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 a thousand 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 runtimes. While the severity of this problem has indeed been raised [34, 57] and asymptotic error bounds for 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 article, 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 [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, employs 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 $C = C(H_j)$ 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 $\rho$ 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 pre-chosen 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 minimized. An arbitrary quantum circuit of the NISQ device used to run the VQA is a sequence of training $[W_l(\theta)]$ and classically-encoded $[V_l(x)]$ unitary operators, where $\theta_l$ are trainable parameters and $x$ are nontrainable ones.

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

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

and measurement observable $O$. As specific examples of VQAs, in variational quantum eigensolver (VQE) problems [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 $C = \langle 0 |U_{\theta}^\dagger H U_{\theta} |0\rangle$, 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 $f(\theta;x)$ for different classical-data encoding.
parameters \( \mathbf{x} \). Given \( m \) of these parameters \( \{\mathbf{x}_j\}_{j=1}^m \), the quality of the learning algorithm is defined by more sophisticated cost functions such as the mean squared-error
\[
C = \frac{1}{m} \sum_{j=1}^{m} \| f_Q(\mathbf{\theta}, \mathbf{x}_j) - f(\mathbf{x}_j) \|^2.
\]

Gradient-based (and Hessian-based) optimization routines in VQAs would then rely on the computation accuracies of \( \partial_{\mathbf{\theta}} f_Q \) and \( \partial_{\mathbf{\theta}} f_Q \), with all arguments dropped from hereon for notational simplicity unless otherwise necessary. The estimations of gradient \( \partial_{\mathbf{\theta}} f_Q \) and Hessian components \( \partial_{\mathbf{\theta}}(\partial_{\mathbf{\theta}} f_Q) \) require the specifications of actual physical ansätze structures that make up the \( W_1 \) and \( V_1 \) operators. A very common type of parametrized quantum-circuit multiqubit \textit{ansätze} encode training parameters \( \mathbf{\theta} \) 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 \( R(\mathbf{\phi}) \) defined by the trivariate angular parameter \( \mathbf{\phi} = (\phi_1 \phi_2 \phi_3)^\top \), \( R(\mathbf{\phi}) = R_Z(\phi_3)R_Y(\phi_2)R_Z(\phi_3) \), into a sequence of rotations about the \( Y \)- and \( Z \)-axes, so that each basic Pauli rotation gate \( R_z(\phi_3) = e^{-i\phi_3 \sigma_z}/2 \) is attributed to the relevant single-qubit Pauli operator \( \sigma_z \) of the \( j \)th axis.

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

- The structures of \( V_1 \) that house the nontrainable parameters \( \mathbf{x} \) can be arbitrary.
- 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 shall assume that each \( O_k \) is sampled independently for the same set of PEPCQ parameters.

We note that the set of \( W_n \),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 \( \langle W_n \rangle \) are approximately those from the Haar measure [73] over the \( \mathbb{U}(2^n) \) group.

III. FIGURE OF MERIT: MEAN SQUARED-ERROR

Throughout the article, 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 \( \langle \cdots \rangle \), which, for instance, could mean averages according to the Haar measure. Numerical or vectorial averages over NISQ-sampling distributions shall be denoted by \( \mathbb{E}[\cdots] \). Those over the nontrainable parameters \( \mathbf{x} \) are denoted by \( \cdots \).

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

\[
\text{MSE}(Y) = \mathbb{E}[\hat{Y} - Y]^2,
\]

where \( \hat{Y} \) 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 shall 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, [36]). Using the spectral decomposition of the Pauli observable \( O_k = \sum_{l=0}^{d-1} |o_{kl}| \alpha_{kl}, \langle \alpha_{kl} \rangle \), sampling the individual function components

\[
f_Q,\alpha(\mathbf{\theta}; \mathbf{x}) = \sum_{l=0}^{d-1} o_{kl} \langle |o_{kl}| U_{\mathbf{\theta}}(\mathbf{x}) \rangle^2 = \sum_{l=0}^{d-1} o_{k|l} p_{kl}(\mathbf{\theta}; \mathbf{x})
\]

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].

![FIG. 2. Examples of “hardware-efficient” setups for trainable unitary operators \( W_1 \) (with two repeated units) and those of fixed encodings \( V_1 \), shown here for a four-qubit system with \( \mathbf{\theta}_1 = (\theta_{11}, \theta_{22}, \theta_{24})^\top \) and \( \mathbf{x} = (x_1 \ldots x_4)^\top \). Each single-qubit green block in \( W_1 \) represents a rotation operator defined by an angle 3-tuple \( \phi \). For instance, \( R(\phi) = R_Z(\theta_{11})R_Y(\theta_{22})R_Z(\theta_{24}) \). The operator \( V_1 \) may also take similar structures, where \( \mathbf{x} \) is encoded into entanglement-free unitary operators characterized by their respective generators \( H_{kl} \) along with CNOT gates.](image-url)
of \( f_Q = \sum_k h_k f_{Q,k}/\|h\| \) is equivalent to sampling the circuit probabilities \( \sum_{d=0}^{d-1} \mu_{kl}(\theta; \mathbf{x}) = 1 \), where the eigenvalues \( \alpha_{kl} = \pm 1 \). Since \( Y - Y = \sum_k h_k (\hat{Y}_k - Y_k)/\|h\| \), it can be deduced easily, as a result of independence sampling with the \( O_k \)'s, that MSE of \( Y = \sum_k \alpha_{kl}^2 \text{MSE}(|Y_k|)/\|h\|^2 \) if the estimators \( \hat{Y}_k \)'s are unbiased (\( \mathbb{E}[\hat{Y}_k] = Y_k \)).

A usual physical circumstance when sampling a circuit function defined by the measurement observable \( O \) is when detectors measuring the probabilities \( p_{kl}(\theta; \mathbf{x}) \) register “click’s” 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 \( n_{kl}(\theta; \mathbf{x}) = n_{kl}(\theta) = n_{kl}/N \to p_{kl}(\theta; \mathbf{x}) \) corresponding to the number of “clicks” \( n_{kl}(\theta; \mathbf{x}) \) registered at the \( l \)th detector, with \( N = \sum_{l=0}^{d-1} n_{kl}(\theta; \mathbf{x}) \). Thus, the average accuracy in estimating \( f_Q \) using the unbiased estimator

\[
\hat{f}_Q = \sum_k \frac{h_k}{\|h\|} \hat{f}_{Q,k} = \sum_k h_k \frac{1}{\|h\|} \sum_{l=0}^{d-1} \alpha_{kl} n_{kl}(\theta; \mathbf{x})
\]

is quantified by the MSE.

We caution the reader that the MSEs involve circuit averages of gradient and Hessian components that are generally non-trivial 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 operator 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 single upper bounds are calculable. While explicit details on how circuit averaging is carried out are supplied in Appendix A, we simply state that throughout the main text, unless otherwise stated, all analytical MSE expressions shall refer to the so-called two-design sandwich (TDS) condition (Case I in Fig. 7), where every gradient operator is sandwiched between two two-design trainable modules. Upper-bound expressions for all other cases are obtained from Table 1 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 [74, 75] approximates the gradient components \( \partial_{\mu,l} f_{Q,k} \) for each observable basis operator \( O_k \) according to

\[
\text{FD}_{\mu,l} f_{Q,k} = \frac{f_{Q,k}(\theta_{\mu,l} + \epsilon/2; \mathbf{x}) - f_{Q,k}(\theta_{\mu,l} - \epsilon/2; \mathbf{x})}{\epsilon} = \text{sinc}(\epsilon/2) \partial_{\mu,l} f_{Q,k}
\]

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 \( \epsilon \) 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 \( \text{sinc}(\epsilon/2) \approx 1 - \epsilon^2/24 \) for a small \( \epsilon \). 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 \( \text{FD}_{\mu',l'} f_{Q,k} \) on \( \text{FD}_{\mu,l} f_{Q,k} \) in (5), reads

\[
\text{FD}_{\mu,l} \text{FD}_{\mu',l'} f_{Q,k} = \left[ \text{sinc}(\epsilon/2) \right]^4 \partial_{\mu,l} \partial_{\mu',l'} f_{Q,k} = f_Q(\theta_{\mu,l} + \epsilon/2, \theta_{\mu',l'} + \epsilon/2; \mathbf{x}) - f_Q(\theta_{\mu,l} + \epsilon/2, \theta_{\mu',l'} - \epsilon/2; \mathbf{x})
\]

\[
- f_Q(\theta_{\mu,l} - \epsilon/2, \theta_{\mu',l'} + \epsilon/2; \mathbf{x}) + f_Q(\theta_{\mu,l} - \epsilon/2, \theta_{\mu',l'} - \epsilon/2; \mathbf{x})
\]

\[
\frac{\epsilon^2}{2} \frac{\epsilon^2}{2} \frac{\epsilon^2}{2} \frac{\epsilon^2}{2}
\]

The corresponding function estimators that enter \( \text{FD}_{\mu,l} f_{Q,k} \) and \( \text{FD}_{\mu',l'} f_{Q,k} \) 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-\( \epsilon \) gradient and Hessian approximators of the FD strategy, collectively denoted by \( \hat{Y}_{k,\epsilon} = \text{FD}_{\mu,l} f_{Q,k} \), \( \text{FD}_{\mu',l'} f_{Q,k} \), incur errors from both finite-copy sampling and nonzero-\( \epsilon \) approximation. The corresponding MSE is hence decomposable into these two error types:

\[
\text{MSE}(Y) = \mathbb{E} \left[ \left( \hat{Y}_{k,\epsilon} - Y_k \right)^2 \right] + \mathbb{E} \left[ \left( \hat{Y}_{k,\epsilon} - Y_k \right)^2 \right] \text{approx. error}
\]

\[
\Delta^2_{\text{copy}}(Y_{k,\epsilon}) + \Delta^2_{\epsilon}(Y_k)
\]

where \( \Delta^2_{\text{copy}}(Y_k) = 0 \), 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 \( \Delta^2_{\text{copy}} \) 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-\( \epsilon \) approximation error \( \Delta^2_{\epsilon} \) on the other hand, requires the evaluation of \( \mathbb{E} \left[ \left( \partial_{\mu,l} f_{Q,k} \right)^2 \right] \) and \( \mathbb{E} \left[ \left( \partial_{\mu,l} \partial_{\mu',l'} f_{Q,k} \right)^2 \right] \) 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 \( \text{FD}_{\mu,l} f_{Q,k} \), and \( N_T = 3N \) for \( 4N \) respectively for the diagonal and off-diagonal \( \text{FD}_{\mu,l} \text{FD}_{\mu',l'} f_{Q,k} \) 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 con-
dition:

\[
\text{MSE}_{\text{FD}}(\partial f_Q) = \frac{4d}{N_T(d+1)e^2} + \frac{d^2[1 - \text{sinc}(\epsilon/2)]^2}{2(d+1)(d^2 - 1)},
\]

\[
\text{MSE}_{\text{FD}}(\partial^2 f_Q) = \frac{18d}{N_T(d+1)e^4} + \frac{d^2(1 - [\text{sinc}(\epsilon/2)]^2)^2}{4(d+1)(d^2 - 1)},
\]

(8)


diagonal Hessian components

\[
\text{MSE}_{\text{FD}}(\partial^2 f_Q) = \frac{16d}{N_T(d+1)e^4} + \frac{d^4(1 - [\text{sinc}(\epsilon/2)]^2)^2}{(d+1)(d^2 - 1)^2},
\]

(9)

(10)

\[
\text{MSE}_{\text{FD}}(\partial f_Q) = \frac{4d}{N_T(d+1)e^2} + \frac{d^2[1 - \text{sinc}(\epsilon/2)]^2}{2(d+1)(d^2 - 1)},
\]

(8)

(10)

\[
\text{MSE}_{\text{FD}}(\partial^2 f_Q) = \frac{18d}{N_T(d+1)e^4} + \frac{d^2(1 - [\text{sinc}(\epsilon/2)]^2)^2}{4(d+1)(d^2 - 1)},
\]

(9)

(10)

\[
\text{MSE}_{\text{FD}}(\partial^2 f_Q) = \frac{16d}{N_T(d+1)e^4} + \frac{d^4(1 - [\text{sinc}(\epsilon/2)]^2)^2}{(d+1)(d^2 - 1)^2},
\]

(8)

(9)

(10)

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

We may also consider a type of generalized difference (GD) estimation strategy [76] where the corresponding nonzero-\( \epsilon \) gradient and Hessian approximators,

\[
\text{GDgrad}_{\mu,l}^j f_Q,k \equiv \sum_{j=1}^J c_j \text{FD}_{\mu,l}^j f_Q,k,
\]

(9)

\[
\text{GDHess}_{\mu,l,\mu',l'}^j f_Q,k \equiv \sum_{j=1}^J c_j \text{FD}_{\mu,l}^j \text{FD}_{\mu',l'}^j f_Q,k,
\]

are weighted sums of FD approximators of integer multiples of the step size \( \epsilon \), which become the true gradient and Hessian components when \( \epsilon = 0 \) under the normalization constraint \( \sum_{j=1}^J c_j = c^T \mathbf{1} = 1 \) for the coefficient column \( c \), where \( \mathbf{1} \) 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 A4.

\[
\text{MSE}_{\text{PS}}(\partial f_Q) = \frac{d}{N_T(d+1)(\sin s)^2} \geq \frac{d}{N_T(d+1)},
\]

(11)

(12)

(13)

\[
\text{MSE}_{\text{PS}}(\partial f_Q) = \frac{9d}{8N_T(d+1)(\sin s)^4} \geq \frac{9d}{8N_T(d+1)}.
\]

(11)

(12)

(13)

These MSEs are minimized when \( s = \pi/2 \), which is the standard shift value that we shall 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 (\( \Delta_q^2 = 0 \)), unlike FD or GD methods. Because of this, it is a belief that FD, for instance, which requires small \( \epsilon \) 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 [61], the scaled parameter-shift (SPS) estimators were introduced. These estimators are essentially scaled versions of the PS estimators, where \( \text{SPS}_{\mu,l} f_Q = \lambda \text{PS}_{\mu,l} f_Q \) and \( \text{SPS}_{\mu,l} f_Q = \lambda \text{SPS}_{\mu,l} f_Q \) are characterized by an additional prefactor \( \lambda \) 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:

\[
\text{MSE}_{\text{SPS}}(\partial f_Q) = \frac{\text{finite-copy error}}{N_T (d + 1)^2 (\sin s)^2} + \frac{\text{approx. error}}{2(d + 1)^2 (d^2 - 1)^2},
\]

\[
\text{MSE}_{\text{SPS}}(\partial^2 f_Q) = \frac{9d\lambda^2}{8N_T (d + 1)(\sin s)^4} + \frac{d^2 (1 - \lambda)^2}{2(d + 1)(d^2 - 1)^2},
\]

\[
\text{(diagonal Hessian components)}
\]

\[
\text{MSE}_{\text{SPS}}(\partial f_Q') = \frac{d^2}{N_T (d + 1)(\sin s)^2} + \frac{d^4 (1 - \lambda)^2}{4(d + 1)(d^2 - 1)^2},
\]

\[
\text{(off-diagonal Hessian components)}
\]

(14)

It is clear that \( s = \pi/2 \) will optimize all SPS MSEs. We shall see that \( \lambda \) 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 Sec. 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 \( \epsilon \). The second observation is that the MSEs are oscillatory functions of \( \epsilon \) by virtue of the unitary encoding. The third observation has to do with the choice of \( \epsilon \). If one picks very small \( \epsilon \) values, then the finite-copy error \( \Delta_{\text{copy}}^2 \) dominates as \( O(1/\text{poly} \ \epsilon) \). If one, instead, picks larger \( \epsilon \) values, then the approximation error \( \Delta_{\epsilon}^2 \) 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 \( \epsilon_{\text{opt}} \) through minimizing \( \text{MSE}_{\text{FD}} \) over \( \epsilon \) requires the solutions of transcendental equations that do not generally admit analytical forms. Numerical optimization methods are therefore in order. Nevertheless, if \( [1 - \sin\epsilon/2]^2 \approx \epsilon^4/576 \) and \( [1 - \sin\epsilon/2]^2 \approx \epsilon^4/144 \) for small right-hand sides, we may obtain analytical approximations of both \( \epsilon_{\text{opt}} \) and \( \text{MSE}_{\text{FD}} \) for optimal FD gradient and Hessian estimation by approximating \( \Delta_{\epsilon}^2 \) to the smallest order \( O(\epsilon^4) \). Minimizing the resulting leading-order expansions of all \( \text{MSE}_{\text{FD}} \) therefore gives

\[
\epsilon_{\text{opt}} \approx \left( \frac{2304d^2 - 1}{N_T + d} \right)^{1/2} \text{ (\( \partial f_Q \) estimation)};
\]

\[
\epsilon_{\text{opt}} \approx \left( \frac{5184d^2 - 1}{N_T + d} \right)^{1/2} \text{ (\( \partial^2 f_Q \) estimation)};
\]

\[
\epsilon_{\text{opt}} \approx \left( \frac{9216(d^2 - 1)^2}{N_T d^3} \right)^{1/2} \text{ (\( \partial f_Q' \) estimation)}.
\]

\[
(15)
\]

The corresponding approximately optimal MSEs are

\[
\text{MSE}_{\text{FD, opt}}(\partial f_Q) \approx \left( \frac{3}{32} \right)^{1/3} \frac{d^4/3}{(d + 1)(d^2 - 1)^{3/2} N_T^{2/3}};
\]

\[
\text{MSE}_{\text{FD, opt}}(\partial^2 f_Q) \approx \left( \frac{2}{3} \right)^{1/3} \frac{d^{3/2}}{2(d + 1)(d^2 - 1)^{1/2} N_T^{1/2}};
\]

\[
\text{MSE}_{\text{FD, opt}}(\partial f_Q') \approx \left( \frac{3}{3} \right)^{1/3} \frac{d^{3/2}}{3(d + 1)(d^2 - 1) N_T^{1/2}}.
\]

(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 \( \text{MSE}_{\text{FD, opt}} \) (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 through B.4 (summarized in Tab. I) that \( \langle \overline{(\partial f_Q)^2} \rangle \), \( \langle \overline{(\partial^2 f_Q)^2} \rangle \) and \( \langle \overline{((\partial f_Q)^2)} \rangle \) 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 counter-intuitive. 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 Tab. 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 \( d \):

\[
\text{opt. FD:} \quad \left< E \left[ \overline{\partial f_Q^2} \right] \right> \to \sin\left(\epsilon_{\text{opt}}/2\right)^2 \frac{1}{2d} + \frac{4}{N_T \epsilon_{\text{opt}}^2},
\]

\[
\text{PS:} \quad \left< E \left[ \overline{\partial f_Q^2} \right] \right> \to \frac{1}{2d} + \frac{1}{N_T} \to \frac{1}{N_T}.
\]

(17)

(18)

Since from numerical evidence, \( \epsilon_{\text{opt}} \) 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 \( \text{MSE}_{\text{PS}} \) is constant for the same \( N_T \).

As a reminder, \( \epsilon_{\text{opt}} \) 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 shall 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 [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 $\epsilon$ and $\epsilon$. For any $\epsilon$, it is easy to carry out the minimization of the MSEs in (A10) over normalized $\epsilon$, as every GD MSE takes the form

$$\text{MSE}_{GD} = c^\top M c. $$

Upon the standard usage of a Lagrange multiplier (see Appendix C.4), it is straightforward to arrive at

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

where $c_{opt} = M^{-1} 1 / (1^\top M^{-1} 1)$ and $M$ 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 $c_{opt}$, although numerical minimization of $\text{MSE}_{GD}$ over $\epsilon$ to obtain the optimal $\text{MSE}_{GD, opt}$ 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 $\text{MSE}_{GD, opt}$ as $N_T$ and $d$ grows. For this purpose, we first make use of the Cauchy–Schwarz inequality $(1^\top 1)^2 \leq 1^\top M 1 1^\top M^{-1} 1$ to arrive at the bound

$$\text{MSE}_{GD, opt} = \min_{\epsilon > 0} \min_{c|c} \text{MSE}_{GD}$$

$$= \min_{\epsilon > 0} (1^\top M^{-1} 1)^{-1} \leq \frac{1}{J^2} \min_{\epsilon > 0} 1^\top M 1. \quad (20)$$

The inequality is saturated when $J = 1$. Next, under the TDS condition and $N_T \gg d$ (or small $\epsilon_{opt}$), we can obtain the following explicit upper bounds:

$$\text{MSE}_{GD, opt} (\partial f_Q) \leq \frac{1}{4} \left[ \frac{d^4 (J + 1)^2 (2J + 1)^2}{6 N_T^2 J^2 (d + 1)^3 (d^2 - 1)} \right]^{1/3} H_{J, 2}^{2/3},$$

$$\text{MSE}_{GD, opt} (\partial \partial f_Q) \leq \frac{J + 1}{36 J} \left( \frac{2dJ + d}{d + 1} \right)^{3/2} \sqrt{\frac{2H_{J, 2}^2 + H_{J, 4}}{N_T (d - 1)}},$$

$$\text{MSE}_{GD, opt} (\partial \partial f_Q) \leq \frac{(J + 1)(2J + 1)}{18 (d + 1)(d^2 - 1)} \left( \frac{d^2 H_{J, 4}}{N_T J} \right)^{1/2}, \quad (21)$$

where $H_{j, k} = \sum_{m=1}^j m^{-k}$ 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: $\text{MSE}_{GD, opt}$ 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 $\text{MSE}_{GD, opt}$. As examples, Fig. 3 illustrates the graphs of the respective $\text{MSE}_{GD, opt}$ 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 $\text{MSE}_{GD, opt}$.

### C. Optimal SPS estimators

The minimization of (14) with respect to $\lambda$, 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 $\lambda$ on the right-hand sides. Hence, one arrives at
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 \( \psi_l = 1 \) 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 \( (L = 5 \text{ and four-repeated units of single-qubit and CNOT gates for each } W_l) \) and 500 numerical experiments per PEPQC. All PEPQCs possess the Hilbert-space dimension \( d = 2^8 \), which are used to estimate \( \partial f_{Q}^{\text{estimation}}, \partial \partial f_{Q}^{\text{estimation}}, \partial \partial \partial f_{Q}^{\text{estimation}} \), \( \partial f_{Q}^{\text{estimation}} \) [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 \( \text{(Case I)} \), 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.

VI. PERFORMANCE

A. Applications in quantum supervised learning and quantum eigensolver problems

We shall 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 MSE as

\[ \lambda_{\text{opt}} = \frac{dN_T}{2d^2 + dN_T - 2} \] (\( \partial f_Q^{\text{estimation}} \) estimation),

\[ \lambda_{\text{opt}} = \frac{9d^2 + 4dN_T - 9}{4d^2} \] (\( \partial \partial f_Q^{\text{estimation}} \) estimation),

\[ \lambda_{\text{opt}} = \frac{d^3N_T}{4(d^2 - 1)^2 + d^3N_T} \] (\( \partial \partial \partial f_Q^{\text{estimation}} \) estimation),

and realizes that \( \lambda_{\text{opt}} \to 1 \) as \( d \ll N_T \to \infty \), and \( \lambda_{\text{opt}} \to 0 \) as \( N_T \ll d \to \infty \). 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

\[ \text{MSE}_{\text{SPS, opt}}(\partial f_Q) = \frac{d^2}{(d + 1)(2d^2 + dN_T - 1)} \],

\[ \text{MSE}_{\text{SPS, opt}}(\partial \partial f_Q) = \frac{9d^2}{2(d + 1)(9d^2 + 4dN_T - 9)} \] (diagonal Hessian components),

\[ \text{MSE}_{\text{SPS, opt}}(\partial \partial \partial f_Q) = \frac{d^4}{(d + 1)[4(d^2 - 1)^2 + d^3N_T]} \] (off-diagonal Hessian components).

For a fixed \( d \), these optimized SPS MSEs converge to the PS MSEs in \( (13) \), as they should.
The document discusses the estimation performances of various VQA schemes, specifically highlighting the estimation errors for the two examples of VQA discussed earlier. The combined errors $\Delta_{\text{copy}}^2$ and $\Delta_2^2$ are compared, with $N_\text{T}$ being greater than a certain critical value $N_\text{c}$ (defined as the value of $N_\text{T}$ for which $\text{MSE}_{\text{FD}}(\text{GD}, \text{opt}) = \text{MSE}_{\text{PS}}$). The contribution of $\Delta_{\text{copy}}^2$ is shown to be small enough to be dominated by the nonzero-$\epsilon$ error $\Delta_2^2$, suggesting that the advantage of an $\epsilon$-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, the document answers the question: “How does $N_\star$ scale with the number of qubits $n$?” Basic intuition suggests that since the FD and GD estimation errors decrease with $n$ according to Sec. V A, while the PS ones do not [recall (13)], the critical value $N_\text{T} = N_\star$ 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 $N_\star \gg d$, we find that

$$N_\star \cong \frac{32(d^2 - 1)}{3d} \quad (\partial f_Q \text{ estimation}) ;$$

$$N_\star \cong \frac{81(d^2 - 1)}{16d} \quad (\partial \theta f_Q \text{ estimation}) ;$$

$$N_\star \cong \frac{9(d^2 - 1)^2}{d^3} \quad (\partial \delta f_Q \text{ estimation}) . \quad (24)$$

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 $N_\star$:

$$N_\star \gtrsim \frac{384J^2(d^2 - 1)}{d(J + 1)^2(2J + 1)^2H_{J,2}} \quad (\partial f_Q \text{ estimation}) ;$$

$$N_\star \gtrsim \frac{6561J^2(d^2 - 1)}{4d(J + 1)^2(2J + 1)^3(2H_{J,2} + H_{J,4})} \quad (\partial \theta f_Q \text{ est.}) ;$$

$$N_\star \gtrsim \frac{6561J(d^2 - 1)^2}{16d^3(J + 1)^2(2J + 1)^2H_{J,4}} \quad (\partial \delta f_Q \text{ estimation}) . \quad (25)$$

While Eqs. (24) and (25) strictly hold only when $N_\star \gg d$, they analytically show, at least in this regime, that $N_\star \gtrsim O(2^n)$. The plots in Fig. 5 clearly show an exponential increase in $N_\star$ with respect to $n$ regardless of whether $\epsilon_{\text{opt}}$ is found with numerical MSE optimization or large $N_\text{T}$ approximation. In particular, for the optimized FD strategy, the differences between the exact numerically obtained $N_\star$’s and those from (24) are very small. In general, $N_\star$ is a useful measure for the sampling efficiency of a particular optimized scheme in question. A larger $N_\star$ implies that the optimized numerical scheme gives a smaller estimation error for a larger range of $1 \leq N_\text{T} \leq N_\star$ in contrast to the analytical PS strategy. A scheme that exhibits an exponentially growing $N_\star$ with respect to $n$ is therefore a much more statistically favorable one over PS for scalable VQAs. Figure 5 illustrates the exponential growth in $N_\star$ 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 $N_\star$ 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 PITFALLS

The results in this article show that numerical estimators possessing free parameters ($\epsilon$ for FD and GD and $\lambda$ 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.

For optimized FD and GD schemes, we reiterate that $\epsilon_{\text{opt}}$ 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 $\epsilon$, so that numerical methods for their minimization are the only resort. Otherwise, analytical approximations of $\epsilon_{\text{opt}}$ 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 $\epsilon_{\text{opt}}$ is typically not small, regardless of whether numerical optimization or analytical approximations are invoked. In fact, as $d$ increases, numerical experience shows that $\epsilon_{\text{opt}}$ grows roughly exponentially with $n$. One can already witness this behavior approximately from (15). This is, apparently, at odds with the usual narrative that $\epsilon_{\text{opt}}$ should be as small as possible, preferably $\epsilon_{\text{opt}} \to 0$, 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. VA) that $\epsilon_{\text{opt}}$ is neither minuscule nor astronomical, but somewhere in between. However, for a fixed $N_T$, as $d$ or $n$ increases, $\epsilon_{\text{opt}}$ tends to larger values because the nonzero-$\epsilon_{\text{opt}}$ approximation error $\Delta_{\text{opt}}^2$ approaches zero while the finite-copy error $\Delta_{\text{copy}}^2$ is asymptotically constant. The value of $\epsilon_{\text{opt}}$ 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. 1). 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 $\epsilon_{\text{opt}}$ 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 ad-
dressed by this article. The analysis of the MSE throughout the manuscript 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, the landscape of $f_Q(\theta; x)$ from a randomized two-design circuit, for instance, is almost flat, so that the problem of distinguishing function magnitudes in different $\theta$ search directions still persists, even with accurate statistical estimations.

An absurd hypothetical scenario would be $n \to \infty$, where the initialization $\langle f_Q^2 \rangle \cong 0 \cong f_Q$, 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

$$D_{\theta_0} = \left\langle \frac{\max\left\{\text{Var}\left[f_Q(\theta + \theta_0; x)\right]\right\}}{\left|f_Q(\theta + \theta_0; x) - f_Q(\theta - \theta_0; x)\right|^2}\right\rangle$$

for some chosen displacement $\theta_0$, when speaking of trainability. This quantifies the worst-case average relative spread (variance) of $f_Q(\theta \pm \theta_0; x)$, which if too large, results in the failure of distinguishing between $f_Q(\theta + \theta_0; x)$ and $f_Q(\theta - \theta_0; x)$ in the direction set by $\theta_0$ 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 $O(1/poly \; d)$, so that $D_{\theta_0} \geq O((poly \; d)/N)$. For large circuits, $N$ must thus at least be exponentially large in $n$ for trainability.

Efforts in ameliorating the effects of barren plateaus would therefore require deeper understanding in both the circuit ansatz [77, 78], and $\theta$-initialization and optimization strategies [79], 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 article. With that said, it cannot be overemphasized that both accurate estimation and trainability are equally important in quantum computation.
cially 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 article.

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 all gradient and Hessian computation from turning into random guesses for a fixed number of sampling copies as the quantum-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, the absence of approximation errors with this analytical rule 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-mitigating power for optimized numerical schemes shall 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 (Grant nos. NRF-2020R1A2C1008609, NRF-2020K2A9A1A06102946, NRF-2019R1A6A1A10073437 and NRF-2022M3E4A1076099) via the Institute of Applied Physics at Seoul National University.

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 $W_j$ that each has a $O(\text{poly}(n, 2))$ circuit depth, so that a randomized $W_j$ 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

$$\int (dU)_{\text{Haar}} U_{j_1 \ldots j_n}^* U_{j_1 \ldots j_n} U_{j_1 k_1} U_{j_2 k_2} = \frac{\delta_{j_1 j_1'} \delta_{j_2 j_2'} \delta_{k_1 k_1'} \delta_{k_2 k_2'} + \delta_{j_1 j_1'} \delta_{j_2 j_2'} \delta_{k_1 k_1'} \delta_{k_2 k_2'} - \delta_{j_1 j_1'} \delta_{j_2 j_2'} \delta_{k_1 k_1'} \delta_{k_2 k_2'}}{d^2 - 1}$$

(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

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

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

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

where $\tau$ 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,

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

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

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

$$
\partial_{\mu_l} f_{Q,k} = \frac{\partial}{\partial \theta} \left( \sum_{\ell=L}^{L+1} V_\ell W_\ell \right) = \frac{1}{2} \left( 0 | B W_{l+1}^{(2)} \partial_{\mu_l} W_l^{(1)} A O_k AV_l W_l B | 0 \right)
$$

(c.c.),

$$
\partial_{\mu_l}^2 f_{Q,k} = \frac{1}{2} \left( 0 | B W_{l+1}^{(2)} \partial_{\mu_l} W_l^{(1)} V_l^\dagger A O_k AV_l W_l B | 0 \right) - \frac{1}{2} f_{Q,k}, \quad (A5)
$$

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

$$
(\partial_{\mu_l})^{2k+1} f_{Q,k} = (-1)^k (\partial_{\mu_l}) f_{Q,k}, \quad (A6)
$$

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

$$
f_{Q,k}(\theta_{\mu_l} + \theta_0; x) = f_{Q,k}(\theta_{\mu_l}; x) + \sin \theta_0 \partial_{\mu_l} f_{Q,k}(\theta_{\mu_l}; x) + (1 - \cos \theta_0) (\partial_{\mu_l})^2 f_{Q,k}(\theta_{\mu_l}; x).
$$

(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 shall 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 shall 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 twodesign-approximable training modules sandwiching every training 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

\[
\text{GDgrad}^{J_x}_{\mu,l} f_{Q,k} \equiv \sum_{j=1}^{J} c_j \sin(j \epsilon/2) \partial_{\mu,l} f_{Q,k}, \tag{A8}
\]

\[
\text{GDHess}^{J_x}_{\mu,l,\mu',l'} f_{Q,k} \equiv \sum_{j=1}^{J} c_j [\sin(j \epsilon/2)]^2 \partial_{\mu,l} \partial_{\mu',l'} f_{Q,k}. \tag{A9}
\]

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

\[
\text{MSE}_{\text{GD}}(\cdot) = c^T (M \cdot) c, \quad M > 0,
\]

\[
M_{\partial f_{Q,k}} = \frac{4Jd}{N_T(d+1)\epsilon^2} A_{\partial f_{Q,k}} + \frac{d^2 v_1 v_1^T}{2(d+1)(d^2 - 1)},
\]

\[
M_{\partial \partial f_{Q,k}} = \frac{(2J+1)d}{N_T(d+1)\epsilon^4} A_{\partial \partial f_{Q,k}} + \frac{d^2 v_2 v_2^T}{2(d+1)(d^2 - 1)},
\]

\[
M_{\partial \partial \partial f_{Q,k}} = \frac{16Jd}{N_T(d+1)\epsilon^4} A_{\partial \partial \partial f_{Q,k}} + \frac{d^4 v_2 v_2^T}{4(d+1)(d^2 - 1)^2},
\]

where

\[
v_1 = I - \begin{pmatrix}
\sin(\epsilon/2) \\
\sin(\epsilon) \\
\vdots \\
\sin(J \epsilon/2)
\end{pmatrix}, \quad v_2 = I - \begin{pmatrix}
[\sin(\epsilon/2)]^2 \\
[\sin(\epsilon)]^2 \\
\vdots \\
[\sin(J \epsilon/2)]^2
\end{pmatrix},
\]

\[
A_{\partial f_{Q,k}} = \begin{pmatrix}
1 & 0 & \cdots & 0 \\
0 & 1/\epsilon^2 & 0 & \cdots \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 1/\epsilon^2
\end{pmatrix}, \quad A_{\partial \partial f_{Q,k}} = A_{\partial f_{Q,k}}^2,
\]

\[
A_{\partial \partial \partial f_{Q,k}} = 2 \begin{pmatrix}
1 & 0 & \cdots & 0 \\
0 & 1/\epsilon^2 & 0 & \cdots \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 1/\epsilon^2
\end{pmatrix} + \begin{pmatrix}
1 \\
0 \\
\vdots \\
0
\end{pmatrix} \begin{pmatrix}
1/\epsilon^2 \\
0 \\
\vdots \\
0
\end{pmatrix} \cdot \begin{pmatrix}
1/\epsilon^2 \\
0 \\
\vdots \\
0
\end{pmatrix}. \tag{A11}
\]

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 \(\epsilon\) and normalized \(c\) that minimize the relevant operational MSE \(\text{MSE}_{\text{GD}}\) listed in (A10).

Appendix B: Quantum-circuit averages \(\langle \cdot \rangle\)

1. Inner-product average of translated \(f_{Q,k}\)

As a warm-up 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 \(\theta_0\) and \(\theta_0'\) on the same randomized parameter \(\theta_{\mu,l}\). From Eq. (A7),

\[
\langle f_{Q,k}(\theta_{\mu,l} + \theta_0; \mathbf{x}) f_{Q,k}(\theta_{\mu,l} + \theta_0'; \mathbf{x}) \rangle = \langle f_{Q,k} \rangle + (2 - \cos \theta_0 - \cos \theta_0') \langle f_{Q,k}(\partial_{\mu,l})^2 f_{Q,k} \rangle + \sin \theta_0 \sin \theta_0' \langle (\partial_{\mu,l} f_{Q,k})^2 \rangle + (1 - \cos \theta_0)(1 - \cos \theta_0') \langle ([\partial_{\mu,l}]^2 f_{Q,k})^2 \rangle, \tag{B1}
\]

where we shall 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

\[
\langle f_{Q,k} \partial_{\mu,l} f_{Q,k} \rangle = \frac{i}{2(d+1)} \langle 0 | B^{(2)\otimes 2} W_l^{(2)\otimes 2} \otimes \sigma_{\mu} W_l^{(2)\otimes 2} B^{\otimes 2} | 0 \rangle + c.c. \tag{B2}
\]

Since \(W_l^{(2)\otimes 2} \otimes \sigma_{\mu} W_l^{(2)\otimes 2}\) 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 \(\langle f_{Q,k} \partial_{\mu,l} f_{Q,k} \rangle_{1,1} = 0\) for both cases.

\[
\langle f_{Q,k} \partial_{\mu,l} f_{Q,k} \rangle \text{ for Case III}
\]

For this case, we define the operators

\[
\sigma' = W_L^{(1)\otimes 1} V_{L}^{(1)} Q_{L} V_{L}^{(1)}, \tag{B3}
\]

\[
Q_{III} = W_L^{(2)\otimes 2} \sigma' \otimes \sigma_{\mu} W_L^{(2)\otimes 2}. \tag{B4}
\]

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

\[
\text{tr}\{V_{L-1}^{(1)\otimes 2} Q_{III} V_{L-1}^{(1)}\} = \text{tr}\{\sigma' \otimes \sigma_{\mu} \sigma'\} = 0,
\]

\[
\text{tr}\{V_{L-1}^{(1)\otimes 2} Q_{III} V_{L-1}^{(1)}\} = \text{tr}\{\sigma' \sigma_{\mu} \sigma'\} = \text{tr}\{\sigma_{\mu}\} = 0 \tag{B5}
\]

become apparent, giving us \(\langle f_{Q,k} \partial_{\mu,l} f_{Q,k} \rangle_{11} = 0\).
By repeating the above calculations, we also find that \( \langle (\partial_{\mu,L} f_{Q,k}) | (\partial_{\mu,L} f_{Q,k}) \rangle_{1,II,III} = 0 \). 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 \( \langle f_{Q,k} (\partial_{\mu,L}) f_{Q,k} \rangle, \langle (\partial_{\mu,L} f_{Q,k})^2 \rangle \) and \( \langle (\partial_{\mu,L} f_{Q,k})^2 \rangle \). First, the average of

\[
f_{Q,k} (\partial_{\mu,L})^2 f_{Q,k} = \frac{1}{2} \langle 0 | B^{\otimes 2} W_l^{(2)} \otimes 1 \otimes \sigma_{\mu L} W_l^{(2)} \otimes V_l^{\otimes 2} \rangle - f_{Q,k}^2 \frac{2}{d(d+1)},
\]

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

\[
\langle f_{Q,k} (\partial_{\mu,L})^2 f_{Q,k} \rangle \text{ for Case I}
\]

With (A4),

\[
\langle f_{Q,k} (\partial_{\mu,L})^2 f_{Q,k} \rangle_1 = \frac{d}{2(d^2 - 1)} \left( \langle 0 | B^{\dagger} W_l^{(2)} \sigma_{\mu L} W_l^{(2)} B | 0 \rangle - \frac{d}{2(d^2 - 1)} \right),
\]

or \( \langle f_{Q,k} (\partial_{\mu,L})^2 f_{Q,k} \rangle_1 = -\frac{d^2}{2(d+1)(d^2 - 1)} \). (B8)

\[
\langle f_{Q,k} (\partial_{\mu,L})^2 f_{Q,k} \rangle \text{ for Case II}
\]

We note that

\[
\gamma_{II} = \langle 0 | W_l^{(2)} \sigma_{\mu L} W_l^{(2)} | 0 \rangle \leq 1,
\]

so that

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

\[
\langle f_{Q,k} (\partial_{\mu,L})^2 f_{Q,k} \rangle \text{ for Case III}
\]

For this case, consider the operator

\[
Q'_{III} = W_L^{(2)} \sigma_{\mu L} W_L^{(2)} \otimes \sigma_{\mu L} W_L^{(2)} \otimes V_{L-1}^{\otimes 2}.
\]

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

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

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 \( \langle (\partial_{\mu,L} f_{Q,k})^2 \rangle \) in the following subsections, and simply list the final answers:

\[
\langle f_{Q,k} (\theta_{\mu L} + \theta_0; \mathbf{x}) f_{Q,k} (\theta_{\mu L} + \theta'_0; \mathbf{x}) \rangle = \frac{d^2 \left[ \cos \left( \frac{\theta_{\mu L} - \theta'_0}{2} \right) \right]^2 - 1}{(d+1)(d^2 - 1)}
\]

for Case I,

\[
\frac{d(1 - \gamma_{II}) \left[ \cos \left( \frac{\theta_{\mu L} - \theta'_0}{2} \right) \right]^2 + d \gamma_{II} - 1}{d^2 - 1}
\]

for Case II,

\[
\frac{[d - \gamma_{III}(\mathbf{x})] \left[ \cos \left( \frac{\theta_{\mu L} - \theta'_0}{2} \right) \right]^2 + \gamma_{III}(\mathbf{x})}{d(d+1)}
\]

for Case III. (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.

\[
\langle \partial_{\mu,L} f_{Q,k} \rangle \text{ for Cases I and II}
\]

Again, as \( \langle W_L^{(2)} V_L^{(2)} O_k V_L^{(2)} W_L^{(2)} \rangle_{\text{Haar}} = 0 \) for a Pauli \( O_k \),

\[
\langle \partial_{\mu,L} f_{Q,k} \rangle_{1,II} = 0. \quad \text{(B15)}
\]

\[
\langle \partial_{\mu,L} f_{Q,k} \rangle \text{ for Case III}
\]

We inspect the operator

\[
Q''_{III} = W_L^{(2)} \sigma_{\mu L} W_L^{(2)} \otimes V_{L-1}^{(2)} O_k V_L W_L^{(2)}.
\]

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

\[
\langle \partial_{\mu,L} f_{Q,k} \rangle_{III} = \frac{1}{2d} \left( \text{tr}\{\sigma_{\mu L} \sigma'_{\mu L}\} - \frac{1}{2d} \langle \text{tr}\{\sigma_{\mu L} \sigma'_{\mu L}\} \rangle \right) = 0. \quad \text{(B17)}
\]

Now, for the latter:

\[
\langle (\partial_{\mu,L} f_{Q,k})^2 \rangle \text{ for Case I}
\]
Averaging over \( W_L \) yields

\[
\langle [\partial_{\mu_L} f_{Q, k}]^2 \rangle = - \frac{d}{4(d^2 - 1)} \langle 0 | B^{\dagger \otimes 2} W_i^{(2)} \otimes 2 \sigma_{\mu L}^2 W_i^{(2)} \otimes 2 B^{\otimes 2} | 0 \rangle + \frac{d}{4(d^2 - 1)} + \text{c.c.},
\]

or

\[
\langle [\partial_{\mu_L} f_{Q, k}]^2 \rangle_1 = \frac{d^2}{2(d + 1)(d^2 - 1)}. \tag{B19}
\]

\( \langle [\partial_{\mu_L} f_{Q, k}]^2 \rangle \) for Case II

From (B9), we simply obtain

\[
\langle [\partial_{\mu_L} f_{Q, k}]^2 \rangle_2 = \frac{d(1 - \gamma_{11})}{2(d^2 - 1)} \leq \frac{d}{2(d^2 - 1)}. \tag{B20}
\]

\( \langle [\partial_{\mu_L} f_{Q, k}]^2 \rangle \) for Case III

For this case, properties of the operators

\[
Q^{m}_{\text{III}(a)} = W_L^{(2)} \otimes 2 \sigma_{\mu L}^2 \sigma_{\tau}^2 W_L^{(2)} \otimes 2,
\]

\[
Q^{m}_{\text{III}(b)} = W_L^{(2)} \otimes 2 1 \otimes \sigma_{\mu L} \sigma_{\tau'}^2 \sigma_{\mu L} \otimes 1 W_L^{(2)} \otimes 2 \tag{B21}
\]

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

\[
\text{tr}\{Q^{m}_{\text{III}(a)}\} = \text{tr}\{\sigma_{\mu L} \sigma' \}^2 = \text{tr}\{Q^{m}_{\text{III}(b)}\}. \tag{B22}
\]

For the trace properties with the swap operator, they are

\[
\text{tr}\{V_{L-1}^{\dagger \otimes 2} Q^{m}_{\text{III}(a)} V_{L-1}^{\otimes 2} \} = \text{tr}\{\sigma_{\mu L} \sigma' \sigma_{\mu L} \sigma' \} = \gamma_{11}(x),
\]

\[
\text{tr}\{V_{L-1}^{\dagger \otimes 2} Q^{m}_{\text{III}(b)} V_{L-1}^{\otimes 2} \} = \text{tr}\{\sigma' \sigma_{\mu L} \sigma' \sigma_{\mu L} \} = d. \tag{B23}
\]

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

\[
\gamma_{11}(x) = \langle \text{tr}\{[\sigma_{\mu L} \sigma']^2\} \rangle,
\]

\[
\gamma'_{11}(x) = \langle \text{tr}\{[\sigma_{\mu L} \sigma']^2\} \rangle,
\]

\[
\langle [\partial_{\mu_L, l} f_{Q, k}]^2 \rangle_{\text{III}} = - \frac{\gamma_{11}(x) + \gamma_{11}(x) + d}{2d(d + 1)} + \frac{\gamma_{11}(x) + d}{2d(d + 1)}
\]

\[
= \frac{d - \gamma_{11}(x)}{2d(d + 1)} \leq \frac{1}{d + 1}. \tag{B24}
\]

The final inequality is of the Cauchy–Schwarz type

\[
\text{tr}\{[\sigma_{\mu L} \sigma']^2\}^2 \leq \text{tr}\{\sigma_{\mu L}^2\} \text{tr}\{[\sigma' \sigma_{\mu L}]^2\} = d^2, \tag{B25}
\]

or \(-d \leq \gamma_{11}(x) \leq d\).

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 \( \langle \partial \partial f_{Q, k}\rangle_{1, II, III} = 0 \). For the squared averages, since

\[
\begin{aligned}
\langle [\partial_{\mu_L, l} f_{Q, k}]^2 \rangle &= \frac{1}{4} \left[ f_{Q, k}^2 + D_1 - 2 D_{II} \right], \\
D_1 &= \langle 0 | B^{\dagger \otimes 2} W_i^{(2)} \otimes 2 \sigma_{\mu L}^2 W_i^{(2)} \otimes 2 B^{\otimes 2} | 0 \rangle, \\
D_{II} &= \langle 0 | B^{\dagger \otimes 2} W_i^{(2)} \otimes 2 1 \otimes \sigma_{\mu L} W_i^{(2)} \otimes 2 V_i^{\otimes 2} | 0 \rangle,
\end{aligned}
\]

we need the averages of \( D_1 \) and \( D_{II} \) in all three cases.

\( \langle [\partial_{\mu_L, l} f_{Q, k}]^2 \rangle \) for Case I

That \( \langle D_1 \rangle = 1/(d + 1) \) is immediate. On the other hand,

\[
\begin{aligned}
\langle D_{II} \rangle &= \frac{d}{d^2 - 1} \langle 0 | B^{\dagger \otimes 2} W_i^{(2)} \otimes 2 \sigma_{\mu L}^2 W_i^{(2)} \otimes 2 B^{\otimes 2} | 0 \rangle \\
&= \frac{1}{d^2 - 1} - \frac{1}{(d + 1)(d^2 - 1)}. \tag{B27}
\end{aligned}
\]

Altogether,

\[
\langle [\partial_{\mu_L, l} f_{Q, k}]^2 \rangle_1 = \frac{d^2}{2(d + 1)(d^2 - 1)}. \tag{B28}
\]

\( \langle [\partial_{\mu_L, l} f_{Q, k}]^2 \rangle \) for Case II

If \( l = 1 \), then \( \langle D_1 \rangle = 1/(d + 1) \) still, but

\[
\langle D_{II} \rangle = \frac{d - d_{II}}{d^2 - 1}, \tag{B29}
\]

so that

\[
\langle [\partial_{\mu_L, l} f_{Q, k}]^2 \rangle_1 \leq \frac{1}{d + 1}. \tag{B30}
\]

where \( d_{II} \) is as defined in (B9).

\( \langle [\partial_{\mu_L, l} f_{Q, k}]^2 \rangle \) for Case III

From (B11) and the operator definition

\[
Q^{m}_{\text{III}} = W_L^{(2)} \otimes 2 \sigma_{\mu L}^2 \sigma_{\tau}^2 \sigma_{\mu L} W_L^{(2)} \otimes 2, \tag{B31}
\]

whose trace properties include \( \text{tr}\{Q^{m}_{\text{III}}\} = 0 \) and

\[
\text{tr}\{V_{L-1}^{\dagger \otimes 2} Q^{m}_{\text{III}} V_{L-1}^{\otimes 2} \} = \text{tr}\{\sigma_{\mu L} \sigma' \sigma_{\mu L} \} = d. \tag{B32}
\]
With these,
\[
\langle D_1 \rangle = \langle 0 | B^{\otimes 2} Q''_{II} B^{\otimes 2} | 0 \rangle = \frac{1}{d + 1},
\]
\[
\langle D_{II} \rangle = \langle 0 | B^{\otimes 2} Q''_{II} B^{\otimes 2} | 0 \rangle = \frac{\gamma_{II}(x)}{d(d + 1)}, \tag{B33}
\]
which finally brings us to
\[
\langle (\partial_{\mu,L})^2 f_{Q,k} |^2 \rangle_{III} = \frac{d - \gamma_{II}(x)}{2d(d + 1)} \leq \frac{1}{d + 1}. \tag{B34}
\]

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:
\[
\partial_{\mu,L} \partial_{\mu',L'} f_{Q,k} = [(1) + (1)_{c.c.} + (2) + (2)_{c.c.}],
\]
\[
(1) = -\frac{1}{4} \langle 0 | C \otimes W_{I}^{(2)} \otimes \sigma_{\mu',L'} W_{I}^{(1)} \otimes W_{I}^{(2)} \otimes \sigma_{\mu,L} W_{I}^{(1)} \otimes V_{I}^{(1)} \otimes V_{I}^{(2)} A^\dagger O_{k} A W_{I} B_{I} V_{I} B_{I} C | 0 \rangle,
\]
\[
(2) = \frac{1}{4} \langle 0 | C \otimes W_{I}^{(1)} \otimes W_{I}^{(2)} \otimes \sigma_{\mu',L'} W_{I}^{(1)} \otimes V_{I}^{(1)} \otimes A^\dagger O_{k} A W_{I} B_{I} V_{I} B_{I} C | 0 \rangle. \tag{B35}
\]

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

\[\langle |\partial_{\mu,L} \partial_{\mu',L'} f_{Q,k}|^2 \rangle \text{ for Cases I and IV} \]

Let \( \tilde{\sigma}_{\mu} = W_{I}^{(1)} V_{I}^{(1)} B_{I} W_{I}^{(2)} \otimes \sigma_{\mu,L} W_{I}^{(2)} B_{I} W_{I}^{(1)} \). Then upon denoting \( \bar{\gamma} = \text{tr}\{ (\sigma_{\mu',L'} \tilde{\sigma}_{\mu})^2 \} \) and \( \gamma' = \text{tr}\{ (\sigma_{\mu',L'} \tilde{\sigma}_{\mu}) \} \), we get
\[
\langle (1)^2 + (1)_{c.c.} \rangle = \langle (2)^2 + (2)_{c.c.} \rangle = \frac{1}{8d(d + 1)} \left( \bar{\gamma} + \gamma' \right),
\]
\[
2 \langle (1) \times (1)_{c.c.} \rangle = \frac{d^2 + d - 1}{8(d + 1)(d^2 - 1)} - \frac{(\bar{\gamma})}{8d(d + 1)(d^2 - 1)},
\]
\[
2 \langle (1) \times (2) \rangle = \frac{d^2 + d - 1}{8(d + 1)(d^2 - 1)} - \frac{(\gamma') - (d \gamma')}{8d(d + 1)(d^2 - 1)},
\]
\[
2 \langle (1) \times (2) \rangle = \frac{d^2 + d - 1}{8(d + 1)(d^2 - 1)} - \frac{(\bar{\gamma} + \gamma')}{8d(d + 1)(d^2 - 1)},
\]
\[
2 \langle (2) \times (2) \rangle = 2 \langle (1) \times (1)_{c.c.} \rangle. \tag{B36}
\]

These amount to
\[
\langle |\partial_{\mu,L} \partial_{\mu',L'} f_{Q,k}|^2 \rangle_{IV} = \frac{d^2 + (\gamma')}{4(d + 1)(d^2 - 1)} \leq \frac{d^2}{2(d + 1)(d^2 - 1)}. \tag{B37}
\]

In Case I, \( (\gamma') = d^2/(d^2 - 1) \), so that
\[
\langle |\partial_{\mu,L} \partial_{\mu',L'} f_{Q,k}|^2 \rangle_{I} = \frac{d^4}{4(d + 1)(d^2 - 1)^2}. \tag{B38}
\]

\[\langle |\partial_{\mu,L} \partial_{\mu',L'} f_{Q,k}|^2 \rangle \text{ for Cases II and V} \]

Let us define the shorthand notations \( \alpha = \langle 0 | W_{I}^{(2)} \sigma_{\mu',L'} \tilde{\sigma}_{\mu,L} B_{I} W_{I}^{(2)} | 0 \rangle, \beta = \langle 0 | W_{I}^{(2)} \otimes \sigma_{\mu',L'} \tilde{\sigma}_{\mu,L} \otimes \sigma_{\mu',L'} \tilde{\sigma}_{\mu,L} W_{I}^{(2)} | 0 \rangle \), and \( \gamma = \langle 0 | W_{I}^{(2)} \sigma_{\mu',L'} W_{I}^{(2)} | 0 \rangle \). Then,
\[
\langle (1)^2 + (1)_{c.c.} \rangle = \langle (2)^2 + (2)_{c.c.} \rangle = \frac{\text{Re}\{ (\alpha^2) \}}{8(d + 1)^2},
\]
\[
2 \langle (1) \times (1)_{c.c.} \rangle = \frac{d}{8(d - 1)} - \langle (\alpha^2) \rangle,
\]
\[
2 \langle (1) \times (2) \rangle = \frac{d}{8(d - 1)} - \langle (\alpha^2) \rangle,
\]
\[
2 \langle (1) \times (2) \rangle = \frac{d}{8(d - 1)} - \langle (\alpha^2) \rangle,
\]
\[
2 \langle (2) \times (2) \rangle = 2 \langle (1) \times (1)_{c.c.} \rangle, \tag{B39}
\]

such that
\[
\langle |\partial_{\mu,L} \partial_{\mu',L'} f_{Q,k}|^2 \rangle_{V} = \frac{d(1 - \text{Re}\{ (\alpha^2) \})}{4(d^2 - 1)} \leq \frac{3d}{4(d^2 - 1)}. \tag{B40}
\]

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 \( \langle \beta \rangle = \langle \gamma^2 \rangle / (d + 1) \), yielding
\[
\langle |\partial_{\mu,L} \partial_{\mu',L'} f_{Q,k}|^2 \rangle_{II} = \frac{d^3(1 - \gamma^2)}{4(d^2 - 1)^2} \leq \frac{d^3}{4(d^2 - 1)^2}. \tag{B41}
\]

\[\langle |\partial_{\mu,L} \partial_{\mu',L'} f_{Q,k}|^2 \rangle \text{ for Cases III and VII} \]

For these two cases, let us focus on the Pauli operators
\[
\sigma_{0} = W_{I}^{(1)} V_{I}^{(1)} A^\dagger O_{k} A V_{I} W_{I}^{(1)} ,
\]
\[
\sigma_{1} = V_{I}^{(1)} B_{I} W_{I}^{(2)} \otimes \sigma_{\mu,L} W_{I}^{(2)} B_{I} V_{I} ,
\]
\[
\sigma_{2} = V_{I}^{(1)} B_{I} W_{I}^{(2)} \otimes \sigma_{\mu,L} W_{I}^{(2)} B_{I} V_{I}. \tag{B42}
\]

In terms of these Pauli operators and the rotated versions \( \tilde{\sigma}_{j} = W_{I}^{(1)} \sigma_{j} W_{I}^{(1)} \), we define a new set of parameters:
\( \alpha' = \text{tr}\{ (\sigma_{\mu,L} \tilde{\sigma}_{2} \tilde{\sigma}_{1})^2 \}, \beta' = \text{tr}\{ (\sigma_{\mu,L} \tilde{\sigma}_{2})^2 \} \). These lead to
$$\langle (1)^2 + (1)^2_{c.c.} \rangle = \frac{\Re \left\{ \langle 2 \rangle \cdot (2) \cdot (2)_{c.c.} \right\}}{8d(d+1)} ,$$

$$2 \langle (1) \times (1)_{c.c.} \rangle = \frac{\langle (1)^2 \rangle + d}{8d(d+1)} ,$$

$$2 \langle (1) \times (2) \rangle = \frac{-\langle (1)^2 \rangle + \langle 2 \rangle \cdot (1) \cdot (1)_{c.c.} \rangle + \langle 2 \rangle \cdot (2)_{c.c.} \rangle + 2 \langle (1) \times (1) \cdot (2)_{c.c.} \rangle ,$$

$$\langle |x|^2 \rangle = \Re \left\{ \langle 2 \rangle \cdot (2) \cdot (2)_{c.c.} \rangle \right\} ,$$

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

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

$$\langle |\partial_{\mu,i} \partial_{\mu',i'} f_{Q,k} |^2 \rangle \leq \frac{1}{2d(d+1)} .$$

**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 distribution**

**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 $\{|v_{kl}\rangle\}_{l=0}^{d-1}$ of $O_k$, where relative frequencies

$v_{kl}(\theta; x) = n_{kl}(\theta; x)/N \rightarrow p_{kl}(\theta; x)$ are recorded. Explicitly, these relative frequencies satisfy the following basic sta-

tistical identities:

$$E[v_{kl}(\theta; x)] = p_{kl}(\theta; x),$$

$$E[v_{kl}(\theta; x)v_{kl}(\theta'; x')] = (1-\delta_{\theta,\theta'}\delta_{x,x'})p_{kl}(\theta; x)p_{kl}(\theta'; x') + \frac{\delta_{\theta,\theta'}\delta_{x,x'}}{N}p_{kl}(\theta; x) + (N-1)p_{kl}(\theta; x)p_{kl}(\theta; x),$$

$$E[v_{kl}(\theta; x)v_{kl'\neq k}(\theta'; x')] = p_{kl}(\theta; x)p_{k'l'}(\theta'; x').$$

(C1)

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 esti-

mator in (4). Together with the multinomial identities in (C1),

this leads to

$$\text{MSE}(f_Q) = \frac{1}{N_T} \left(1 - \langle f_Q^2 \rangle \right),$$

(C2)

where $N_T = N$ is the total number of copies per basis ob-

servable needed to complete one function estimation evalu-

ated at the circuit parameters $\theta$. For two-design appro-

achable PEPQCs, by making use of Eq. (A4), we quickly find that

$$\langle f_{Q,k}(\theta_{ml}; x)^2 \rangle = \langle f_{Q,k}(\theta_{ml} + \theta_0; x)^2 \rangle = \frac{1}{d+1}$$

for any $\theta_0$ (see Appendix B1), so that

$$\text{MSE}(f_Q) = \frac{d}{N_T(d+1)}.$$ 

(C3)

3. Optimally-tuned FD estimators for general cases

In their most general forms, the FD MSES, which are line-

ar combinations of finite-copy and nonzero-$\epsilon$ approximation

squared-errors, read

$$\text{MSE}_{FD}(\partial f_Q) = \frac{4d}{N_T(d+1)^2} + \langle (\partial f_Q)^2 \rangle g_1(\epsilon/2),$$

$$\text{MSE}_{FD}(\partial \partial f_Q) = \frac{18d^2}{N_T(d+1)^2} + \langle (\partial \partial f_Q)^2 \rangle g_2(\epsilon/2),$$

$$\text{MSE}_{FD}(\partial \partial \partial f_Q) = \frac{16d}{N_T(d+1)^2} + \langle (\partial \partial \partial f_Q)^2 \rangle g_2(\epsilon/2),$$

with $g_1(\epsilon/2) = \left[1 - \text{sinc}(\epsilon/2)^2\right]^2$, $g_2(\epsilon/2) = \left[1 - \text{sinc}(\epsilon/2)^2\right]^2$, and all the three average terms $\langle (\partial f_Q)^2 \rangle$, $\langle (\partial \partial f_Q)^2 \rangle$, and $\langle (\partial \partial \partial f_Q)^2 \rangle$ have been rigorously worked out in Appendices B2 through B4.

Keeping in their arbitrary forms, we can proceed to mini-

mize all these MSES and derive optimal FD estimators in the

regime $\epsilon \ll 1$. Using the Taylor approximations $g_1(\epsilon/2) \approx \epsilon^3/576$ and $g_2(\epsilon/2) \approx \epsilon^4/144$, we identify two different functional structures

$$\text{MSE}_{FD}(\partial f_Q) \approx \frac{A_1}{\epsilon^2} + \frac{\epsilon^4}{576} B_1,$$

$$\text{MSE}_{FD}(\partial \partial f_Q), \text{MSE}_{FD}(\partial \partial \partial f_Q) \approx \frac{A_2}{\epsilon^4} + \frac{\epsilon^4}{144} B_2,$$

(C6)

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

Now, optimizing over $\epsilon$ for both kinds of structures,

$$\min_{\epsilon} \left\{ \frac{A_1}{\epsilon^2} + \frac{\epsilon^4}{576} B_1 \right\}^{1/3} \text{with } \epsilon = \left( \frac{288A_1}{B_1} \right)^{1/6},$$

$$\min_{\epsilon} \left\{ \frac{A_2}{\epsilon^4} + \frac{\epsilon^4}{144} B_2 \right\}^{1/3} \text{with } \epsilon = \left( \frac{144A_2}{B_2} \right)^{1/8}.$$ 

(C7)

These give the complete set of approximate $\epsilon$-ops for all types of sampling inasmuch as

$$\epsilon_{\text{opt}}(\partial f_Q) \approx \left[ \frac{1152d}{N_T(d+1)} \right]^{1/6},$$

$$\epsilon_{\text{opt}}(\partial \partial f_Q) \approx \left[ \frac{2592d}{N_T(d+1)} \right]^{1/8},$$

$$\epsilon_{\text{opt}}(\partial \partial \partial f_Q) \approx \left[ \frac{2304d}{N_T(d+1)} \right]^{1/8},$$

(C8)

along with the optimized MSES:

$$\text{MSE}_{FD,\text{opt}}(\partial f_Q) = \left[ \frac{3d^2}{16N_T(d+1)^2} \right]^{1/3},$$

$$\text{MSE}_{FD,\text{opt}}(\partial \partial f_Q) = \sqrt{\frac{d}{N_T(d+1)^2}},$$

$$\text{MSE}_{FD,\text{opt}}(\partial \partial \partial f_Q) = \frac{2}{3} \sqrt{\frac{d}{N_T(d+1)}}.$$ 

(C9)

The final task is then to substitute the correct expressions of

$\langle (\partial f_{Q,k})^2 \rangle, \langle (\partial \partial f_{Q,k})^2 \rangle$ and $\langle (\partial \partial \partial f_{Q,k})^2 \rangle$ that are applicable to the relevant case in point as listed in Fig. 7. Enjoying the fruits of our labor in Appendices B2 through B4, summarized in Tab. 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 minimizes 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 esti-

mators for any case require substitutions of the answers from

Tab. I, which may be done if so desired. For benchmark-

ing with the PS strategy, however, all one needs is to recog-

nize that Tab. I implies that $\langle (\partial f_{Q,k})^2 \rangle, \langle (\partial \partial f_{Q,k})^2 \rangle, \langle (\partial \partial \partial f_{Q,k})^2 \rangle \leq$
\( O(1/d) \) for large \( d \), so that

\[
\text{MSE}_{\text{FD, opt}}(\partial f_Q) \lesssim O\left( \frac{1}{N_T^{2/3}d^{1/3}} \right),
\]

\[
\text{MSE}_{\text{FD, opt}}(\partial \partial f_Q), \text{MSE}_{\text{FD, opt}}(\partial \partial f_Q') \lesssim O\left( \frac{1}{N_T^{1/2}d^{1/2}} \right),
\]

(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 \( \text{MSE}_{\text{GD}} = e^\top M e \) in their MSEs in (A10) and (A11). The only additional step one needs to perform is the minimization over normalized \( e \). To do this, we first introduce the Lagrange function

\[
D = e^\top M e - 2 \lambda (e^\top I - 1)
\]

(C11)

that is to be optimized, where \( \lambda \) is the Lagrange multiplier that takes care of the normalization constraint, and the factor of 2 in front of \( \lambda \) is introduced for convenience that shall become clear very soon. An arbitrary variation of \( D \) over \( e \) gives

\[
\delta D = \delta e^\top M e + e^\top M \delta e - \lambda (\delta e^\top I + I^\top \delta e).
\]

(C12)

A minimization of \( D \), akin to the constrained minimization of \( \text{MSE}_{\text{GD}} \), is done by setting \( \delta D = 0 \), resulting in \( M e = \lambda I \). As \( M \) is invertible, solving \( \lambda \) yields the extremal equation

\[
e = \frac{M^{-1} I}{I^\top M^{-1} I}.
\]

(C13)

In other words, the optimal \( \text{MSE}_{\text{GD}} \) may be obtained by first substituting \( e \) \( \text{MSE}_{\text{GD}} \) with the optimal one defined in (C13), and, next, minimizing the result over \( \epsilon \).

Just like for FD estimators, this second minimization over \( \epsilon \) 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 \( \text{MSE}_{\text{GD}} \). Thankfully, the Cauchy–Schwarz inequality remains our dearest friend for this task, awarding us with \((1^\top I)^2 \leq 1^\top M I 1^\top M^{-1} I \), and consequently \( \text{MSE}_{\text{GD, opt}} \leq J^{-2} \min_{\epsilon>0} 1^\top M I \).

We now take advantage of the fact that the \( J \)-dimensional Ms, which are more precisely listed in (A10), give rise to \( M = 1^\top M I \) that are multidimensional analogs of the right-hand sides in (C5), namely

\[
M_{\partial f_Q} = \frac{4JdH_{J,2}}{N_T(d+1)\epsilon^2} + \epsilon^4 \langle (\partial f_{Q,k})^2 \rangle J^2(J+1)^2(2J+1)^2,
\]

\[
M_{\partial \partial f_Q} = \frac{(2J+1)d}{N_T(d+1)\epsilon^2}(4H_{J,2}^2 + 2H_{J,4}) + \epsilon^4 \langle (\partial \partial f_{Q,k})^2 \rangle J^2(J+1)^2(2J+1)^2,
\]

\[
M_{\partial \partial f_Q} = \frac{16JdH_{J,4}}{N_T(d+1)\epsilon^2} + \epsilon^4 \langle (\partial \partial f_{Q,k})^2 \rangle J^2(J+1)^2(2J+1)^2,
\]

(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

\[
\text{MSE}_{\text{GD, opt}}(\partial f_Q) \lesssim \frac{1}{J^2} \left[ \frac{(\langle \partial f \rangle)^2}{192N_T^2(1+d)^2} \right]^{1/3},
\]

\[
\text{MSE}_{\text{GD, opt}}(\partial \partial f_Q) \lesssim \frac{1}{36J^2} \left[ \frac{(\langle \partial \partial f \rangle)^2}{N_T(d+1)} \right],
\]

\[
\text{MSE}_{\text{GD, opt}}(\partial \partial f_Q') \lesssim \frac{1}{9J^2} \left[ \frac{(\langle \partial \partial f' \rangle)^2}{N_T(d+1)} \right],
\]

(C15)

all of which reduce to the corresponding equalities in (C9) for

\[
J = 1 \text{ as } H_{1,k} = 1. \text{ Unsurprisingly, for large } d,
\]

\[
\text{MSE}_{\text{GD, opt}}(\partial f_Q) \lesssim O\left( \frac{1}{N_T^{2/3}d^{1/3}} \right),
\]

\[
\text{MSE}_{\text{GD, opt}}(\partial \partial f_Q), \text{MSE}_{\text{GD, opt}}(\partial \partial f_Q') \lesssim O\left( \frac{1}{N_T^{1/2}d^{1/2}} \right).
\]

(C16)
| k  | $h_k$ | $O_{k,1}$ | $O_{k,2}$ | $O_{k,3}$ | $O_{k,4}$ | $O_{k,5}$ | $O_{k,6}$ | $O_{k,7}$ | $O_{k,8}$ |
|----|-------|----------|----------|----------|----------|----------|----------|----------|----------|
| 1  | -0.180625859 | 1 | 1 | 1 | 1 | 1 | Z | 1 |
| 2  | -0.180625859 | 1 | 1 | 1 | 1 | 1 | Z | 1 |
| 3  | -0.159587991 | 1 | 1 | 1 | 1 | 1 | Z | 1 |
| 4  | -0.159587991 | 1 | 1 | 1 | 1 | 1 | Z | 1 |
| 5  | 0.174193924 | 1 | 1 | Z | 1 | 1 | 1 | 1 |
| 6  | 0.174193924 | 1 | 1 | Z | 1 | 1 | 1 | 1 |
| 7  | 0.227570968 | Z | 1 | 1 | 1 | 1 | 1 | 1 |
| 8  | 0.227570968 | Z | 1 | 1 | 1 | 1 | 1 | 1 |
| 9  | 0.112704888 | 1 | 1 | 1 | 1 | Z | 1 | Z |
| 10 | 0.112704888 | 1 | 1 | 1 | 1 | Z | 1 | Z |
| 11 | 0.119520678 | Z | 1 | 1 | 1 | Z | 1 | 1 |
| 12 | 0.119520678 | Z | 1 | 1 | 1 | Z | 1 | 1 |
| 13 | 0.134100069 | Z | 1 | 1 | 1 | Z | 1 | 1 |
| 14 | 0.134100069 | Z | 1 | 1 | 1 | Z | 1 | 1 |
| 15 | 0.137351346 | Z | 1 | 1 | 1 | 1 | Z | 1 |
| 16 | 0.137351346 | Z | 1 | 1 | 1 | 1 | Z | 1 |
| 17 | 0.137670707 | 1 | 1 | Z | 1 | 1 | 1 | 1 |
| 18 | 0.137670707 | 1 | 1 | Z | 1 | 1 | 1 | 1 |
| 19 | 0.141387333 | 1 | 1 | 1 | Z | 1 | 1 | Z |
| 20 | 0.141387333 | 1 | 1 | 1 | Z | 1 | 1 | Z |
| 21 | 0.147230746 | 1 | 1 | 1 | Z | 1 | 1 | 1 |
| 22 | 0.147230746 | 1 | 1 | 1 | Z | 1 | 1 | 1 |
| 23 | 0.149265817 | 1 | 1 | 1 | Z | 1 | 1 | 1 |
| 24 | 0.149731063 | 1 | 1 | Z | 1 | 1 | 1 | Z |
| 25 | 0.149731063 | 1 | 1 | Z | 1 | 1 | 1 | Z |
| 26 | 0.151377428 | Z | 1 | 1 | 1 | 1 | 1 | Z |
| 27 | 0.151377428 | Z | 1 | 1 | 1 | 1 | 1 | Z |
| 28 | 0.153453599 | 1 | 1 | 1 | Z | 1 | Z | Z |
| 29 | 0.155818816 | 1 | 1 | Z | 1 | 1 | 1 | Z |
| 30 | 0.155818816 | 1 | 1 | Z | 1 | 1 | 1 | Z |
| 31 | 0.167560719 | Z | 1 | 1 | 1 | 1 | 1 | Z |
| 32 | 0.167560719 | Z | 1 | 1 | 1 | 1 | 1 | Z |
| 33 | 0.181433521 | Z | 1 | 1 | 1 | 1 | 1 | Z |
| 34 | 0.181433521 | Z | 1 | 1 | 1 | 1 | 1 | Z |
| 35 | 0.193911464 | Z | 1 | 1 | 1 | 1 | 1 | Z |
| 36 | 0.220039773 | 1 | Z | Z | 1 | 1 | 1 | 1 |
| 37 | 0.220039773 | 1 | Z | Z | 1 | 1 | 1 | 1 |
| 38 | 0.220039773 | 1 | Z | Z | 1 | 1 | 1 | 1 |
| 39 | 0.228682446 | 1 | Z | X | Y | X | X | X |
| 40 | 0.228682446 | 1 | Z | X | Y | X | X | X |
| 41 | 0.228682446 | 1 | Z | X | Y | X | X | X |
| 42 | 0.228682446 | 1 | Z | X | Y | X | X | X |
| 43 | 0.228682446 | 1 | Z | X | Y | X | X | X |
| 44 | 0.228682446 | 1 | Z | X | Y | X | X | X |
| 45 | 0.228682446 | 1 | Z | X | Y | X | X | X |
| 46 | 0.228682446 | 1 | Z | X | Y | X | X | X |
| 47 | 0.228682446 | 1 | Z | X | Y | X | X | X |
| 48 | 0.228682446 | 1 | Z | X | Y | X | X | X |

| $k$ | $h_k$ | $O_{k,1}$ | $O_{k,2}$ | $O_{k,3}$ | $O_{k,4}$ | $O_{k,5}$ | $O_{k,6}$ | $O_{k,7}$ | $O_{k,8}$ |
|----|-------|----------|----------|----------|----------|----------|----------|----------|----------|
| 49 | 0.006087753 | 1 | 1 | Y | X | 1 | 1 | X | Y |
| 50 | 0.006087753 | 1 | 1 | Y | X | 1 | 1 | X | Y |
| 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 |
| 56 | 0.013872802 | X | Y | X | Y | 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 | 0.01736736 | X | Y | 1 | 1 | 1 | 1 | Y | X |
| 60 | 0.01736736 | X | Y | 1 | 1 | 1 | 1 | Y | X |

**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 $\text{sto-3g}$, 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: ($-0.3099$, $-0.0638$, 0.0), O: (1.5780, 0.8540, 0.0), H: (2.7909, $-0.5159$, 0.0). All coefficients are generated using the pennylane.qchem Python package.*

* [Visit https://pennylane.ai/qml/demos/tutorial_quantum_chemistry.html](https://pennylane.ai/qml/demos/tutorial_quantum_chemistry.html)
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 \( \text{MSE}_{\text{SPS}}(\cdot) \) expressions read

\[
\begin{align*}
\text{MSE}_{\text{SPS}}(\partial f_Q) &= \frac{d\lambda^2}{N_T(d+1)} + \langle (\partial f_Q)^2 \rangle (1-\lambda)^2, \\
\text{MSE}_{\text{SPS}}(\partial \partial f_Q) &= \frac{9d\lambda^2}{8N_T(d+1)} + \langle (\partial \partial f_Q)^2 \rangle (1-\lambda)^2, \\
\text{MSE}_{\text{SPS}}(\partial \partial f_Q) &= \frac{d\lambda^2}{N_T(d+1)} + \langle (\partial \partial f_Q)^2 \rangle (1-\lambda)^2, \\
\end{align*}
\]

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

\[
\begin{align*}
\text{MSE}_{\text{SPS, opt}}(\partial Q) &= \frac{d\langle (\partial f_Q)^2 \rangle}{d + (d+1) \langle (\partial f_Q)^2 \rangle N_T}, \\
\text{MSE}_{\text{SPS, opt}}(\partial \partial Q) &= \frac{9d (\partial f_Q)^2}{9d + 8(d+1) \langle (\partial f_Q)^2 \rangle N_T}, \\
\text{MSE}_{\text{SPS, opt}}(\partial \partial f_Q) &= \frac{d\langle (\partial \partial f_Q)^2 \rangle}{d + (d+1) \langle (\partial \partial f_Q)^2 \rangle N_T}. \\
\end{align*}
\]

It is straightforward to verify that \( \text{MSE}_{\text{SPS, opt}}(\partial Q), \text{MSE}_{\text{SPS, opt}}(\partial \partial Q) \) and \( \text{MSE}_{\text{SPS, opt}}(\partial \partial f_Q) \) 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

\[
\begin{align*}
\frac{\partial \text{MSE}_{\text{SPS, opt}}(\partial f_Q)}{\partial \langle (\partial f_Q)^2 \rangle} &= \frac{d^2}{[d + (d+1) \langle (\partial f_Q)^2 \rangle N_T]^2}, \\
\frac{\partial \text{MSE}_{\text{SPS, opt}}(\partial \partial f_Q)}{\partial \langle (\partial f_Q)^2 \rangle} &= \frac{81d^2}{[9d + 8(d+1) \langle (\partial f_Q)^2 \rangle N_T]^2}, \\
\frac{\partial \text{MSE}_{\text{SPS, opt}}(\partial \partial f_Q)}{\partial \langle (\partial f_Q)^2 \rangle} &= \frac{d^2}{[d + (d+1) \langle (\partial f_Q)^2 \rangle N_T]^2},
\end{align*}
\]

are all nonnegative. It then follows from Tab. 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 \( N_T \gg d \), we find that \( \text{MSE}_{\text{SPS, opt}}(\cdot) \rightarrow \text{MSE}_{\text{PS}}(\cdot) \).

Appendix D: Simplified electronic description of a water molecule

In the quantum-eigensolver scenario, the observable \( O = H - h_0 f \), with \( h_0 = \text{tr} \{ H \} / d \), is defined as a trace-subtracted Hamiltonian operator \( H \) that describes the dynamics of electrons in a water molecule. Under the Hartree–Fock approximation [80], every electron in the molecule is treated as an independent particle that experiences both the Coulomb potential in the molecule 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 Tab. II.
[13] A. Kitaev, Fault-tolerant quantum computation by anyons, *Ann. Phys.* **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 Information* **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, MA, 2004) pp. 247–266.

[19] D. Aharonov and M. Ben-Or, Fault-Tolerant Quantum Computation with Constant Error Rate, *SIAM Journal on Computing* **38**, 1207 (2008).

[20] E. Knill, Approximation by quantum circuits (1995), arXiv:quant-ph/9508006 [quant-ph].

[21] J. Preskill, Quantum Computing in the NISQ era and beyond, *Quantum* **2**, 78 (2018).

[22] T. R. Bromley, J. M. Arrazola, S. Jahangiri, J. Izac, N. Quesada, A. D. Gran, M. Schulz, 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, *Chemical Physics Letters* **219**, 343 (1994).

[25] T. Kadokawa 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, NY, USA, 2011) p. 333–342.

[27] S. Aaronson, A linear-optical proof that the permanent is #P-hard, *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **467**, 4183–4193 (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 Physics* **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**, L304041 (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 Reviews Physics* **3**, 625 (2021).

[33] Y. Cao, J. Romero, J. P. Olson, M. Degroote, P. D. Johnson, M. Kiefertövá, 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, *Chemical Reviews* **119**, 10856 (2019), pMID: 31469277.

[34] S. Endo, Z. Cai, S. C. Benjamin, and X. Yuan, Hybrid quantum-classical algorithms and quantum error mitigation, *Journal of the Physical Society of Japan* **90**, 032001 (2021).

[35] S. McArdrle, 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 Communications* **5**, 4213 (2014).

[37] D. Wecker, M. B. Hastings, and M. Troyer, Progress towards practical quantum variational algorithms, *Phys. Rev. X* **5**, 303 (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 (2014), arXiv:arXiv:1411.4028 [quant-ph].

[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. Schulz, J. Sinayskiy, and F. Petruccione, An introduction to quantum machine learning, *Contemporary Physics* **56**, 172 (2015).

[42] M. Schulz and N. Killoran, Quantum machine learning in feature hilbert spaces, *Phys. Rev. Lett.* **122**, 040504 (2019).

[43] G. Carleo, I. Cirac, K. Cramer, L. Daudet, M. Schulz, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, *Rev. Mod. Phys.* **91**, 045002 (2019).

[44] P. Date, Quantum discriminator for binary classification (2020), arXiv: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 (2021).

[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. Haridl, 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. Haridlová, 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*, Vol. 2 (1998) pp. 1213–1216 vol.2.

[53] S. Amari, Natural Gradient Works Efficiently in Learning, *Neural Computation* **10**, 251 (1998).

[54] B. Koczo and S. C. Benjamin, Quantum natural gradient generalised to non-unitary circuits (2019).
[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. Research 2, 043246 (2020).
[57] S. E. Smart and D. A. Mazzotti, 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 Communications 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 Communications 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. Research 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 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. Applied 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, Communications in Mathematical Physics 291, 257 (2009).
[73] Z. Puchała and J. Miszczak, Symbolic integration with respect to the haar measure on the unitary groups, Bulletin of the Polish Academy of Sciences: Technical Sciences 65, 21 (2017).
[74] P. J. Olver, Introduction to Partial Differential Equations (Springer Science & Business Media, Switzerland, 2014).
[75] G. G. Guerreschi and M. Smelyanskiy, Practical optimization for hybrid quantum-classical algorithms (2017).
[76] B. Swartz and B. Wendroff, Generalized finite-difference schemes, Mathematics of Computation 23, 37 (1969).
[77] K. Zhang, M.-H. Hsieh, L. Liu, and D. Tao, Toward trainability of quantum neural networks (2020).
[78] C. Zhao and X.-S. Gao, Analyzing the barren plateau phenomenon in training quantum neural networks with the ZX-calculus, Quantum 5, 466 (2021).
[79] T. Haug and M. S. Kim, Optimal training of variational quantum algorithms without barren plateaus (2021).
[80] R. Seeger and J. A. Pople, Self-consistent molecular orbital methods. XVIII. Constraints and stability in Hartree–Fock theory, The Journal of Chemical Physics 66, 3045 (1977).