mathbf{x}) &\equiv \langle \text{tr}\{(\sigma_{\mu L} \sigma')^2\} \rangle, \\ \gamma'_{\text{III}}(\mathbf{x}) &\equiv \langle \text{tr}\{\sigma_{\mu L} \sigma'\}^2 \rangle, \\ \langle |\partial_{\mu,L} f_{Q,k}|^2 \rangle_{\text{III}} &= -\frac{\gamma'_{\text{III}}(\mathbf{x}) + \gamma_{\text{III}}(\mathbf{x})}{2d(d+1)} + \frac{\gamma'_{\text{III}}(\mathbf{x}) + d}{2d(d+1)} \\ &= \frac{d - \gamma_{\text{III}}(\mathbf{x})}{2d(d+1)} \leq \frac{1}{d+1}. \end{aligned} \quad (\text{B24})$$

The final inequality is of the Cauchy-Schwarz type

$$\text{tr}\{[\sigma_{\mu L} \sigma']^2\} \leq \text{tr}\{\sigma_{\mu L}^2\} \text{tr}\{[\sigma' \sigma_{\mu L} \sigma']^2\} = d^2, \quad (\text{B25})$$

or $-d \leq \gamma_{\text{III}}(\mathbf{x}) \leq d$.

3. Averages of diagonal Hessian components $\partial \partial f_{Q,k}$

From (A5), just like $\partial f_{Q,k} = 0$, it can be easily confirmed that $\langle \partial \partial f_{Q,k} \rangle_{\text{I,II,III}} = 0$. For the squared averages, since

$$[(\partial_{\mu,l})^2 f_{Q,k}]^2 = \frac{1}{4} [f_{Q,k}^2 + D_{\text{I}} - 2D_{\text{II}}],$$

$$\begin{aligned} D_{\text{I}} &= \langle \mathbf{0} | B^{\dagger \otimes 2} W_l^{(2) \dagger \otimes 2} \sigma_{\mu l}^{\otimes 2} W_l^{(1) \dagger \otimes 2} V_l^{\dagger \otimes 2} A^{\dagger \otimes 2} O^{\otimes 2} \\ &\times A^{\otimes 2} V_l^{\otimes 2} W_l^{(1) \otimes 2} \sigma_{\mu l}^{\otimes 2} W_l^{(2) \otimes 2} B^{\otimes 2} | \mathbf{0} \rangle, \end{aligned}$$

$$\begin{aligned} D_{\text{II}} &= \langle \mathbf{0} | B^{\dagger \otimes 2} W_l^{(2) \dagger \otimes 2} \mathbf{1} \otimes \sigma_{\mu l} W_l^{(1) \dagger \otimes 2} V_l^{\dagger \otimes 2} A^{\dagger \otimes 2} \\ &\times O_k^{\otimes 2} A^{\otimes 2} V_l^{\otimes 2} W_l^{(1) \otimes 2} \mathbf{1} \otimes \sigma_{\mu l} W_l^{(2) \otimes 2} B^{\otimes 2} | \mathbf{0} \rangle, \end{aligned} \quad (\text{B26})$$

we need the averages of D_{I} and D_{II} in all three cases.

$\langle |\partial_{\mu,l} f_{Q,k}|^2 \rangle$ for Case I

That $\langle D_{\text{I}} \rangle = 1/(d+1)$ is immediate. On the other hand,

$$\begin{aligned} \langle D_{\text{II}} \rangle &= \frac{d}{d^2-1} \langle \langle \mathbf{0} | B^{\dagger \otimes 2} W_l^{(2) \dagger \otimes 2} \sigma_{\mu l}^{\otimes 2} W_l^{(2) \otimes 2} B^{\otimes 2} | \mathbf{0} \rangle \rangle \\ &- \frac{1}{d^2-1} \\ &= -\frac{1}{(d+1)(d^2-1)}. \end{aligned} \quad (\text{B27})$$

Altogether,

$$\langle |\partial_{\mu,l} f_{Q,k}|^2 \rangle_{\text{I}} = \frac{d^2}{2(d+1)(d^2-1)}. \quad (\text{B28})$$

$\langle |\partial_{\mu,l} f_{Q,k}|^2 \rangle$ for Case II

If $l = 1$, then $\langle D_{\text{I}} \rangle = 1/(d+1)$ still, but

$$\langle D_{\text{II}} \rangle = \frac{d\gamma_{\text{II}} - 1}{d^2 - 1}, \quad (\text{B29})$$

so that

$$\langle |\partial_{\mu,l} f_{Q,k}|^2 \rangle_{\text{II}} = \frac{d - d\gamma_{\text{II}}}{2(d^2-1)} \leq \frac{d}{2(d^2-1)}, \quad (\text{B30})$$

where γ_{II} is as defined in (B9).

$\langle |\partial_{\mu,l} f_{Q,k}|^2 \rangle$ for Case III

From (B11) and the operator definition

$$Q_{\text{III}}'''' = W_L^{(2) \dagger \otimes 2} \sigma_{\mu L}^{\otimes 2} \sigma'^{\otimes 2} \sigma_{\mu L}^{\otimes 2} W_L^{(2) \otimes 2}, \quad (\text{B31})$$

whose trace properties include $\text{tr}\{Q_{\text{III}}''''\} = 0$ and

$$\text{tr}\{V_{L-1}^{\dagger \otimes 2} Q_{\text{III}}'''' V_{L-1}^{\otimes 2} \tau\} = \text{tr}\{\sigma_{\mu L} \sigma' \sigma' \sigma_{\mu L}\} = d. \quad (\text{B32})$$

With these,

$$\begin{aligned} \langle D_{\text{I}} \rangle &= \langle \langle \mathbf{0} | B^{\dagger \otimes 2} Q_{\text{III}}'''' B^{\otimes 2} | \mathbf{0} \rangle \rangle = \frac{1}{d+1}, \\ \langle D_{\text{II}} \rangle &= \langle \langle \mathbf{0} | B^{\dagger \otimes 2} Q_{\text{III}}'''' B^{\otimes 2} | \mathbf{0} \rangle \rangle = \frac{\gamma_{\text{III}}(x)}{d(d+1)}, \end{aligned} \quad (\text{B33})$$

which finally brings us to

$$\langle |\partial_{\mu,L} f_{Q,k}|^2 \rangle_{\text{III}} = \frac{d - \gamma_{\text{III}}(\mathbf{x})}{2d(d+1)} \leq \frac{1}{d+1}. \quad (\text{B34})$$

4. Averages of off-diagonal Hessian components $\partial_{\mu,l} \partial_{\mu',l'} f_{Q,k}$

The averages $|\partial_{\mu,l} \partial_{\mu',l'} f_{Q,k}|^2$ are to be computed for a total of seven cases depicted in Fig. 7. Each component consists of a summation of four pieces:

$$\partial_{\mu,l} \partial_{\mu',l'} f_{Q,k} = [(1) + (1)_{\text{c.c.}} + (2) + (2)_{\text{c.c.}}],$$

$$(1) = -\frac{1}{4} \langle \mathbf{0} | C^{\dagger} W_l^{(2) \dagger} \sigma_{\mu l'} W_l^{(1) \dagger} V_l^{\dagger} B^{\dagger} W_l^{(2) \dagger} \sigma_{\mu l} W_l^{(1)} \rangle$$

$$\begin{aligned}
 & \times V_l^\dagger A^\dagger O_k A V_l W_l B V_l W_l C |\mathbf{0}\rangle, \\
 (2) &= \frac{1}{4} \langle \mathbf{0} | C^\dagger W_l^\dagger V_l^\dagger B^\dagger W_l^{(2)\dagger} \sigma_{\mu l} W_l^{(1)\dagger} V_l^\dagger A^\dagger O_k A \\
 & \times V_l W_l B V_l^\dagger W_l^{(1)\dagger} \sigma_{\mu' l'} W_l^{(2)\dagger} C |\mathbf{0}\rangle. \quad (B35)
 \end{aligned}$$

Its squared average thus involves 16 terms. We will list all of their combinations conveniently for every case.

$\langle |\partial_{\mu, l} \partial_{\mu', l'} f_{Q, k}|^2 \rangle$ for Cases I and IV

Let $\tilde{\sigma}_{\mu l} = W_l^{(1)\dagger} V_l^\dagger B^\dagger W_l^{(2)\dagger} \sigma_{\mu l} W_l^{(2)} B V_l W_l^{(1)}$. Then upon denoting $\tilde{\gamma} = \text{tr}\{(\sigma_{\mu' l'} \tilde{\sigma}_{\mu l})^2\}$ and $\tilde{\gamma}' = \text{tr}\{\sigma_{\mu' l'} \tilde{\sigma}_{\mu l}\}^2$, we get

$$\begin{aligned}
 \langle (1)^2 + (1)_{c.c.}^2 \rangle &= \langle (2)^2 + (2)_{c.c.}^2 \rangle = \frac{1}{8d(d+1)^2} \langle \tilde{\gamma} + \tilde{\gamma}' \rangle, \\
 2\langle (1) \times (1)_{c.c.} \rangle &= \frac{d^2 + d - 1}{8(d+1)(d^2 - 1)} - \frac{\langle \tilde{\gamma}' \rangle}{8d(d+1)(d^2 - 1)}, \\
 2\langle (1) \times (2) \rangle &= \frac{\langle \tilde{\gamma}' - d\tilde{\gamma} \rangle}{8d(d+1)(d^2 - 1)} + \frac{1}{8(d+1)(d^2 - 1)}, \\
 2\langle (1) \times (2)_{c.c.} \rangle &= \frac{\langle \tilde{\gamma} + \tilde{\gamma}' \rangle}{8d(d+1)(d^2 - 1)} - \frac{d}{8(d+1)(d^2 - 1)}, \\
 2\langle (2) \times (2)_{c.c.} \rangle &= 2\langle (1) \times (1)_{c.c.} \rangle. \quad (B36)
 \end{aligned}$$

These amount to

$$\begin{aligned}
 \langle |\partial_{\mu, l} \partial_{\mu', l'} f_{Q, k}|^2 \rangle_{IV} &= \frac{d^2 + \langle \tilde{\gamma}' \rangle}{4(d+1)(d^2 - 1)} \\
 &\leq \frac{d^2}{2(d+1)(d^2 - 1)}. \quad (B37)
 \end{aligned}$$

In Case I, $\langle \tilde{\gamma}' \rangle = d^2/(d^2 - 1)$, so that

$$\langle |\partial_{\mu, l} \partial_{\mu', l'} f_{Q, k}|^2 \rangle_I = \frac{d^4}{4(d+1)(d^2 - 1)^2}. \quad (B38)$$

$\langle |\partial_{\mu, l} \partial_{\mu', l'} f_{Q, k}|^2 \rangle$ for Cases II and V

Let us define the shorthand notations $\alpha = \langle \mathbf{0} | W_l^{(2)\dagger} \sigma_{\mu' l'} \tilde{\sigma}_{\mu l} W_l^{(2)} | \mathbf{0} \rangle$, $\beta = \langle \mathbf{0} | W_l^{(2)\dagger \otimes 2} \sigma_{\mu' l'} \tilde{\sigma}_{\mu l} \sigma_{\mu' l'} \otimes \tilde{\sigma}_{\mu l} W_l^{(2) \otimes 2} | \mathbf{0} \rangle$ and $\gamma = \langle \mathbf{0} | W_l^{(2)\dagger} \sigma_{\mu' l'} W_l^{(2)} | \mathbf{0} \rangle$. Then,

$$\begin{aligned}
 \langle (1)^2 + (1)_{c.c.}^2 \rangle &= \langle (2)^2 + (2)_{c.c.}^2 \rangle = \frac{\text{Re}\{\langle \alpha^2 \rangle\}}{8(d+1)}, \\
 2\langle (1) \times (1)_{c.c.} \rangle &= \frac{d}{8(d^2 - 1)} - \frac{\langle |\alpha|^2 \rangle}{8(d^2 - 1)}, \\
 2\langle (1) \times (2) \rangle &= \frac{\langle |\alpha|^2 \rangle}{8(d^2 - 1)} - \frac{d\langle \beta \rangle}{8(d^2 - 1)}, \\
 2\langle (1) \times (2)_{c.c.} \rangle &= \frac{\langle |\alpha|^2 \rangle}{8(d^2 - 1)} - \frac{d\langle \gamma^2 \rangle}{8(d^2 - 1)}, \\
 2\langle (2) \times (2)_{c.c.} \rangle &= 2\langle (1) \times (1)_{c.c.} \rangle, \quad (B39)
 \end{aligned}$$

such that

$$\langle |\partial_{\mu, l} \partial_{\mu', l'} f_{Q, k}|^2 \rangle_V = \frac{d(1 + \text{Re}\{\alpha^2\} - \beta - \gamma^2)}{4(d^2 - 1)} \leq \frac{3d}{4(d^2 - 1)}. \quad (B40)$$

The inequality is deducible from the basic inequalities $0 \leq \gamma^2 \leq 1$, $-1 \leq \beta \leq 1$, and $\text{Re}\{\alpha^2\} \leq |\alpha|^2 \leq 1$.

When Case II holds, one arrives at $\langle \alpha^2 \rangle = \langle \gamma^2 \rangle / (d + 1)$ and $\langle \beta \rangle = (d\langle \gamma^2 \rangle - 1) / (d^2 - 1)$, yielding

$$\langle |\partial_{\mu, l} \partial_{\mu', l'} f_{Q, k}|^2 \rangle_{II} = \frac{d^3(1 - \langle \gamma^2 \rangle)}{4(d^2 - 1)^2} \leq \frac{d^3}{4(d^2 - 1)^2}. \quad (B41)$$

$\langle |\partial_{\mu, l} \partial_{\mu', l'} f_{Q, k}|^2 \rangle$ for Cases III and VII

For these two cases, let us focus on the Pauli operators

$$\begin{aligned}
 \sigma_0 &\equiv W_l^{(1)\dagger} V_l^\dagger A^\dagger O_k A V_l W_l^{(1)}, \\
 \sigma_1 &\equiv V_l^\dagger B^\dagger W_l^{(2)\dagger} \sigma_0 W_l^{(2)} B V_l, \\
 \sigma_2 &\equiv V_l^\dagger B^\dagger W_l^{(2)\dagger} \sigma_{\mu l} W_l^{(2)} B V_l. \quad (B42)
 \end{aligned}$$

In terms of these Pauli operators and the rotated versions $\tilde{\sigma}_j = W_l^{(1)\dagger} \sigma_j W_l^{(1)}$, we define a new set of parameters: $\alpha' = \text{tr}\{\sigma_{\mu' l'} \tilde{\sigma}_2 \tilde{\sigma}_1\}$, $\beta' = \text{tr}\{(\sigma_{\mu' l'} \tilde{\sigma}_2 \tilde{\sigma}_1)^2\}$. These lead to

$$\begin{aligned}
 \langle (1)^2 + (1)_{c.c.}^2 \rangle &= \langle (2)^2 + (2)_{c.c.}^2 \rangle = \frac{\text{Re}\{\langle \alpha'^2 + \beta' \rangle\}}{8d(d+1)}, \\
 2\langle (1) \times (1)_{c.c.} \rangle &= \frac{\langle |\alpha'|^2 \rangle + d}{8d(d+1)}, \\
 2\langle (1) \times (2) \rangle &= -\frac{\langle \alpha'^2 + \text{tr}\{(\sigma_1 \sigma_2)^2 \} \rangle}{8d(d+1)}, \\
 2\langle (1) \times (2)_{c.c.} \rangle &= -\frac{\langle |\alpha'|^2 + \beta' \rangle}{8d(d+1)}, \\
 2\langle (2) \times (2)_{c.c.} \rangle &= 2\langle (1) \times (1)_{c.c.} \rangle, \quad (B43)
 \end{aligned}$$

giving us

$$\langle |\partial_{\mu, l} \partial_{\mu', l'} f_{Q, k}|^2 \rangle_{VII} = \frac{d - \langle \text{tr}\{(\sigma_1 \sigma_2)^2 \} \rangle}{4d(d+1)} \leq \frac{1}{2(d+1)}, \quad (B44)$$

where we employed the Cauchy-Schwarz inequality $-d \leq \text{tr}\{(\sigma_1 \sigma_2)^2\} \leq d$.

As $\text{tr}\{(\sigma_1 \sigma_2)^2\}$ is independent of B , we see that Case III offers no further tightening to the above inequality:

$$\langle |\partial_{\mu, l} \partial_{\mu', l'} f_{Q, k}|^2 \rangle_{III} \leq \frac{1}{2(d+1)}. \quad (B45)$$

$\langle |\partial_{\mu, l} \partial_{\mu', l'} f_{Q, k}|^2 \rangle$ for Case VI

In this final case, do expressions are

$$\begin{aligned}
 \langle (1)^2 + (1)_{c.c.}^2 \rangle &= \langle (2)^2 + (2)_{c.c.}^2 \rangle = \frac{1}{8} \text{Re}\{\langle x^2 \rangle\}, \\
 2\langle (1) \times (1)_{c.c.} \rangle &= \frac{1}{8} \langle |x|^2 \rangle, \\
 2\langle (1) \times (2) \rangle &= -\frac{1}{8} \langle xy \rangle, \\
 2\langle (1) \times (2)_{c.c.} \rangle &= -\frac{1}{8} \langle xy^* \rangle, \\
 2\langle (2) \times (2)_{c.c.} \rangle &= \frac{1}{8} \langle |y|^2 \rangle, \quad (B46)
 \end{aligned}$$

with

$$w_{\mu' l'} = \langle \mathbf{0} | W_l^{(2)\dagger} \sigma_{\mu' l'} W_l^{(2)} | \mathbf{0} \rangle^2 \leq 1,$$

TABLE I. Summary table listing all averages (and their upper bounds) of squared-magnitudes for all types of gradient and Hessian components. We remind the reader that the notation $\partial_{\mu,l}\partial_{\mu',l'}$ refers to the off-diagonal components, where either $\mu \neq \mu'$, or $l \neq l'$, or both.

Case	$\langle \partial_{\mu,l} f_{Q,k} ^2 \rangle$	$\langle (\partial_{\mu,l})^2 f_{Q,k} ^2 \rangle$	$\langle \partial_{\mu,l}\partial_{\mu',l'} f_{Q,k} ^2 \rangle$
I	$\frac{d^2}{2(d+1)(d^2-1)}$	$\frac{d^2}{2(d+1)(d^2-1)}$	$\frac{d^4}{4(d+1)(d^2-1)^2}$
II	$\leq \frac{d}{2(d^2-1)}$	$\leq \frac{d}{2(d^2-1)}$	$\leq \frac{d^3}{4(d^2-1)^2}$
III	$\leq \frac{1}{d+1}$	$\leq \frac{1}{d+1}$	$\leq \frac{1}{2(d+1)}$
Case	$\langle \partial_{\mu,l}\partial_{\mu',l'} f_{Q,k} ^2 \rangle$		
IV	$\leq \frac{d^2}{2(d+1)(d^2-1)}$		
V	$\leq \frac{3d}{4(d^2-1)}$		
VI	$\leq \frac{d}{2(d^2-1)}$		
VII	$\leq \frac{1}{2(d+1)}$		

$$\begin{aligned}
\langle x^2 \rangle &= w_{\mu'l'} \frac{\langle \text{tr}\{\sigma_0\sigma_{\mu l}\}^2 + \text{tr}\{\sigma_0\sigma_{\mu l}\sigma_0\sigma_{\mu l}\} \rangle}{d(d+1)}, \\
\langle |x|^2 \rangle &= \langle |y|^2 \rangle = \langle \text{tr}\{\sigma_0\sigma_{\mu l}\}^2 \rangle \frac{w_{\mu'l'} - 1/d}{d(d^2-1)} + \frac{d - w_{\mu'l'}}{d^2-1}, \\
\langle \text{Re}\{xy\} \rangle &= \frac{\text{tr}\{\sigma_0\sigma_{\mu l}\}^2 (dw_{\mu'l'} - 1)}{d(d^2-1)} \\
&\quad + \frac{\text{tr}\{(\sigma_0\sigma_{\mu l})^2\} (d - w_{\mu'l'})}{d(d^2-1)}, \\
\langle \text{Re}\{xy^*\} \rangle &= \frac{w_{\mu'l'}}{d+1} \left(1 + \frac{\text{tr}\{\sigma_0\sigma_{\mu l}\}^2}{d} \right). \tag{B47}
\end{aligned}$$

Combining these expressions gives us

$$\begin{aligned}
\langle |\partial_{\mu,l}\partial_{\mu',l'} f_{Q,k}|^2 \rangle_{\text{VI}} &= \frac{(1 - w_{\mu'l'})(d - \text{tr}\{\sigma_0\sigma_{\mu l}\sigma_0\sigma_{\mu l}\})}{4(d^2-1)} \\
&\leq \frac{d}{2(d^2-1)}. \tag{B48}
\end{aligned}$$

5. Summary table of all gradient and Hessian averages

Without referring to the details in Appendices B 2 through B 4, Table I summarizes the important averages of squared magnitudes for all gradient- and Hessian-component types. Evidently, all averages are at most $O(1/d)$, which is crucial for revealing the beneficial statistical properties of optimized FD and GD estimators.

APPENDIX C: OPTIMIZED SAMPLING ERRORS FOR ALL CASES

1. Multinomial sampling distribution

If each d -dimensional Pauli basis operator O_k of the variational quantum circuit is measured independently, then each measurement of a fixed number of sampling copies N is the eigenbasis $\{|o_{kl}\rangle\}_{l=0}^{d-1}$ of O_k , where relative frequencies $v_{kl}(\boldsymbol{\theta}; \mathbf{x}) = n_{kl}(\boldsymbol{\theta}; \mathbf{x})/N \rightarrow p_{kl}(\boldsymbol{\theta}; \mathbf{x})$ are recorded. Explicitly, these relative frequencies satisfy the following basic statistical

identities:

$$\begin{aligned}
\mathbb{E}[v_{kl}(\boldsymbol{\theta}; \mathbf{x})] &= p_{kl}(\boldsymbol{\theta}; \mathbf{x}), \\
\mathbb{E}[v_{kl}(\boldsymbol{\theta}; \mathbf{x})v_{k'l'}(\boldsymbol{\theta}'; \mathbf{x}')] &= (1 - \delta_{\theta,\theta'}\delta_{x,x'})p_{kl}(\boldsymbol{\theta}; \mathbf{x})p_{k'l'}(\boldsymbol{\theta}'; \mathbf{x}') \\
&\quad + \frac{\delta_{\theta,\theta'}\delta_{x,x'}}{N} [\delta_{l,l'}p_{kl}(\boldsymbol{\theta}; \mathbf{x}) \\
&\quad + (N-1)p_{kl}(\boldsymbol{\theta}; \mathbf{x})p_{k'l'}(\boldsymbol{\theta}; \mathbf{x})], \\
\mathbb{E}[v_{kl}(\boldsymbol{\theta}; \mathbf{x})v_{k' \neq k, l'}(\boldsymbol{\theta}'; \mathbf{x}')] &= p_{kl}(\boldsymbol{\theta}; \mathbf{x})p_{k'l'}(\boldsymbol{\theta}'; \mathbf{x}). \tag{C1}
\end{aligned}$$

The above identities hold whenever circuits of different training parameters are sampled independently.

2. Function estimation

We start with the estimation of f_Q using the unbiased estimator in (4). Together with the multinomial identities in (C1), this leads to

$$\mathcal{D}(f_Q) = \frac{1}{N_T} \left(1 - \overline{f_{Q,k}^2} \right), \tag{C2}$$

where $N_T = N$ is the total number of copies per basis observable needed to complete one function estimation evaluated at the circuit parameters θ . For two-design approximable PEPQCs, by making use of Eq. (A4), we quickly find that

$$\langle f_{Q,k}(\theta_{\mu l}; \mathbf{x})^2 \rangle = \langle f_{Q,k}(\theta_{\mu l} + \theta_0; \mathbf{x})^2 \rangle = \frac{1}{d+1} \tag{C3}$$

for any θ_0 (see Appendix B 1), so that

$$\mathcal{D}(f_Q) = \frac{d}{N_T(d+1)}. \tag{C4}$$

3. Optimally tuned FD estimators for general cases

In their most general forms, the FD MSEs, which are linear combinations of finite-copy and nonzero- ϵ approximation squared errors, read

$$\begin{aligned}
\mathcal{D}_{\text{FD}}(\partial f_Q) &= \frac{4d}{N_T(d+1)\epsilon^2} + \langle (\partial f_Q)^2 \rangle g_1(\epsilon/2), \\
\mathcal{D}_{\text{FD}}(\partial \partial f_Q) &= \frac{18d}{N_T(d+1)\epsilon^4} + \langle (\partial \partial f_Q)^2 \rangle g_2(\epsilon/2), \\
\mathcal{D}_{\text{FD}}(\partial \partial' f_Q) &= \frac{16d}{N_T(d+1)\epsilon^4} + \langle (\partial \partial' f_Q)^2 \rangle g_2(\epsilon/2), \tag{C5}
\end{aligned}$$

with $g_1(\epsilon/2) = [1 - \text{sinc}(\epsilon/2)]^2$, $g_2(\epsilon/2) = \{1 - [\text{sinc}(\epsilon/2)]^2\}^2$, and all the three average terms $\langle (\partial f_Q)^2 \rangle$, $\langle (\partial \partial f_Q)^2 \rangle$ and $\langle (\partial \partial' f_Q)^2 \rangle$ have been rigorously worked out in Appendices B 2–B 4.

Keeping in their arbitrary forms, we can proceed to minimize all these MSEs and derive optimal FD estimators in the regime $\epsilon \ll 1$. Using the Taylor approximations $g_1(\epsilon/2) \cong$

$\epsilon^4/576$ and $g_2(\epsilon/2) \cong \epsilon^4/144$, we identify two different functional structures

$$\begin{aligned} \mathcal{D}_{\text{FD}}(\partial f_Q) &\cong \frac{A_1}{\epsilon^2} + \frac{\epsilon^4}{576} B_1, \\ \mathcal{D}_{\text{FD}}(\partial\partial f_Q), \mathcal{D}_{\text{FD}}(\partial\partial' f_Q) &\cong \frac{A_2}{\epsilon^4} + \frac{\epsilon^4}{144} B_2, \end{aligned} \quad (\text{C6})$$

where $A_1, A_2, B_1, B_2 > 0$.

Now, optimizing over ϵ for both kinds of structures,

$$\begin{aligned} \min_{\epsilon} \left\{ \frac{A_1}{\epsilon^2} + \frac{\epsilon^4 B_1}{576} \right\} &= \left(\frac{3A_1^2 B_1}{256} \right)^{1/3} \text{ with } \epsilon = \left(\frac{288A_1}{B_1} \right)^{1/6}, \\ \min_{\epsilon} \left\{ \frac{A_2}{\epsilon^4} + \frac{\epsilon^4 B_2}{144} \right\} &= \frac{\sqrt{A_2 B_2}}{6} \text{ with } \epsilon = \left(\frac{144A_2}{B_2} \right)^{1/8}. \end{aligned} \quad (\text{C7})$$

These give the complete set of approximate ϵ_{opt} s for all types of sampling inasmuch as

$$\begin{aligned} \epsilon_{\text{opt}}(\partial f_Q) &\cong \left[\frac{1152d}{\langle (\partial f_Q)^2 \rangle N_T(d+1)} \right]^{1/6}, \\ \epsilon_{\text{opt}}(\partial\partial f_Q) &\cong \left[\frac{2592d}{\langle (\partial\partial f_Q)^2 \rangle N_T(d+1)} \right]^{1/8}, \\ \epsilon_{\text{opt}}(\partial\partial' f_Q) &\cong \left[\frac{2304d}{\langle (\partial\partial' f_Q)^2 \rangle N_T(d+1)} \right]^{1/8}, \end{aligned} \quad (\text{C8})$$

along with the optimized MSEs:

$$\begin{aligned} \mathcal{D}_{\text{FD,opt}}(\partial f_Q) &= \left[\frac{3d^2 \langle (\partial f_Q)^2 \rangle}{16N_T^2(d+1)^2} \right]^{1/3}, \\ \mathcal{D}_{\text{FD,opt}}(\partial\partial f_Q) &= \sqrt{\frac{d \langle (\partial\partial f_Q)^2 \rangle}{2N_T(d+1)}}, \\ \mathcal{D}_{\text{FD,opt}}(\partial\partial' f_Q) &= \frac{2}{3} \sqrt{\frac{d \langle (\partial\partial' f_Q)^2 \rangle}{N_T(d+1)}}. \end{aligned} \quad (\text{C9})$$

The final task is then to substitute the correct expressions of $\langle \partial f_{Q,k} \rangle^2$, $\langle \partial\partial f_{Q,k} \rangle^2$ and $\langle \partial\partial' f_{Q,k} \rangle^2$ that are applicable to the relevant case in point as listed in Fig. 7. Enjoying the fruits of our labor in Appendices B 2–B 4, summarized in Table I, we observe that only Case I supplies exact analytical expressions of these averages, whereas all other cases provide only upper bounds. Optimally tuned estimators, therefore, refer to either those that minimize the exact expression of two-design-averaged MSEs in Case I, or MSE upper bounds in all other cases.

Analytical formulations of approximately optimal FD estimators for any case require substitutions of the answers from Table I, which may be done if so desired. For benchmarking with the PS strategy, however, all one needs is to recognize that Table I implies that $\langle \partial f_{Q,k} \rangle$, $\langle \partial\partial f_{Q,k} \rangle$, $\langle \partial\partial' f_{Q,k} \rangle \leq O(1/d)$ for large d , so that

$$\begin{aligned} \mathcal{D}_{\text{FD,opt}}(\partial f_Q) &\lesssim O\left(\frac{1}{N_T^{2/3} d^{1/3}}\right), \\ \mathcal{D}_{\text{FD,opt}}(\partial\partial f_Q), \mathcal{D}_{\text{FD,opt}}(\partial\partial' f_Q) &\lesssim O\left(\frac{1}{N_T^{1/2} d^{1/2}}\right), \end{aligned} \quad (\text{C10})$$

all scales exponentially with n , which are compatible with the barren-plateau phenomenon that also commensurately scales all gradient- and Hessian-component squared-magnitudes exponentially with n .

4. Optimally tuned GD estimators for general cases

The generalization of FD estimators as defined by (9) and (10), carries the same basic functional structure $\mathcal{D}_{\text{GD}} = \mathbf{c}^\top \mathbf{M} \mathbf{c}$ in their MSEs in (A10) and (A11). The only additional step one needs to perform is the minimization over normalized \mathbf{c} . To do this, we first introduce the Lagrange function

$$\mathcal{L} = \mathbf{c}^\top \mathbf{M} \mathbf{c} - 2\lambda (\mathbf{c}^\top \mathbf{I} - 1) \quad (\text{C11})$$

that is to be optimized, where λ is the Lagrange multiplier that takes care of the normalization constraint, and the factor of 2 in front of λ is introduced for convenience that will become clear very soon. An arbitrary variation of \mathcal{L} over \mathbf{c} gives

$$\delta \mathcal{L} = \delta \mathbf{c}^\top \mathbf{M} \mathbf{c} + \mathbf{c}^\top \mathbf{M} \delta \mathbf{c} - \lambda (\delta \mathbf{c}^\top \mathbf{I} + \mathbf{I}^\top \delta \mathbf{c}). \quad (\text{C12})$$

A minimization of \mathcal{L} , akin to the constrained minimization of \mathcal{D}_{GD} , is done by setting $\delta \mathcal{L} = 0$, resulting in $\mathbf{M} \mathbf{c} = \lambda \mathbf{I}$. As \mathbf{M} is invertible, solving λ yields the extremal equation

$$\mathbf{c} = \frac{\mathbf{M}^{-1} \mathbf{I}}{\mathbf{I}^\top \mathbf{M}^{-1} \mathbf{I}}. \quad (\text{C13})$$

In other words, the optimal \mathcal{D}_{GD} s may be obtained by first substituting \mathbf{c} in \mathcal{D}_{GD} with the optimal one defined in (C13), and, next, minimizing the result over ϵ .

Just like for FD estimators, this second minimization over ϵ can be done numerically, the results of which are used in Figs. 3 and 4. To benchmark GD estimators with the PS ones, we may again resort to looking at upper bounds of \mathcal{D}_{GD} . Thankfully, the Cauchy-Schwarz inequality remains our dearest friend for this task, awarding us with $(\mathbf{I}^\top \mathbf{I})^2 \leq \mathbf{I}^\top \mathbf{M} \mathbf{I} \mathbf{I}^\top \mathbf{M}^{-1} \mathbf{I}$, and consequently $\mathcal{D}_{\text{GD,opt}} \leq J^{-2} \min_{\epsilon > 0} \mathbf{I}^\top \mathbf{M} \mathbf{I}$.

We now take advantage of the fact that the J -dimensional \mathbf{M} s, which are more precisely listed in (A10), give rise to $\mathbf{M} \equiv \mathbf{I}^\top \mathbf{M} \mathbf{I}$ that are multidimensional analogs of the right-hand sides in (C5), namely

$$\begin{aligned} M_{\partial f_Q} &\cong \frac{4Jd H_{J,2}}{N_T(d+1)\epsilon^2} + \epsilon^4 \langle (\partial f_{Q,k})^2 \rangle \frac{J^2(J+1)^2(2J+1)^2}{20736}, \\ M_{\partial\partial f_Q} &\cong \frac{(2J+1)d}{N_T(d+1)\epsilon^4} (4H_{J,2}^2 + 2H_{J,4}) \\ &\quad + \epsilon^4 \langle (\partial\partial f_{Q,k})^2 \rangle \frac{J^2(J+1)^2(2J+1)^2}{5184}, \\ M_{\partial\partial' f_Q} &\cong \frac{16Jd H_{J,4}}{N_T(d+1)\epsilon^4} \\ &\quad + \epsilon^4 \langle (\partial\partial' f_{Q,k})^2 \rangle \frac{J^2(J+1)^2(2J+1)^2}{5184}, \end{aligned} \quad (\text{C14})$$

for $\epsilon \ll 1$, where $H_{j,k} = \sum_{m=1}^j m^{-k}$ is the generalized harmonic number. Recalling the results in (C7), we write down

the optimized upper bounds

$$\begin{aligned}
\mathcal{D}_{\text{GD,opt}}(\partial f_Q) &\lesssim \frac{1}{J^2} \left[\frac{\langle (\partial f)^2 \rangle d^2 J^4 (1+J)^2 (1+2J)^2 H_{J,2}^2}{192 N_T^2 (1+d)^2} \right]^{1/3}, \\
\mathcal{D}_{\text{GD,opt}}(\partial \partial f_Q) &\lesssim \frac{1}{36 J^2} \sqrt{\frac{\langle (\partial \partial f)^2 \rangle d J^2 (J+1)^2 (2J+1)^3 (4H_{J,2}^2 + 2H_{J,4})}{N_T (d+1)}}, \\
\mathcal{D}_{\text{GD,opt}}(\partial \partial' f_Q) &\lesssim \frac{1}{9 J^2} \sqrt{\frac{\langle (\partial \partial' f)^2 \rangle d J^3 (1+J)^2 (1+2J)^2 H_{J,4}}{N_T (d+1)}}, \tag{C15}
\end{aligned}$$

all of which reduce to the corresponding equalities in (C9) for $J = 1$ as $H_{1,k} = 1$. Unsurprisingly, for large d ,

$$\begin{aligned}
\mathcal{D}_{\text{GD,opt}}(\partial f_Q) &\lesssim O\left(\frac{1}{N_T^{2/3} d^{1/3}}\right), \\
\mathcal{D}_{\text{GD,opt}}(\partial \partial f_Q), \mathcal{D}_{\text{GD,opt}}(\partial \partial' f_Q) &\lesssim O\left(\frac{1}{N_T^{1/2} d^{1/2}}\right). \tag{C16}
\end{aligned}$$

These inequalities are sufficient to again show that the optimal MSEs for the GD estimators scale with the respective component squared magnitudes.

5. Optimally tuned SPS estimators for general cases

For general cases, the $\mathcal{D}_{\text{SPS}}(\cdot)$ expressions read

$$\begin{aligned}
\mathcal{D}_{\text{SPS}}(\partial f_Q) &= \frac{d\lambda^2}{N_T(d+1)} + \langle (\partial f_Q)^2 \rangle (1-\lambda)^2, \\
\mathcal{D}_{\text{SPS}}(\partial \partial f_Q) &= \frac{9d\lambda^2}{8N_T(d+1)} + \langle (\partial \partial f_Q)^2 \rangle (1-\lambda)^2, \\
&\quad \text{(diagonal Hessian components)} \\
\mathcal{D}_{\text{SPS}}(\partial \partial' f_Q) &= \frac{d\lambda^2}{N_T(d+1)} + \langle (\partial \partial' f_Q)^2 \rangle (1-\lambda)^2, \\
&\quad \text{(off-diagonal Hessian components)} \tag{C17}
\end{aligned}$$

where $s = \pi/2$. After optimizing the scaling prefactors,

$$\begin{aligned}
\mathcal{D}_{\text{SPS,opt}}(\partial f_Q) &= \frac{d \langle (\partial f_Q)^2 \rangle}{d + (d+1) \langle (\partial f_Q)^2 \rangle N_T}, \\
\mathcal{D}_{\text{SPS,opt}}(\partial \partial f_Q) &= \frac{9d \langle (\partial \partial f_Q)^2 \rangle}{9d + 8(d+1) \langle (\partial \partial f_Q)^2 \rangle N_T}, \\
&\quad \text{(diagonal Hessian components)} \\
\mathcal{D}_{\text{SPS,opt}}(\partial \partial' f_Q) &= \frac{d \langle (\partial \partial' f_Q)^2 \rangle}{d + (d+1) \langle (\partial \partial' f_Q)^2 \rangle N_T}. \\
&\quad \text{(off-diagonal Hessian components)} \tag{C18}
\end{aligned}$$

It is straightforward to verify that $\mathcal{D}_{\text{SPS,opt}}(\partial f_Q)$, $\mathcal{D}_{\text{SPS,opt}}(\partial \partial f_Q)$, and $\mathcal{D}_{\text{SPS,opt}}(\partial \partial' f_Q)$ are respectively monotonically increasing in $\langle (\partial f_Q)^2 \rangle$, $\langle (\partial \partial f_Q)^2 \rangle$, and $\langle (\partial \partial' f_Q)^2 \rangle$, since the derivatives

$$\begin{aligned}
\frac{\partial \mathcal{D}_{\text{SPS,opt}}(\partial f_Q)}{\partial \langle (\partial f_Q)^2 \rangle} &= \frac{d^2}{[d + (d+1) \langle (\partial f_Q)^2 \rangle N_T]^2}, \\
\frac{\partial \mathcal{D}_{\text{SPS,opt}}(\partial \partial f_Q)}{\partial \langle (\partial \partial f_Q)^2 \rangle} &= \frac{81d^2}{[9d + 8(d+1) \langle (\partial \partial f_Q)^2 \rangle N_T]^2}, \\
\frac{\partial \mathcal{D}_{\text{SPS,opt}}(\partial \partial' f_Q)}{\partial \langle (\partial \partial' f_Q)^2 \rangle} &= \frac{d^2}{[d + (d+1) \langle (\partial \partial' f_Q)^2 \rangle N_T]^2}, \tag{C19}
\end{aligned}$$

are all non-negative. It then follows from Table I that the SPS MSEs are always smaller than the PS MSEs as $\langle (\partial f_Q)^2 \rangle$, $\langle (\partial \partial f_Q)^2 \rangle$, and $\langle (\partial \partial' f_Q)^2 \rangle$ are all less than one. Based on (C18), in the limit $N_T \gg d$, $\mathcal{D}_{\text{SPS,opt}}(\cdot) \rightarrow \mathcal{D}_{\text{PS}}(\cdot)$.

APPENDIX D: SIMPLIFIED ELECTRONIC DESCRIPTION OF A WATER MOLECULE

In the quantum-eigsolver scenario, the observable $O = H - h_0 1$, with $h_0 = \text{tr}\{H\}/d$, is defined as a trace-subtracted Hamilton operator H that describes the dynamics of electrons in a water molecule. Under the Hartree-Fock approximation [81], every electron in the molecule is treated as an independent particle that experiences both the Coulomb potential from the nuclei and a mean field generated by all other electrons. The results in Fig. 4(b) are generated by imposing an additional restriction on the electronic excitation to four active orbitals. An application of the Jordan-Wigner transformation turns the resulting Hartree-Fock Hamilton operator into a linear combination of multiqubit Pauli operators weighted by coefficients listed in Table II.

TABLE II. The complete sheet of all 96 multiqubit Pauli components and their respective weights (of magnitudes larger than 10^{-3}) that constitute the observable O describing a neutral water molecule in the minimal basis set $sto-3g$, where electronic excitations are restricted to four active orbitals. The two hydrogen (H) atoms and one oxygen (O) atom are geometrically positioned according to the respective spatial (x, y, z) coordinates—H: $(-0.0399, -0.0038, 0.0)$, O: $(1.5780, 0.8540, 0.0)$, H: $(2.7909, -0.5159, 0.0)$. All coefficients are generated using the QCHEM module in the “PennyLane” PYTHON package [74].

k	h_k	O_{k1}	O_{k2}	O_{k3}	O_{k4}	O_{k5}	O_{k6}	O_{k7}	O_{k8}
1	-0.180625859	1	1	1	1	1	1	Z	1
2	-0.180625859	1	1	1	1	1	1	1	Z
3	-0.159587991	1	1	1	1	Z	1	1	1
4	-0.159587991	1	1	1	1	1	Z	1	1
5	0.174193924	1	1	Z	1	1	1	1	1
6	0.174193924	1	1	1	Z	1	1	1	1
7	0.227570968	Z	1	1	1	1	1	1	1
8	0.227570968	1	Z	1	1	1	1	1	1
9	0.112704888	1	1	1	1	Z	1	Z	1
10	0.112704888	1	1	1	1	1	Z	1	Z
11	0.119520678	Z	1	1	1	Z	1	1	1
12	0.119520678	1	Z	1	1	1	Z	1	1
13	0.134010069	Z	1	1	1	1	1	Z	1
14	0.134010069	1	Z	1	1	1	1	1	Z
15	0.137351346	Z	1	1	1	1	Z	1	1
16	0.137351346	1	Z	1	1	Z	1	1	1
17	0.137670707	1	1	Z	1	Z	1	1	1
18	0.137670707	1	1	1	Z	1	Z	1	1
19	0.141387333	1	1	1	1	Z	1	1	Z
20	0.141387333	1	1	1	1	1	Z	Z	1
21	0.147230746	1	1	Z	1	1	Z	1	1
22	0.147230746	1	1	1	Z	Z	1	1	1
23	0.149265817	1	1	1	1	Z	Z	1	1
24	0.149731063	1	1	Z	1	1	1	Z	1
25	0.149731063	1	1	1	Z	1	1	1	Z
26	0.151377428	Z	1	1	1	1	1	1	Z
27	0.151377428	1	Z	1	1	1	1	Z	1
28	0.154354359	1	1	1	1	1	1	Z	Z
29	0.155818816	1	1	Z	1	1	1	1	Z
30	0.155818816	1	1	1	Z	1	1	Z	1
31	0.167560719	Z	1	Z	1	1	1	1	1
32	0.167560719	1	Z	1	Z	1	1	1	1
33	0.181433521	Z	1	1	Z	1	1	1	1
34	0.181433521	1	Z	Z	1	1	1	1	1
35	0.19391146	Z	Z	1	1	1	1	1	1
36	0.220039773	1	1	Z	Z	1	1	1	1
37	-0.028682446	1	1	1	1	Y	Y	X	X
38	-0.028682446	1	1	1	1	X	X	Y	Y
39	-0.017830668	Y	Y	1	1	X	X	1	1
40	-0.017830668	X	X	1	1	Y	Y	1	1
41	-0.01736736	Y	Y	1	1	1	1	X	X
42	-0.01736736	X	X	1	1	1	1	Y	Y
43	-0.013872802	Y	Y	X	X	1	1	1	1
44	-0.013872802	X	X	Y	Y	1	1	1	1
45	-0.00956004	1	1	Y	Y	X	X	1	1
46	-0.00956004	1	1	X	X	Y	Y	1	1
47	-0.006087753	1	1	Y	Y	1	1	X	X
48	-0.006087753	1	1	X	X	1	1	Y	Y

TABLE II. (Continued.)

k	h_k	O_{k1}	O_{k2}	O_{k3}	O_{k4}	O_{k5}	O_{k6}	O_{k7}	O_{k8}
49	0.006087753	1	1	Y	X	1	1	X	Y
50	0.006087753	1	1	X	Y	1	1	Y	X
51	0.00956004	1	1	Y	X	X	Y	1	1
52	0.00956004	1	1	X	Y	Y	X	1	1
53	0.01130811	1	Y	Z	Z	1	Y	1	1
54	0.01130811	1	X	Z	Z	1	X	1	1
55	0.013872802	Y	X	X	Y	1	1	1	1
56	0.013872802	X	Y	Y	X	1	1	1	1
57	0.01736736	Y	X	1	1	1	1	X	Y
58	0.01736736	X	Y	1	1	1	1	Y	X
59	0.017830668	Y	X	1	1	X	Y	1	1
60	0.017830668	X	Y	1	1	Y	X	1	1
61	0.028682446	1	1	1	1	Y	X	X	Y
62	0.028682446	1	1	1	1	X	Y	Y	X
63	0.029818179	Y	Z	Z	1	Y	1	1	1
64	0.029818179	X	Z	Z	1	X	1	1	1
65	0.029818179	1	Y	1	Z	Z	Y	1	1
66	0.029818179	1	X	1	Z	Z	X	1	1
67	0.030109333	Y	Z	1	Z	Y	1	1	1
68	0.030109333	X	Z	1	Z	X	1	1	1
69	0.030109333	1	Y	Z	1	Z	Y	1	1
70	0.030109333	1	X	Z	1	Z	X	1	1
71	0.030791132	Y	1	Z	Z	Y	1	1	1
72	0.030791132	X	1	Z	Z	X	1	1	1
73	0.043763244	Y	Z	Z	Z	Y	1	1	1
74	0.043763244	X	Z	Z	Z	X	1	1	1
75	0.043763244	1	Y	Z	Z	Z	Y	1	1
76	0.043763244	1	X	Z	Z	Z	X	1	1
77	-0.0145648	1	Y	Z	Z	X	1	X	Y
78	-0.0145648	1	Y	Z	Z	Y	1	Y	Y
79	-0.0145648	1	X	Z	Z	X	1	X	X
80	-0.0145648	1	X	Z	Z	Y	1	Y	X
81	0.010541633	Y	Z	Z	Z	Y	1	1	Z
82	0.010541633	X	Z	Z	Z	X	1	1	Z
83	0.010541633	1	Y	Z	Z	Z	Y	Z	1
84	0.010541633	1	X	Z	Z	Z	X	Z	1
85	0.01130811	Y	Z	Z	Z	Y	Z	1	1
86	0.01130811	X	Z	Z	Z	X	Z	1	1
87	0.025106432	Y	Z	Z	Z	Y	1	Z	1
88	0.025106432	X	Z	Z	Z	X	1	Z	1
89	0.025106432	1	Y	Z	Z	Z	Y	1	Z
90	0.025106432	1	X	Z	Z	Z	X	1	Z
91	0.030791132	Z	Y	Z	Z	Z	Y	1	1
92	0.030791132	Z	X	Z	Z	Z	X	1	1
93	-0.0145648	Y	Z	Z	Z	Z	Y	X	X
94	-0.0145648	X	Z	Z	Z	Z	X	Y	Y
95	0.0145648	Y	Z	Z	Z	Z	X	X	Y
96	0.0145648	X	Z	Z	Z	Z	Y	Y	X

- [1] I. Chuang and M. Nielsen, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, 2000)
- [2] T. D. Ladd, F. Jelezko, R. Laflamme, Y. Nakamura, C. Monroe, and J. L. O'Brien, Quantum computers, *Nature (London)* **464**, 45 (2010).
- [3] E. T. Campbell, B. M. Terhal, and C. Vuillot, Roads towards fault-tolerant universal quantum computation, *Nature (London)* **549**, 172 (2017).
- [4] B. Lekitsch, S. Weidt, A. G. Fowler, K. Mølmer, S. J. Devitt, C. Wunderlich, and W. K. Hensinger, Blueprint for a microwave trapped ion quantum computer, *Sci. Adv.* **3**, e1601540 (2017).
- [5] D. E. Deutsch, A. Barenco, and A. Ekert, Universality in quantum computation, *Proc. Roy. Soc. Lond. A* **449**, 669 (1995).
- [6] A. Barenco, C. H. Bennett, R. Cleve, D. P. DiVincenzo, N. Margolus, P. Shor, T. Sleator, J. A. Smolin, and H. Weinfurter, Elementary gates for quantum computation, *Phys. Rev. A* **52**, 3457 (1995).
- [7] B.-G. Englert, C. Kurtsiefer, and H. Weinfurter, Universal unitary gate for single-photon two-qubit states, *Phys. Rev. A* **63**, 032303 (2001).
- [8] S. D. Bartlett, B. C. Sanders, S. L. Braunstein, and K. Nemoto, Efficient Classical Simulation of Continuous Variable Quantum Information Processes, *Phys. Rev. Lett.* **88**, 097904 (2002).
- [9] A. Sawicki, L. Mattioli, and Z. Zimborás, Universality verification for a set of quantum gates, *Phys. Rev. A* **105**, 052602 (2022).
- [10] L. K. Grover, A fast quantum mechanical algorithm for database search, in *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '96 (Association for Computing Machinery, New York, 1996), p. 212–219.
- [11] P. W. Shor, Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer, *SIAM J. Comput.* **26**, 1484 (1997).
- [12] R. Raussendorf and H. J. Briegel, A One-Way Quantum Computer, *Phys. Rev. Lett.* **86**, 5188 (2001).
- [13] A. Kitaev, Fault-tolerant quantum computation by anyons, *Ann. Phys. (NY)* **303**, 2 (2003).
- [14] R. Raussendorf, J. Harrington, and K. Goyal, Topological fault-tolerance in cluster state quantum computation, *New J. Phys.* **9**, 199 (2007).
- [15] A. Sehrawat, L. H. Nguyen, and B.-G. Englert, Test-state approach to the quantum search problem, *Phys. Rev. A* **83**, 052311 (2011).
- [16] A. Montanaro, Quantum algorithms: An overview, *npj Quantum Inf.* **2**, 15023 (2016).
- [17] E. Knill, R. Laflamme, and W. H. Zurek, Resilient quantum computation, *Science* **279**, 342 (1998).
- [18] D. Franklin and F. T. Chong, Challenges in reliable quantum computing, in *Nano, Quantum and Molecular Computing: Implications to High Level Design and Validation*, edited by S. K. Shukla and R. I. Bahar (Springer US, Boston, 2004), pp. 247–266.
- [19] D. Aharonov and M. Ben-Or, Fault-tolerant quantum computation with constant error rate, *SIAM J. Comput.* **38**, 1207 (2008).
- [20] E. Knill, Approximation by Quantum Circuits, Technical Report LAUR-95-2225, Los Alamos National Laboratory, 1995.
- [21] J. Preskill, Quantum Computing in the NISQ era and beyond, *Quantum* **2**, 79 (2018).
- [22] T. R. Bromley, J. M. Arrazola, S. Jahangiri, J. Izaac, N. Quesada, A. D. Gran, M. Schuld, J. Swinarton, Z. Zabaneh, and N. Killoran, Applications of near-term photonic quantum computers: software and algorithms, *Quantum Sci. Technol.* **5**, 034010 (2020).
- [23] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke, W.-K. Mok, S. Sim, L.-C. Kwek, and A. Aspuru-Guzik, Noisy intermediate-scale quantum algorithms, *Rev. Mod. Phys.* **94**, 015004 (2022).
- [24] A. Finnila, M. Gomez, C. Sebenik, C. Stenson, and J. Doll, Quantum annealing: A new method for minimizing multidimensional functions, *Chem. Phys. Lett.* **219**, 343 (1994).
- [25] T. Kadowaki and H. Nishimori, Quantum annealing in the transverse ising model, *Phys. Rev. E* **58**, 5355 (1998).
- [26] S. Aaronson and A. Arkhipov, The computational complexity of linear optics, in *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, STOC '11 (Association for Computing Machinery, New York, 2011), p. 333–342.
- [27] S. Aaronson, A linear-optical proof that the permanent is #P-hard, *Proc. Roy. Soc. Lond. A* **467**, 3393 (2011).
- [28] C. S. Hamilton, R. Kruse, L. Sansoni, S. Barkhofen, C. Silberhorn, and I. Jex, Gaussian Boson Sampling, *Phys. Rev. Lett.* **119**, 170501 (2017).
- [29] A. Trabesinger, Quantum simulation, *Nature Phys.* **8**, 263 (2012).
- [30] I. M. Georgescu, S. Ashhab, and F. Nori, Quantum simulation, *Rev. Mod. Phys.* **86**, 153 (2014).
- [31] J. Biamonte, Universal variational quantum computation, *Phys. Rev. A* **103**, L030401 (2021).
- [32] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, Variational quantum algorithms, *Nature Rev. Phys.* **3**, 625 (2021).
- [33] Y. Cao, J. Romero, J. P. Olson, M. Degroote, P. D. Johnson, M. Kieferova, I. D. Kivlichan, T. Menke, B. Peropadre, N. P. D. Sawaya, S. Sim, L. Veis, and A. Aspuru-Guzik, Quantum chemistry in the age of quantum computing, *Chem. Rev.* **119**, 10856 (2019).
- [34] S. Endo, Z. Cai, S. C. Benjamin, and X. Yuan, Hybrid quantum-classical algorithms and quantum error mitigation, *J. Phys. Soc. Jpn.* **90**, 032001 (2021).
- [35] S. McArdle, S. Endo, A. Aspuru-Guzik, S. C. Benjamin, and X. Yuan, Quantum computational chemistry, *Rev. Mod. Phys.* **92**, 015003 (2020).
- [36] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, A variational eigenvalue solver on a photonic quantum processor, *Nature Commun.* **5**, 4213 (2014).
- [37] D. Wecker, M. B. Hastings, and M. Troyer, Progress towards practical quantum variational algorithms, *Phys. Rev. A* **92**, 042303 (2015).
- [38] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, The theory of variational hybrid quantum-classical algorithms, *New J. Phys.* **18**, 023023 (2016).
- [39] E. Farhi, J. Goldstone, and S. Gutmann, A Quantum Approximate Optimization Algorithm, Technical Report MIT-CTP/4610, MIT, 2014.
- [40] L. Zhou, S.-T. Wang, S. Choi, H. Pichler, and M. D. Lukin, Quantum Approximate Optimization Algorithm: Performance,

- Mechanism, and Implementation on Near-Term Devices, *Phys. Rev. X* **10**, 021067 (2020).
- [41] M. Schuld, I. Sinayskiy, and F. Petruccione, An introduction to quantum machine learning, *Contemp. Phys.* **56**, 172 (2014).
- [42] M. Schuld and N. Killoran, Quantum Machine Learning in Feature Hilbert Spaces, *Phys. Rev. Lett.* **122**, 040504 (2019).
- [43] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, *Rev. Mod. Phys.* **91**, 045002 (2019).
- [44] P. Date and W. Smith, Quantum discriminator for binary classification, [arXiv:2009.01235](https://arxiv.org/abs/2009.01235) [quant-ph].
- [45] A. Pérez-Salinas, A. Cervera-Lierta, E. Gil-Fuster, and J. I. Latorre, Data re-uploading for a universal quantum classifier, *Quantum* **4**, 226 (2020).
- [46] T. Dutta, A. Pérez-Salinas, J. P. S. Cheng, J. I. Latorre, and M. Mukherjee, Single-qubit universal classifier implemented on an ion-trap quantum device *Phys. Rev. A* **106**, 012411 (2022).
- [47] T. Goto, Q. H. Tran, and K. Nakajima, Universal Approximation Property of Quantum Machine Learning Models in Quantum-Enhanced Feature Spaces, *Phys. Rev. Lett.* **127**, 090506 (2021).
- [48] S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, 2009).
- [49] J. Fiurášek, Maximum-likelihood estimation of quantum measurement, *Phys. Rev. A* **64**, 024102 (2001).
- [50] J. Řeháček, Z. Hradil, E. Knill, and A. I. Lvovsky, Diluted maximum-likelihood algorithm for quantum tomography, *Phys. Rev. A* **75**, 042108 (2007).
- [51] Y. S. Teo, H. Zhu, B.-G. Englert, J. Řeháček, and Z. Hradil, Quantum-State Reconstruction by Maximizing Likelihood and Entropy, *Phys. Rev. Lett.* **107**, 020404 (2011).
- [52] S. Amari and S. Douglas, Why natural gradient? in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No. 98CH36181)*, New Jersey, Vol. 2 (IEEE, 1998), pp. 1213–1216.
- [53] S. Amari, Natural gradient works efficiently in learning, *Neural Comput.* **10**, 251 (1998).
- [54] B. Koczor and S. C. Benjamin, Quantum natural gradient generalised to non-unitary circuits, *Phys. Rev. A* **106**, 062416 (2022).
- [55] J. Stokes, J. Izaac, N. Killoran, and G. Carleo, Quantum natural gradient, *Quantum* **4**, 269 (2020).
- [56] D. Wierichs, C. Gogolin, and M. Kastoryano, Avoiding local minima in variational quantum eigensolvers with the natural gradient optimizer, *Phys. Rev. Res.* **2**, 043246 (2020).
- [57] S. E. Smart and D. A. Mazziotti, Quantum-classical hybrid algorithm using an error-mitigating n -representability condition to compute the mott metal-insulator transition, *Phys. Rev. A* **100**, 022517 (2019).
- [58] B. van Straaten and B. Koczor, Measurement cost of metric-aware variational quantum algorithms, *PRX Quantum* **2**, 030324 (2021).
- [59] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Quantum circuit learning, *Phys. Rev. A* **98**, 032309 (2018).
- [60] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, Evaluating analytic gradients on quantum hardware, *Phys. Rev. A* **99**, 032331 (2019).
- [61] A. Mari, T. R. Bromley, and N. Killoran, Estimating the gradient and higher-order derivatives on quantum hardware, *Phys. Rev. A* **103**, 012405 (2021).
- [62] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, *Nature Commun.* **9**, 4812 (2018).
- [63] A. Arrasmith, M. Cerezo, P. Czarnik, L. Cincio, and P. J. Coles, Effect of barren plateaus on gradient-free optimization, *Quantum* **5**, 558 (2021).
- [64] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, Cost function dependent barren plateaus in shallow parametrized quantum circuits, *Nature Commun.* **12**, 1791 (2021).
- [65] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, Connecting ansatz expressibility to gradient magnitudes and barren plateaus, *PRX Quantum* **3**, 010313 (2022).
- [66] Y. Du, M.-H. Hsieh, T. Liu, and D. Tao, Expressive power of parametrized quantum circuits, *Phys. Rev. Res.* **2**, 033125 (2020).
- [67] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe, Circuit-centric quantum classifiers, *Phys. Rev. A* **101**, 032308 (2020).
- [68] M. Schuld, R. Sweke, and J. J. Meyer, Effect of data encoding on the expressive power of variational quantum-machine-learning models, *Phys. Rev. A* **103**, 032430 (2021).
- [69] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets, *Nature (London)* **549**, 242 (2017).
- [70] K. Mitarai, T. Yan, and K. Fujii, Generalization of the Output of a Variational Quantum Eigensolver By Parameter Interpolation with a Low-Depth Ansatz, *Phys. Rev. Appl.* **11**, 044087 (2019).
- [71] J.-S. Kim, L. S. Bishop, A. D. Córcoles, S. Merkel, J. A. Smolin, and S. Sheldon, Hardware-efficient random circuits to classify noise in a multiqubit system, *Phys. Rev. A* **104**, 022609 (2021).
- [72] A. W. Harrow and R. A. Low, Random quantum circuits are approximate 2-designs, *Commun. Math. Phys.* **291**, 257 (2009).
- [73] Z. Puchała and J. Miszczak, Symbolic integration with respect to the Haar measure on the unitary groups, *Bull. Polish Acad. Sci.* **65**, 21 (2017).
- [74] https://pennylane.ai/qml/demos/tutorial_quantum_chemistry.html
- [75] P. J. Olver, *Introduction to Partial Differential Equations* (Springer Science & Business Media, Switzerland, 2014).
- [76] G. G. Guerreschi and M. Smelyanskiy, Practical optimization for hybrid quantum-classical algorithms, [arXiv:1701.01450](https://arxiv.org/abs/1701.01450) [quant-ph].
- [77] B. Swartz and B. Wendroff, Generalized finite-difference schemes, *Mathematics of Computation* **23**, 37 (1969).
- [78] K. Zhang, M.-H. Hsieh, L. Liu, and D. Tao, Toward trainability of quantum neural networks, [arXiv:2011.06258](https://arxiv.org/abs/2011.06258) [quant-ph].
- [79] C. Zhao and X.-S. Gao, Analyzing the barren plateau phenomenon in training quantum neural networks with the ZX-calculus, *Quantum* **5**, 466 (2021).
- [80] T. Haug and M. S. Kim, Optimal training of variational quantum algorithms without barren plateaus, [arXiv:2104.14543](https://arxiv.org/abs/2104.14543) [quant-ph].
- [81] R. Seeger and J. A. Pople, Self-consistent molecular orbital methods. XVIII. Constraints and stability in Hartree-Fock theory, *J. Chem. Phys.* **66**, 3045 (1977).