Robustness of optimized numerical estimation schemes for noisy variational quantum algorithms

Y. S. Teo
Department of Physics and Astronomy, Seoul National University, 08826 Seoul, South Korea

(Received 15 October 2023; accepted 5 January 2024; published 26 January 2024)

With a finite amount of measurement data acquired in variational quantum algorithms, the statistical benefits of several optimized numerical estimation schemes, including the scaled parameter-shift (SPS) rule and finite-difference (FD) method, for estimating gradient and Hessian functions over analytical schemes [unscaled parameter-shift (PS) rule] were reported by the present author [Phys. Rev. A **107**, 042421 (2023)]. We continue the saga by exploring the extent to which these numerical schemes remain statistically more accurate for a given number of sampling copies in the presence of noise. For noise-channel error terms that are independent of the circuit parameters, we demonstrate that *without any knowledge* about the noise channel, using the SPS and FD estimators optimized specifically for noiseless circuits can still give lower mean-squared errors than PS estimators for substantially wide sampling-copy number ranges—specifically for SPS, closed-form mean-squared error expressions reveal that these ranges grow exponentially in the qubit number and reciprocally with a decreasing error rate. Simulations also demonstrate similar characteristics for the FD scheme. Lastly, if the error rate is known, we propose a noise-model-agnostic error-mitigation procedure to optimize the SPS estimators under the assumptions of two-design circuits and circuit-parameter-independent noise-channel error terms. We show that these heuristically optimized SPS estimators can significantly reduce mean-squared-error biases that naive SPS estimators possess even with realistic circuits and noise channels, thereby improving their estimation qualities even further. The heuristically optimized FD estimators possess as much mean-squared-error biases as the naively optimized counterparts and are thus not beneficial with noisy circuits.

DOI: 10.1103/PhysRevA.109.012620

I. INTRODUCTION

Quantum computation is a key theoretical milestone of quantum information theory [1] where prospective quantum computers and devices [2–7] are used to perform tasks with computation power that could in theory surpass classical computers. This prompted the invention of a plethora of quantum-computation and cryptographic algorithms [8–14]. In practice, we are still in the era of *noisy intermediate-scale quantum* (NISQ) devices [15] which run algorithms on noisy circuits and a limited number of working qubits [16–24]. These include the class of *variational quantum algorithms* (VQAs) [25–29] that rely on the interplay between classical and NISQ devices. Examples are quantum eigensolvers designed for quantum chemistry [30–32], combinatorial tasks [33,34], and quantum machine learning [35–43].

Among the multiple problems faced by NISQ devices, efficient circuit sampling and accurate circuit-function estimation are important goals for achieving practical quantum computation [28,44,45]. In recent years, there have been proposals to employ *analytical* estimation schemes, commonly known as the parameter-shift rule (PS) in the quantum-computing community [46–49], to exactly estimate gradients and/or Hessians VQAs that rely on, for instance, steepest gradient-descent [50–53] and quantum natural gradient-descent methods [54–58] in function optimization. On the other hand, *numerical* estimation schemes are frequently criticized because they are statistically biased and introduce approximation errors. Especially for the finite-difference (FD) scheme, the general mindset has been that decreasing approximation errors requires a very small step size and thus a large number of sampling copies to reduce estimation errors.

Contrary to the above consensus, we note that when circuit functions are to be *sampled*, an FD estimator that gives the minimum *mean-squared error* (MSE, synonymous to "estimation error," or "sampling error" as in Ref. [59]) will have an optimized step size that is *not* small—statistical bias is generally necessary to minimize the MSE [60–64]. For Pauli-encoded parametrized quantum circuits (PEPQCs) in which variable parameters are encoded on single-qubit Pauli rotation gates, the present author argued in Ref. [59] that the additional free parameter in a numerical estimator, such as one of those of the FD scheme (or its generalized versions) or the scaled PS (SPS) scheme, should be optimized by minimizing the respective MSE averaged over two-design circuits [65–67]. In situations where barren plateaus exist [68–71], that is when the circuit function and its gradient and Hessian magnitudes drop exponentially with the qubit number, these optimized estimators offer exponentially lower mean-squared errors relative to those from PS for a *fixed* number of sampling copies.

In this work, we demonstrate that optimized numerical estimators can still be statistically more accurate than analytical ones when quantum circuits are subjected to noise channels, which is part of a crucial research topic that is intimately related to the possibility of a quantum advantage using noisy circuits, especially when the qubit number is large [72–85]. After recalling the concepts of gradient and Hessian estimation in Sec. II and noisy quantum circuits in Sec. III, we first supply closed-form MSE expressions for both the FD and SPS schemes averaged over two-design circuits for a given

noise error rate η and qubit number n in Sec. IV A. With these, we show in Sec. IV B that if SPS estimators optimized for noiseless quantum circuits are used to estimate gradient and Hessian components of noisy circuits, then the critical sampling-copy numbers below which these naively optimized SPS estimators outperform the PS ones all grow as $\mathcal{O}(2^n/\eta)$ with increasingly large n and decreasing η. While FD exhibits no closed-form results for these critical numbers, simulations exhibit similar behaviors for both naively optimized numerical schemes with noisy hardware-efficient circuits.

While the original SPS and FD estimators may be employed when one has absolutely no knowledge about the noise channel, using these naively optimized numerical schemes that are strictly catered only to noiseless circuits will result in asymptotically wrong estimated gradient and Hessian components. In Sec. V, when *only* the noise-channel error rate is known, we introduce a heuristic error mitigation strategy to reduce the noise biases. Based on the assumptions of unitary two-designs circuits and that the noise-channel error terms do not depend on the circuit parameters, this strategy is independent of the kind of circuit noise channel: it minimizes the MSE *upper bound* over the free parameter for the chosen numerical estimator, which depends only on the error rate and not the noise-channel type. Such a procedure is therefore operational since a very accurate and complete description of the noise channel is unnecessary.

Under this error-mitigation strategy, we find that the heuristically optimized SPS scheme offers a much more significant reduction in the MSE relative to the naively optimized SPS and PS schemes even with hardware-efficient quantum circuits and realistic circuit noise. However, the heuristically optimized FD estimators obtained from this error-mitigation strategy are still as noisily biased as their naively optimized counterparts due to the way statistical biases enter the approximation errors. This establishes the heuristically optimized SPS scheme as the preferred choice for noisy gradient and Hessian estimation when the noise-channel error rate is known prior to the estimation.

II. NUMERICAL AND ANALYTICAL GRADIENT AND HESSIAN ESTIMATORS FOR VARIATIONAL QUANTUM ALGORITHMS

A parametrized quantum circuit (PQC) represented by the unitary operator U_θ, where θ is a collection of real parameters characterizing this circuit of a certain *ansatz*, together with some Hermitian measurement observable O, defines a real circuit function $f_Q(\theta) = \langle 0|U_\theta^\dagger O U_\theta|0\rangle$. Here, $|0\rangle\langle 0| = (|0\rangle\langle 0|)^{\otimes n}$ is some n-qubit pure state initialized to the zero bit-string state of the computational basis. A core purpose of VQAs is to minimize $f_Q(\theta)$ over θ. Examples of problems relevant to this task are eigenvalue-minimization schemes such as variational quantum eigensolvers [30–32] and quantum approximate optimization algorithms [33,34], where O is a Hamilton operator of either a physical system or combinatorial problem. The PQCs may include classical-data encoding, as in quantum machine learning [35–43]. Additionally, since the traceless part of O may be written as a sum of traceless Pauli basis operators that are usually each measured independently in an experiment, we consider O as a traceless Pauli operator without loss of generality.

To present the key results and important messages more easily, we consider *Pauli-encoded parametrized quantum circuits* (PEPQCs), which are circuits that encode variable parameters on single-qubit Pauli rotation gates defined by the standard Pauli operators $X \equiv \sigma_x$, $Y \equiv \sigma_y$, and $Z \equiv \sigma_z$. For such PEPQCs and circuits encoded on single-qubit gates of slightly more general Hermitian generators [49], one can *exactly* write down the gradient and Hessian components of $f_Q(\theta)$. If we consider a rather general and universal circuit *ansatz* consisting of L layers, where each layer comprises single-qubit and two-qubit controlled-NOT (CNOT) gates, such that $U_\theta = W_L W_{L-1}\cdots W_2 W_1$, these are

$$
\begin{aligned}
&[\partial_{\mathrm{PS}}]_{\mu,l} f_Q \equiv \partial_{\mu,l} f_Q = \frac{f_Q(\theta_{\mu l} + \pi/2) - f_Q(\theta_{\mu l} - \pi/2)}{2},\\
&[\partial_{\mathrm{PS}}]_{\mu,l}[\partial_{\mathrm{PS}}]_{\mu',l'} f_Q \equiv \partial_{\mu,l}\partial_{\mu',l'} f_Q\\
&\quad= \frac{f_Q\left(\theta_{\mu l} + \frac{\pi}{2}, \theta_{\mu' l'} + \frac{\pi}{2}\right) - f_Q\left(\theta_{\mu l} + \frac{\pi}{2}, \theta_{\mu' l'} - \frac{\pi}{2}\right)}{4}\\
&\qquad- \frac{f_Q\left(\theta_{\mu l} - \frac{\pi}{2}, \theta_{\mu' l'} + \frac{\pi}{2}\right) - f_Q\left(\theta_{\mu l} - \frac{\pi}{2}, \theta_{\mu' l'} - \frac{\pi}{2}\right)}{4},
\end{aligned}
\tag{1}
$$

where the pair (μ, l) labels the μth circuit parameter $\theta_{\mu l}$ located in the unitary operator W_l. All *other unspecified* parameters of f_Q in the above formulas are otherwise untranslated. The right-hand sides of (1) constitute the so-called *parameter-shift* (PS) scheme, which is an *analytical* scheme as it exactly computes the gradient and Hessian components. Since VQAs are iterations of f_Q sampling from a PQC and value updates with a classical computer, these gradient and Hessian components are also estimated from a finite number of sampling copies. We therefore denote the corresponding estimator versions as $\widehat{[\partial_{\mathrm{PS}}]_{\mu,l} f_Q}$ and $[\partial_{\mathrm{PS}}]_{\mu,l}\widehat{[\partial_{\mathrm{PS}}]_{\mu',l'} f_Q}$, where each function estimator $\widehat{f_Q}$ is obtained from measuring N copies of the PQC output state in the eigenbasis of a Pauli observable O. These estimators therefore possess *finite-copy errors*.

There is another class of *numerical* schemes that approximately defines gradient and Hessian components. One of which is the (centralized) finite-difference (FD) scheme (its generalized variants shall not be discussed here):

$$
\begin{aligned}
&[\partial_{\mathrm{FD}}]^{\epsilon}_{\mu,l} f_{Q,k} \equiv \mathrm{sinc}(\epsilon/2)\partial_{\mu l} f_{Q,k}\\
&\quad= \frac{f_{Q,k}(\theta_{\mu l} + \epsilon/2) - f_{Q,k}(\theta_{\mu l} - \epsilon/2)}{\epsilon},\\
&[\partial_{\mathrm{FD}}]^{\epsilon}_{\mu,l}[\partial_{\mathrm{FD}}]^{\epsilon}_{\mu',l'} f_{Q,k} \equiv [\mathrm{sinc}(\epsilon/2)]^2\partial_{\mu,l}\partial_{\mu',l'} f_{Q,k}\\
&\quad= \frac{f_Q\left(\theta_{\mu l} + \frac{\epsilon}{2}, \theta_{\mu' l'} + \frac{\epsilon}{2}\right) - f_Q\left(\theta_{\mu l} + \frac{\epsilon}{2}, \theta_{\mu' l'} - \frac{\epsilon}{2}\right)}{\epsilon^2}\\
&\qquad- \frac{f_Q\left(\theta_{\mu l} - \frac{\epsilon}{2}, \theta_{\mu' l'} + \frac{\epsilon}{2}\right) - f_Q\left(\theta_{\mu l} - \frac{\epsilon}{2}, \theta_{\mu' l'} - \frac{\epsilon}{2}\right)}{\epsilon^2},
\end{aligned}
\tag{2}
$$

for $\epsilon > 0$. Note that the effective multiplicative factors involving $\mathrm{sinc}(\epsilon/2) = 2\sin(\epsilon/2)/\epsilon$ is a special property of

PEPQCs, where

$$f_{\rm Q}(\theta_{\mu,l}+\theta_0) = f_{\rm Q}(\theta_{\mu,l}) + \sin\theta_0\partial_{\mu,l}f_{\rm Q}(\theta_{\mu,l}) + (1-\cos\theta_0)(\partial_{\mu,l})^2 f_{\rm Q}(\theta_{\mu,l}). \quad (3)$$

One may also consider a different numerical scheme where a scalar parameter λ is multiplied to all true gradient and Hessian components. This nabs us the *scaled parameter-shift* (SPS) scheme inasmuch as

$$[\partial_{\rm SPS}]^{\lambda}_{\mu,l} f_{\rm Q} \equiv \lambda[\partial_{\rm PS}]_{\mu,l} f_{\rm Q},$$
$$[\partial_{\rm SPS}]^{\lambda}_{\mu,l}[\partial_{\rm SPS}]^{\lambda}_{\mu',l'} f_{\rm Q} \equiv \lambda[\partial_{\rm PS}]_{\mu,l}[\partial_{\rm PS}]_{\mu',l'} f_{\rm Q}. \quad (4)$$

Notice the difference between how the free parameters enter the SPS and FD schemes.

Unlike the analytical PS scheme, both of these numerical schemes introduce additional *approximation errors* whenever $\epsilon \neq 0$ and $\lambda \neq 1$. Hence, in VQAs, the estimator counterparts $\widehat{[\partial_{\rm FD}]^{\epsilon}_{\mu,l}} f_{\rm Q}$, $[\partial_{\rm FD}]^{\epsilon}_{\mu,l}\widehat{[\partial_{\rm FD}]^{\epsilon}_{\mu',l'}} f_{\rm Q}$, $\widehat{[\partial_{\rm SPS}]^{\lambda}_{\mu,l}} f_{\rm Q}$, and $[\partial_{\rm SPS}]^{\lambda}_{\mu,l}\widehat{[\partial_{\rm SPS}]^{\lambda}_{\mu',l'}} f_{\rm Q}$ will also be *statistically biased* (for instance, the data averages $\overline{\widehat{[\partial_{\rm FD}]^{\epsilon}_{\mu,l}} f_{\rm Q}} \neq \partial_{\mu,l} f_{\rm Q}$ and $\overline{\widehat{[\partial_{\rm SPS}]^{\lambda}_{\mu,l}} f_{\rm Q}} \neq \partial_{\mu,l} f_{\rm Q}$) and result in approximation errors. This means that these estimators possess *both* finite-copy and approximation errors.

As a measure for the estimation quality or accuracy, we investigate the mean-squared error (MSE):

$$\mathcal{D}(\partial f_{\rm Q}) = \overline{\langle(\widehat{[\partial]_{\mu,l} f_{\rm Q}} - \partial_{\mu,l} f_{\rm Q})^2\rangle},$$
$$\mathcal{D}(\partial\partial f_{\rm Q}) = \overline{\langle(\widehat{[\partial]_{\mu,l}[\partial]_{\mu,l}} f_{\rm Q} - (\partial_{\mu,l})^2 f_{\rm Q})^2\rangle},$$
$$\mathcal{D}(\partial\partial' f_{\rm Q}) = \overline{\langle(\widehat{[\partial]_{\mu,l}[\partial]_{\mu',l'}} f_{\rm Q} - \partial_{\mu,l}\partial_{\mu',l'} f_{\rm Q})^2\rangle}. \quad (5)$$

The notation $\langle\ \rangle$ and $\overline{\ }$ respectively refer to averages over θ and measurement data per θ, and $\partial\partial f_{\rm Q}$ and $\partial\partial' f_{\rm Q}$ are shorthand for diagonal and off-diagonal Hessian components.

Technically, a numerical estimator would end up with an MSE that is a *sum* of the finite-copy and approximation errors [see Eq. (B2)]. On the other hand, an analytical estimator only has the finite-copy error as its MSE. So, why are numerical estimators interesting? Well, in textbook scenarios, they are not when $f_{\rm Q}$ is computable exactly, in which case $\epsilon = 0$ and $\lambda = 1$ should be the only sensible options for error-free gradient and Hessian computation. However, they become interesting when $f_{\rm Q}$ has to be sampled from PQCs. Then, the free parameter (ϵ or λ) of a numerical estimator can be chosen as the optimal one that minimizes the MSE.

In Ref. [59], the optimized numerical estimators were shown to give MSEs that drop exponentially in the qubit number n. For finite sampling-copy numbers N, FD estimators outperform the PS ones up to some critical $N = N_*$, beyond which PS becomes more accurate. Furthermore, it has also been shown that SPS always outperforms PS for any given N, making them the more favorable method over FD in practice *especially* when n is large. The intuitive understanding of this difference for noiseless VQA lies in the highly nonlinear dependence on ϵ in both the finite-copy [$\sim 1/(N\epsilon^{\alpha})$] and approximation [$\sim\mathcal{O}(2^{-n})[1-\mathrm{sinc}(\epsilon/2)^{\beta}]^2$] error terms of FD estimators, where $\alpha = 2, 4$ and $\beta = 1, 2$, resulting in optimized MSEs that scale as $2^{-n}N^{-\kappa}$ with $0 < \kappa < 1$ for large N, which will eventually be larger than those of the PS estimators, which scale as $1/N$. Both the finite-copy and approximation error terms of SPS estimators, on the other hand, are only *quadratic* in λ, and this leads to optimized MSEs $\propto 1/N$ when N is large.

Another intuitive guide as to why this is the case: for smaller N values and large qubit number n, SPS's finite-copy errors $\sim\lambda^2/N$ and approximation errors $\sim\mathcal{O}(2^{-n})(1-\lambda)^2 \to 0$ in the presence of barren plateaus, such that the optimal $\lambda \sim \mathcal{O}(2^{-n}) \to 0$ and the optimized SPS MSEs ($\sim\lambda^2/N$) are clearly much smaller than the PS MSEs.

These results demonstrate that the statistical bias in an estimator, *when optimized properly*, is a key ingredient for minimizing its MSE, an insight well-known in sampling theory [60–64].

III. NOISY QUANTUM CIRCUITS

Realistic (PE)PQCs are always susceptible to noise in the form of a noise-channel action, namely, $\rho_\theta = U_\theta|0\rangle\langle 0|U_\theta^\dagger \mapsto \rho'_\theta = \mathcal{E}[\rho_\theta]$, which is completely positive and trace-preserving. The resulting noisy mixed state ρ'_θ can always be written as

$$\rho'_\theta = \rho_{\theta,\eta} = (1-\eta)\rho_\theta + \eta\rho_{\rm err}(\theta,\eta), \quad (6)$$

with η characterizing the error rate, or the strength of the noise-channel map $\mathcal{E}$, and $\rho_{\rm err}(\theta,\eta)$ is the noise-channel error term that is typically a function of both the noiseless state ρ_θ and η, and, thus, also depends on θ.

One example of a realistic noise channel is the successive action of a two-qubit depolarizing channel on an n-qubit quantum state ρ_0 after every two-qubit unitary operation, such as a CNOT-gate operation $U_{{\rm CNOT},jk} = |0\rangle_{jj}\langle 0|1_k + |1\rangle_{jj}\langle 1|X_k$ on qubits j and k:

$$U_{{\rm CNOT},jk}\rho_0 U_{{\rm CNOT},jk}^\dagger \mapsto (1-\eta_{j,k})U_{{\rm CNOT},jk}\rho_0 U_{{\rm CNOT},jk}^\dagger + \frac{\eta_{j,k}}{15}\sum_{1\neq P_{jk}\in\mathcal{P}_2^{(j,k)}} P_{jk}U_{{\rm CNOT},jk}\rho_0 U_{{\rm CNOT},jk}^\dagger P_{jk}, \quad (7)$$

where $\eta_{j,k}$ is the error rate from this CNOT operation and $\mathcal{P}_2^{(j,k)} = \{1, X_j, Y_j, Z_j\}\times\{1, X_k, Y_k, Z_k\}$ is the set of two-qubit Pauli operators and the identity for qubits j and k. Hence, the unitary operator $W_1 = U_{\rm CNOTs}R_1$ for the first layer of an n-qubit circuit *ansatz* (see Fig. 1) consisting of parameter-encoded near-noiseless single-qubit gates $R_1 = R_1^{(1)}\otimes R_2^{(1)}\otimes\cdots\otimes R_n^{(1)}$ followed by an array of n noisy CNOT operations $U_{\rm CNOTs} = U_{{\rm CNOT},n1}U_{{\rm CNOT},n-1n}\cdots U_{{\rm CNOT},23}U_{{\rm CNOT},12}$ gives the noisy state $\rho_\eta^{(1)} = W_1|0\rangle(1-\eta_1)\langle 0|W_1^\dagger + \eta_1\rho_{\rm err}^{(1)}$, with $1-\eta_1 = \prod_{j=1}^n(1-\eta_{j,\ \mathrm{mod}\ (j,n)+1})$. It follows that the noisy version of an L-layered *ansatz* state defined by $U_\theta = W_L W_{L-1}\cdots W_2 W_1$ is given by

$$\rho_{\theta,\eta}^{(L)} = U_\theta|0\rangle(1-\eta)\langle 0|U_\theta^\dagger + \eta\rho_{\rm err}^{(L)}(\theta,\eta),$$
$$\eta = 1 - \prod_{l=1}^{L}\prod_{j=1}^{n}\left[1-\eta_{j,\ \mathrm{mod}\ (j,n)+1}^{(l)}\right]. \quad (8)$$

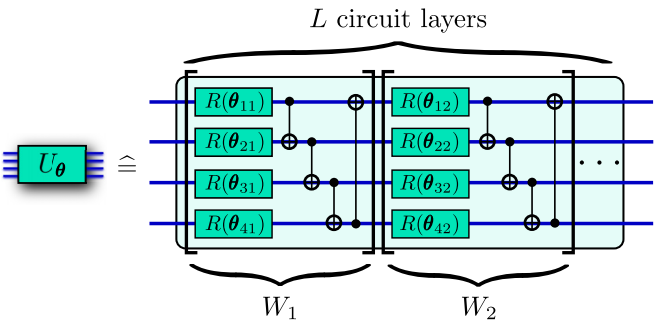

FIG. 1. Schematic of a four-qubit circuit representation of a U_θ unitary operator. Each W_l, or a circuit unitary layer, comprises a chain of parametrized single-qubit rotations R and a CNOT array.

Specifically, when $\eta_{j,k}^{(l)} = \eta_0$ are equal, $\eta = 1 - (1-\eta_0)^{nL}$. For small η_0, we consequently find that $\eta \cong nL\eta_0$.

In all numerical simulations, the circuit *ansatz* as illustrated in Fig. 1 shall be adopted, where the encoded parameters on all single-qubit rotation gates are chosen according to the Haar measure on the qubit unitary group. Furthermore, noisy CNOT gates that bring about (7) and (8) with a constant error rate η_0 per CNOT gate. All single-qubit gates are always taken to have unit fidelity for granted. The corresponding figures of merit are still the MSEs, but this time, the *estimators* are noisy, whereas the *true* components are not. The MSE definitions are otherwise similar to those in (5).

IV. RESULT 1: ADVANTAGES OF USING NUMERICAL ESTIMATORS OPTIMIZED FOR NOISELESS CIRCUITS

A. Noisy mean-square errors of numerical estimators

For the sole purpose of acquiring an analytical understanding of the performance of numerical schemes on a noisy circuit, we first assume that $\rho_{\rm err}^{(L)}(\theta,\eta) = \rho_{\rm err}^{(L)}(\eta)$ does not depend on the circuit parameters θ. This implies that the noisy function

$$f_{{\rm Q},\eta}(\theta) = (1-\eta) f_{\rm Q}(\theta) + \eta g \tag{9}$$

is a sum of the noiseless $f_{\rm Q}(\theta)$ and some constant term $g = {\rm tr}\{\rho_{\rm err}^{(L)}(\eta) O\}$ that is independent of θ. Note that $-1 \leqslant g \leqslant 1$ still generally depends on η, but we suppress this dependence for notational simplicity as it shall be irrelevant for subsequent discussions unless otherwise required.

Figure 2 illustrates that, for PEPQCs (that is, all $R_k^{(l)}$ are products of encoded Pauli rotation gates) with the circuit noise channel of uniform error rate η_0 described in Sec. III, which shall be the noise channel of choice for all subsequent simulations, the distribution of g in θ is generally much flatter than that of $f_{\rm Q}(\theta)$. The constant-g (in θ) assumption thus serves as a reasonable approximation for such a physically motivated noise model. This is slightly elaborated in Appendix A.

The second assumption that we shall make to facilitate the analysis is that all θ averages are well approximated by averages over unitary two-designs. This means that the first and second moments of U_θ coincide with the Haar measure of the unitary group [86,87]. This is a rather good approximation as broad classes of circuits with polynomial and logarithmic circuit depths are approximately unitary two-designs [65–67].

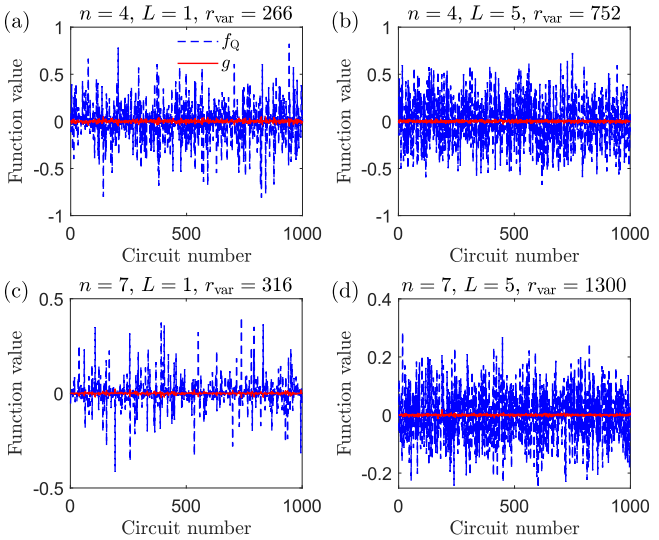

FIG. 2. Distributions of $f_{\rm Q}$ and g in $f_{{\rm Q},\eta}$ for PEPQCs with (a), (b) $n = 4$ and (c), (d) $n = 7$ qubits over 1000 sets of randomly generated PEPQC parameters (Haar-distributed single-qubit unitary rotations) in each figure panel. The ratio $r_{\rm var} = {\rm Var}_\theta[f_{\rm Q}]/{\rm Var}_\theta[g]$ is given in every panel. The CNOT depolarizing error rate is set at $\eta_0 = 0.05$ and the overall error rates $\eta = 1-(1-\eta_0)^{nL}$ are (a) 0.185, (b) 0.642, (c) 0.302, and (d) 0.834. The respective observables are $O = X_1Y_2Z_3X_4$ and $O = X_1Y_2Z_3X_4Y_5Z_6X_7$.

To theoretically establish the performance of FD and SPS, we derive analytical MSE formulas that are strictly relevant to gradient and Hessian components corresponding to parameters whose encoded Pauli rotation gates are each sandwiched by two-design circuit unitary operators (see Fig. 3). This shall be coined the *two-design sandwich* (TDS) condition, which is satisfied for the majority of gradient and Hessian components in the bulk of a sufficiently deep circuit. In Fig. 4, for instance, we show that the simulated average MSEs fit all the TDS-based theoretical expressions well, even though the purposefully chosen gradient and Hessian components do not satisfy the TDS condition.

When O is a traceless Pauli observable, this two-design framework and the TDS condition permits one to obtain

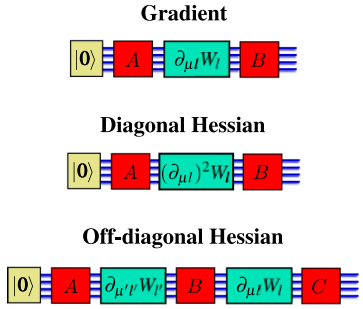

FIG. 3. (Extracted from Ref. [59] and modified.) A visual illustration of the TDS condition. The (red) blocks A, B, and C are unitary operators drawn from a set of unitary two-designs. They sandwich every gradient and Hessian component. This allows for analytical MSE expressions for the numerical estimation schemes.

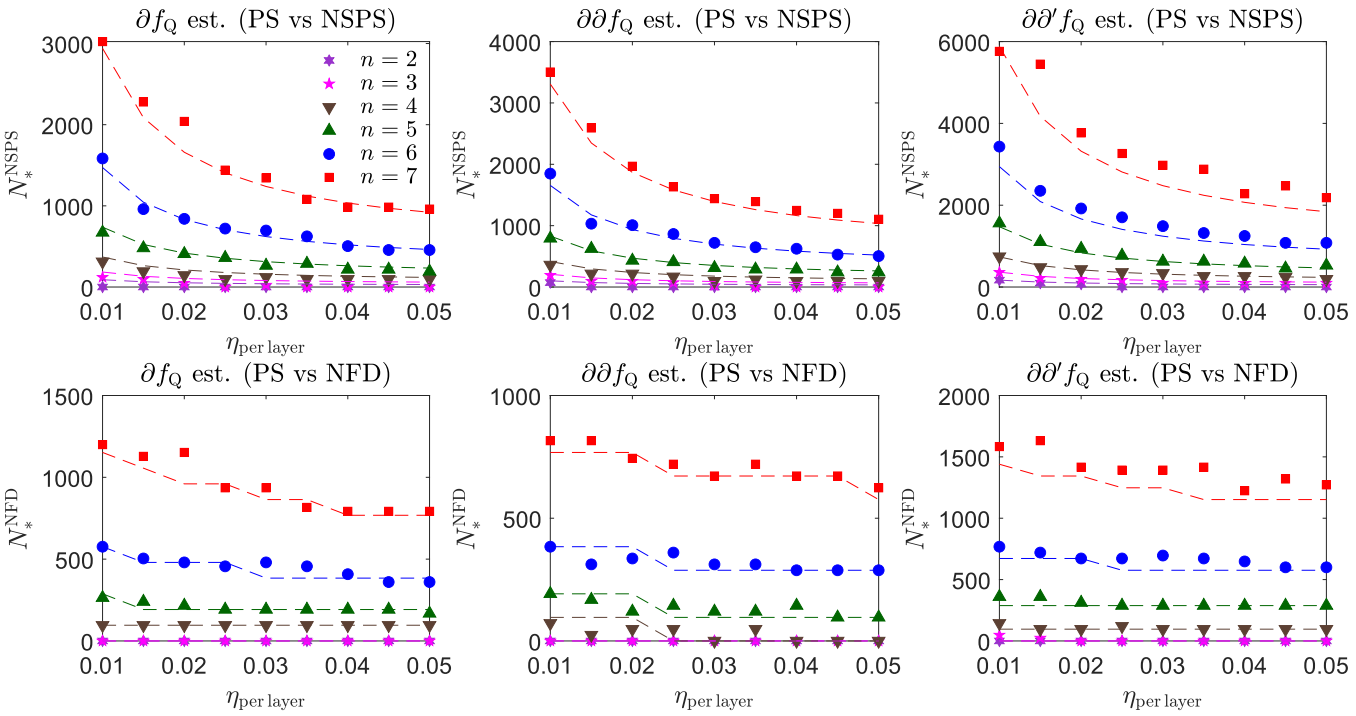

FIG. 4. Monte Carlo simulations generating the respective N_* behaviors in gradient and Hessian estimations with $\eta_{\rm per\ layer}$ for $L = 5$ in the regime of small $\eta_{\rm per\ layer}$. Noisy CNOT gates as in (7) is considered here. The dashed curves in the top figures represent the explicit $N_*^{\rm NSPS}$ expressions for $g = 0$ in (B11), whereas those in the bottom figure trace all $N_*^{\rm NFD}$ values by numerically solving $\mathcal{D}_{\rm NFD}(\cdot) = \mathcal{D}_{\rm PS}(\cdot)$ for $N_{\rm T} \equiv N_*^{\rm NFD}$ when $g = 0$, all derived based on the two main assumptions. The simulation markers are obtained from MSEs averaged over 2000 sets of random PEPQC parameters (Haar-distributed single-qubit unitary rotations) and 1000 sampling experiments per PEPQC parameter set. We discretize the range of $N_{\rm T}$ in steps of 96 copies so that the minimum division is sufficiently large to be distributed to 2, 3, and 4 sampled functions for defining gradient and Hessian estimators as per Sec. IV A. We pick $\mu = 1$ and $l = 2$ to specify the location of the gradient and diagonal Hessian components considered in this figure, and $\mu = 1$, $\mu' = 2$, and $l = l' = 2$ for the off-diagonal Hessian component. As examples, all evaluated gradient and Hessian circuit parameters are encoded onto to the Pauli Y-rotation gate. The exponentially increasing trend of N_* with n is numerically evident. The respective observables O for different n are cyclic repetitions of X, Y, and Z in this order for every qubit following Fig. 2.

PEPQC-based identities that are independent of (μ, l) [59]:

$$\begin{aligned}
\langle f_{\rm Q}\rangle &= 0, \quad \langle f_{\rm Q}^2\rangle = \frac{1}{d+1},\\
\langle|\partial f_{\rm Q}|^2\rangle &= \langle|\partial\partial f_{\rm Q}|^2\rangle = \frac{d^2}{2(d+1)(d^2-1)}\ \text{(TDS)},\\
\langle|\partial\partial' f_{\rm Q}|^2\rangle &= \frac{d^4}{4(d+1)(d^2-1)^2}\ \text{(TDS)}.
\end{aligned} \tag{10}$$

To summarize, in order to acquire some analytical understanding of gradient and Hessian estimation with noisy quantum circuits (in terms of MSE expressions and lemmas), we make the following two physically reasonable assumptions:

(1) The noise channel generates error terms that are independent of the circuit parameters θ (constant g in θ).

(2) The noiseless U_θ is a unitary two-design, where gradient and Hessian components of its parameters conform to the TDS condition such that Eq. (10) hold.

Under these two assumptions, for traceless Pauli observables, we arrive at the following *exact* MSEs for FD:

$$\begin{aligned}
\mathcal{D}_{\rm FD}(\partial f_{\rm Q}) &= \frac{4}{N_{\rm T}\epsilon^2}\left[1-(1-\eta)^2\langle f_{\rm Q}^2\rangle - \eta^2 g^2\right]\\
&\quad + [1-(1-\eta)\mathrm{sinc}(\epsilon/2)]^2\langle|\partial f_{\rm Q}|^2\rangle,\\
\mathcal{D}_{\rm FD}(\partial\partial f_{\rm Q}) &= \frac{18}{N_{\rm T}\epsilon^4}\left[1-(1-\eta)^2\langle f_{\rm Q}^2\rangle - \eta^2 g^2\right]\\
&\quad + \{1-(1-\eta)[\mathrm{sinc}(\epsilon/2)]^2\}^2\langle|\partial\partial f_{\rm Q}|^2\rangle,\\
\mathcal{D}_{\rm FD}(\partial\partial' f_{\rm Q}) &= \frac{16}{N_{\rm T}\epsilon^4}\left[1-(1-\eta)^2\langle f_{\rm Q}^2\rangle - \eta^2 g^2\right]\\
&\quad + \{1-(1-\eta)[\mathrm{sinc}(\epsilon/2)]^2\}^2\langle|\partial\partial' f_{\rm Q}|^2\rangle,
\end{aligned} \tag{11}$$

and those for SPS,

$$\begin{aligned}
\mathcal{D}_{\rm SPS}(\partial f_{\rm Q}) &= \frac{\lambda^2}{N_{\rm T}}\left[1-(1-\eta)^2\langle f_{\rm Q}^2\rangle - \eta^2 g^2\right]\\
&\quad + [1-(1-\eta)\lambda]^2\langle|\partial f_{\rm Q}|^2\rangle,
\end{aligned}$$

$$\mathcal{D}_{\mathrm{SPS}}(\partial\partial f_{\mathrm{Q}}) = \frac{9\lambda^2}{8N_{\mathrm{T}}}\left[1-(1-\eta)^2\langle f_{\mathrm{Q}}^2\rangle - \eta^2 g^2\right] + [1-(1-\eta)\lambda]^2\langle|\partial\partial f_{\mathrm{Q}}|^2\rangle,$$
$$\mathcal{D}_{\mathrm{SPS}}(\partial\partial' f_{\mathrm{Q}}) = \frac{\lambda^2}{N_{\mathrm{T}}}\left[1-(1-\eta)^2\langle f_{\mathrm{Q}}^2\rangle - \eta^2 g^2\right] + [1-(1-\eta)\lambda]^2\langle|\partial\partial' f_{\mathrm{Q}}|^2\rangle. \quad (12)$$

Here, N_{T} is the *total sampling-copy number* for estimating the corresponding components. If N copies are used to estimate the function f_{Q}, then $N_{\mathrm{T}} = 2N$ for gradient-component estimation, and $N_{\mathrm{T}} = 3N$ and $4N$, respectively, for the diagonal and off-diagonal Hessian-component estimation. Setting $\lambda = 1$ gives us the MSEs for PS. The derivation of (11) and (12) may be found in Appendix B.

B. Naively optimized numerical schemes and their estimation advantages

It is now possible to compare the performances of SPS and PS in terms of gradient and Hessian estimation. In the noiseless case ($\eta = 0$), it is a straightforward matter to deduce [59] that SPS estimators with λ taking the following optimal values:

$$\lambda_{\mathrm{opt}} = \frac{dN_{\mathrm{T}}}{2d^2 + dN_{\mathrm{T}} - 2} \leqslant 1\ (\partial f_{\mathrm{Q}}\ \text{estimation}),$$
$$\lambda_{\mathrm{opt}} = \frac{4dN_{\mathrm{T}}}{9d^2 + 4dN_{\mathrm{T}} - 9} \leqslant 1\ (\partial\partial f_{\mathrm{Q}}\ \text{estimation}),$$
$$\lambda_{\mathrm{opt}} = \frac{d^3N_{\mathrm{T}}}{4(d^2-1)^2 + d^3N_{\mathrm{T}}} \leqslant 1\ (\partial\partial' f_{\mathrm{Q}}\ \text{estimation}), \quad (13)$$

will minimize their respective MSEs, and these optimized SPS estimators *always* offer smaller MSEs than those of PS regardless of the value of N_{T}. For $0 < \eta \leqslant 1$, when the *same* λ_{opt} in (13) are naively used in spite of the presence of noise, a smaller *naively optimized SPS* (NSPS) MSE can still be expected when N_{T} is below some critical N_*. So, a larger N_* signifies a more advantageous numerical scheme over PS. One can directly calculate N_* by setting $\mathcal{D}_{\mathrm{NSPS}}(\cdot) - \mathcal{D}_{\mathrm{PS}}(\cdot) = 0$, which is generally a complicated function of η, g, and d. The explicit expressions of N_* when $g = 0$, which is almost exact for our noise model from Fig. 2, are given in (B11).

To get a handle on the behavior of N_* in relation to d and η, we may further consider small η values such that all MSEs of the NSPS schemes may be expanded up to first order in η. In this small-η limit, the leading term $\rho_{\mathrm{err}}^{(L)}(\eta)$ is also a constant in η. The resulting $N_* = N_*^{\mathrm{NSPS}}$ values then read

$$N_*^{\mathrm{NSPS}} = \frac{(d^2-1)}{d\eta}\ (\partial f_{\mathrm{Q}}\ \text{estimation}),$$
$$N_*^{\mathrm{NSPS}} = \frac{9(d^2-1)}{8d\eta}\ (\partial\partial f_{\mathrm{Q}}\ \text{estimation}),$$
$$N_*^{\mathrm{NSPS}} = \frac{2(d^2-1)^2}{d^3\eta}\ (\partial\partial' f_{\mathrm{Q}}\ \text{estimation}). \quad (14)$$

Therefore, we have the following lemma:

Lemma 1. Gradient- and Hessian-estimation performance advantage of NSPS over PS. For an n-qubit two-design PEPQC that satisfies the TDS condition and any noise channel with a θ-independent error term leading to Eq. (9), NSPS outperforms PS if $N_{\mathrm{T}} < N_* = N_*^{\mathrm{NSPS}} \sim \mathcal{O}(2^n/\eta)$.

This result tells us that when the PEPQC is noisy, even without knowing *anything* about the noise channel (such as the error rate η), the NSPS estimators, that is those in (4) evaluated with $\lambda = \lambda_{\mathrm{opt}}$ in (13) can still give more accurate estimation than PS estimators when the sampling-copy number is limited. For circuits of very many qubits, owing to the influence of barren plateaus for universal PEPQC *ansatz̈e*, Lemma 1 informs us that these NSPS estimators are still the more accurate ones for very large copy numbers.

We may also naively optimize the FD estimators in much the same way as we did the NSPS estimators. Accordingly, we may consider the noiseless FD MSEs by setting $\eta = 0$ in all of (11) and minimize each of them over ϵ. Since ϵ enters the MSEs in a transcendental fashion, this minimization is done numerically. The optimal ϵ_{opt} is then a function of N_{T} and d, much like λ_{opt} in (13). If we now use these ϵ_{opt} to define the FD estimators for noisy circuits, we then have the analogous *naively optimized FD* estimators (NFD) and $N_* = N_*^{\mathrm{NSPS}}$ for them can similarly be found by noting the instant $\mathcal{D}_{\mathrm{NFD}}(\cdot) = \mathcal{D}_{\mathrm{PS}}(\cdot)$.

In Fig. 4, we compare the values of N_* for NFD and NSPS schemes by simulating PEPQCs possessing noisy CNOT gates with the channel action in (7). We do this for different number of qubits n given a *fixed error rate per layer*—$\eta_{\mathrm{per\ layer}}$. This allows us to compare MSE performances for various n fairly in an *ansatz*-free manner since the two-qubit gate count that scales with n is absorbed into this "error rate per layer" definition. For the circuit *ansatz* described in Sec. III which we are considering, $\eta_{\mathrm{per\ layer}} \equiv 1-(1-\eta_0)^n$. This is also $\eta_{\mathrm{per\ layer}} \cong n\eta_0$ when η_0 is small. The total error rate is hence $\eta \cong \eta_{\mathrm{per\ layer}} L$. Overall, naively ignoring noise by using the NSPS schemes can still achieve lower MSEs in contrast with PS. Compared with NFD, this happens for larger sampling-copy number ranges (or a larger N_*).

Nonetheless, both NSPS and NFD schemes, which are optimally tuned for noiseless quantum circuits, will evidently introduce *noise biases* (or MSE biases) even when $N_{\mathrm{T}} \to \infty$ if they are employed for noisy circuits. This is a consequence of optimizing over ϵ and λ by ignoring the presence of noise. As these noise biases are permanent systematic errors that cannot be eliminated even when $N_{\mathrm{T}} = \infty$, they are *not to be confused* with statistical biases of the numerical estimators as explained in Sec. II, which, when correctly optimized, can help eliminate noise biases as a matter of fact. In the next section, we discuss how these parameters may be properly tuned for noisy quantum circuits so that the resulting optimized statistical biases actually reduce noise biases.

V. RESULT 2: HEURISTIC ERROR-MITIGATION FOR NUMERICAL GRADIENT AND HESSIAN ESTIMATIONS

In the previous section, we found that numerical gradient and Hessian estimation schemes, when optimized under the negligence of noise in PEPQCs, can still give more accurate estimators for a given total sampling-copy number N_{T} value no larger than some critical value N_*. There is, however,

one more procedure we may carry out to further improve the estimation quality of numerical estimators with noisy NISQ circuits. In this section, we discuss a simple error-mitigation strategy that can be carried out if the error rate η of the overall noise channel is known *a priori*, which may be acquired through an initial device and channel calibration tests. Otherwise, no other explicit knowledge about the type of noise channel acting on a PEPQC is necessary.

Very briefly, this error-mitigation technique involves making a few heuristic assumptions such that the MSE expressions in (11) and (12) are valid, followed by eliminating the dependence of g by considering the upper bounds of these MSE expressions, and finally minimizing these upper bounds to obtain the respective optimal parameters to be used for defining the heuristically optimized numerical estimators given a specific η. As this technique does not require an accurate noise-channel superoperator description, such an error-mitigation scheme is appealingly feasible in real experimental situations, since a full characterization of 16^n real parameters of the noise-channel map is generally impractical for large n.

Let us now describe this error-mitigation protocol with more detail. As the primary step, we make the assumptions listed in Sec. IV A, which we repeat here once more: namely, that (1) the noise channel generates error terms (or g) that are constant in θ, and (2) all U_θ is a unitary two-design in which gradient and Hessian components satisfy the TDS condition, which leads to the MSE expressions in (11) and (12). To execute this error-mitigation strategy without the knowledge about the noise channel or g, the second step we take is to consider the upper bounds of the MSE expressions, which are obtained by simply removing "$\eta^2 g^2$" on the right-hand sides of (11) and (12). After minimizing the MSE upper bounds, the $\lambda_{\mathrm{opt},\eta}$ parameters characterizing the *heuristically optimized SPS* scheme (HSPS) are

$$\lambda_{\mathrm{opt},\eta} = \frac{dN_{\mathrm{T}}(1-\eta)}{2d^2 + dN_{\mathrm{T}} - 2 + \eta(2-\eta)\left(2d - dN_{\mathrm{T}} - \frac{2}{d}\right)} \quad (\partial f_{\mathrm{Q}} \text{ estimation}),$$

$$\lambda_{\mathrm{opt},\eta} = \frac{4dN_{\mathrm{T}}(1-\eta)}{9d^2 + 4dN_{\mathrm{T}} - 9 + \eta(2-\eta)\left(9d - 4dN_{\mathrm{T}} - \frac{9}{d}\right)} \quad (\partial\partial f_{\mathrm{Q}} \text{ estimation}),$$

$$\lambda_{\mathrm{opt},\eta} = \frac{d^3N_{\mathrm{T}}(1-\eta)}{4(d^2-1)^2 + d^3N_{\mathrm{T}} + \eta(2-\eta)\left[\frac{4}{d}(d^2-1)^2 - d^3N_{\mathrm{T}}\right]} \quad (\partial\partial' f_{\mathrm{Q}} \text{ estimation}). \tag{15}$$

When η is known *a priori*, such an error-mitigation strategy of minimizing MSE upper bounds over the parameters that characterize numerical estimators can reduce their noise biases significantly. Furthermore, the simplicity of the λ dependence in all SPS MSEs allows us to acquire a quantitative understanding of these noise biases.

Lemma 2. Gradient- and Hessian-estimation performance advantage of HSPS over NSPS and PS. Suppose an n-qubit two-design PEPQC satisfying the TDS condition is subjected to a noise channel of a fixed error rate $\eta > 0$ and a θ-independent error term leading to Eq. (9). Then when $N_{\mathrm{T}} \to \infty$, the NSPS and PS estimators give nonzero MSEs, which are respectively $\langle|\partial f_{\mathrm{Q}}|^2\rangle\eta^2$ for gradient estimation, $\langle|\partial\partial f_{\mathrm{Q}}|^2\rangle\eta^2$ and $\langle|\partial\partial' f_{\mathrm{Q}}|^2\rangle\eta^2$, respectively, for diagonal and off-diagonal Hessian estimation. On the other hand, the MSEs of HSPS estimators asymptotically approach zero.

This heuristic protocol may also apply to the FD scheme, which would then yield the *heuristically optimized FD* scheme (HFD). However, as in the NFD estimation protocol in Sec. IV B, the $\epsilon_{\mathrm{opt},\eta}$s for all gradient and Hessian estimations have no closed forms and should be obtained by numerically minimizing the upper bounds of the MSEs in (11). Unfortunately, it turns out that the HFD estimators acquired this way are just as noisily biased as the NFD ones. This is equivalently encapsulated in the next lemma:

Lemma 3. HFD and NFD schemes are asymptotically noisy. Suppose an n-qubit two-design PEPQC satisfying the TDS condition is subjected to a noise channel of a fixed error rate $\eta > 0$ and θ-independent error term leading to Eq. (9). Then when $N_{\mathrm{T}} \to \infty$, the NFD and HFD estimators give nonzero MSEs, which are respectively $\langle|\partial f_{\mathrm{Q}}|^2\rangle\eta^2$ for gradient estimation, $\langle|\partial\partial f_{\mathrm{Q}}|^2\rangle\eta^2$ and $\langle|\partial\partial' f_{\mathrm{Q}}|^2\rangle\eta^2$ respectively for diagonal and off-diagonal Hessian estimation.

The interested reader may refer to Appendix C for the simple arguments leading to Lemmas 2 and 3.

One can intuitively understand the reasons behind these two lemmas by observing that, for the HSPS estimators to completely eliminate noise biases in the limit of large N_{T}, we expect the condition $(1-\eta)\lambda_{\mathrm{opt},\eta} = 1$ to hold in order for the approximation error to approach zero, which means that $\lambda_{\mathrm{opt},\eta} > 1$, consistent with the large-N_{T} versions of (15). For the HFD estimators, demanding an asymptotically vanishing approximation error would entail the obedience of the analogous condition $(1-\eta)\mathrm{sinc}(\epsilon_{\mathrm{opt},\eta}/2) = 1$, which can never happen for *any* $\eta > 0$ since the sinc function is always less than or equal to one. The only way out is $\eta = 0$, where we recover the textbook asymptotic limit $\epsilon_{\mathrm{opt},0} = \epsilon_{\mathrm{opt}} \to 0$.

Figure 5 compares the NSPS and NFD schemes with their heuristically optimized counterparts HSPS and HFD using known η_0 or $\eta_{\text{per layer}}$, where the quantum circuit *ansatz* comprises layers of single-qubit gates followed by a complete array of noisy CNOT gates. The zero-noise-bias property of HSPS for constant-g noise channels as in Lemma 2 carries over to other more general noise models, such as the one governed by (7). However, because of the noise biases that persist in HFD estimators, such an error mitigation does improve the MSE. Moreover, the close competition between NFD and HFD and the proximity of their asymptotic noise biases may result in NFD estimators giving lower MSEs than HFD estimators. This is, however, allowed because only the upper bounds of the MSE are minimized to obtain the heuristic schemes, so there is no guarantee for HFD to always do better than NFD. Meanwhile, HSPS estimators would always outperform NSPS and PS for sufficiently large N_{T} if g is roughly a constant in $\boldsymbol{\theta}$.

These findings suggest that, for quantum circuits of sufficient depth and general noisy channels of a *known* error rate, whose error terms are approximately constants in the circuit parameters, HSPS is an advantageous gradient and Hessian

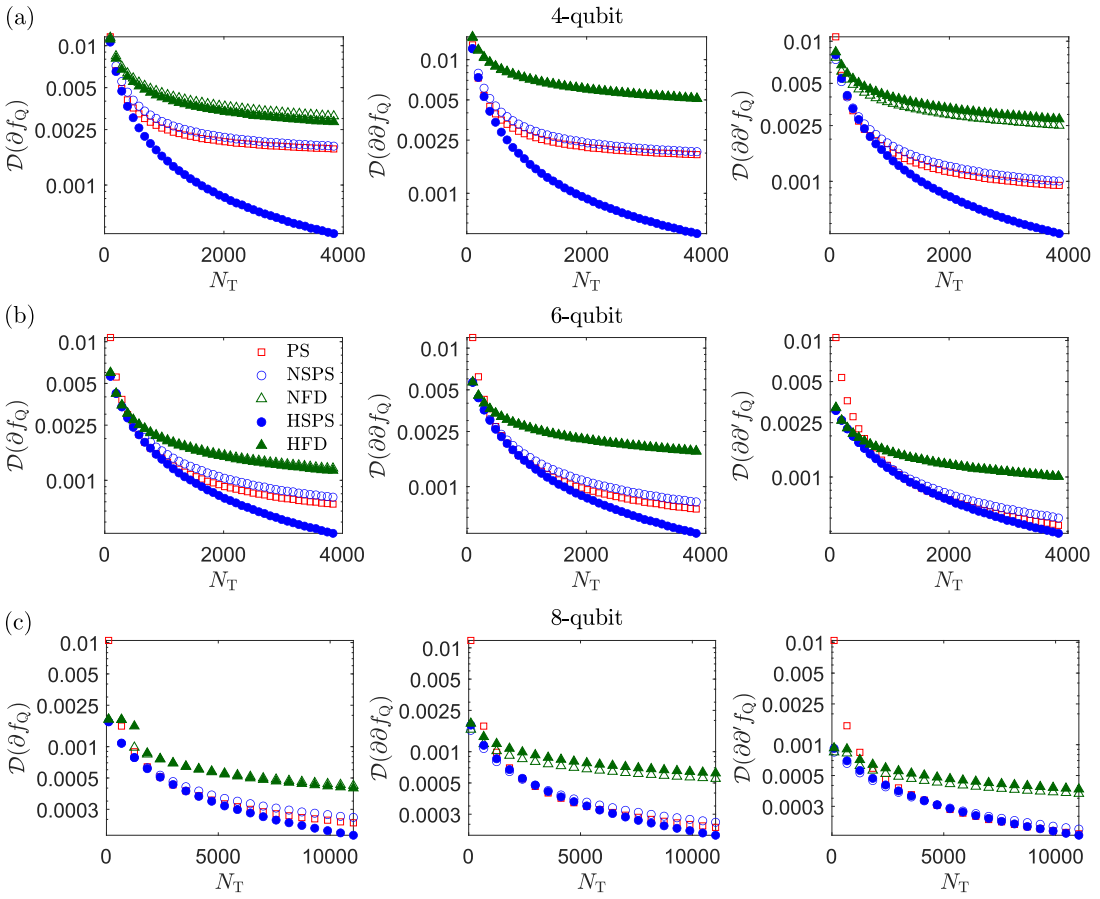

FIG. 5. Monte Carlo simulations comparing the naive (NFD, NSPS) and heuristically optimized (HFD, HSPS) numerical gradient and Hessian estimation schemes with (a) four-qubit, (b) six-qubit, and (c) eight-qubit PEPQCs for $\eta_{\text{per layer}} = 0.05$ and $L = 5$ circuit layers $[\eta = 1 - (1 - \eta_{\text{per layer}})^L = 0.226 \cong 0.25]$. The gradient and diagonal Hessian components are specified by $\mu = 1$ and $l = 2$, and the off-diagonal Hessian component by $\mu = 1$, $\mu' = 2$, and $l = l' = 2$, with evaluated gradient and Hessian circuit parameters encoded to the Y-rotation gate. All MSEs are averaged over 500 sets of random PEPQC parameters (Haar-distributed single-qubit unitary rotations) and 500 sampling experiments per PEPQC parameter set. In these plots, we observe that while HSPS significantly improves the MSEs with respect to NSPS and PS, there is almost no visible differences between HFD and NFD estimators even with the logarithmic-scale plots presented here. The respective observables O for the different n values are cyclic repetitions of X, Y, and Z in this order for every qubit, as in the captions of Figs. 2 and 4.

estimation scheme that can significantly reduce asymptotic noise biases for any given finite number of sampling copies. As the qubit number $n \to \infty$, the asymptotic noise biases of all estimators eventually go to zero, but so are the gradient and Hessian magnitudes anyway, which is a signature of the barren-plateau problem. In this detrimental regime, *no* estimation is feasible.

VI. CONCLUSION

The results of this work and those presented in the prequel article revolve around the use of numerical methods (such as finite-difference and scaled parameter-shift rule) to estimate the gradient and Hessian of quantum-circuit functions in variational quantum algorithms. With hypothetical noiseless circuits, it is known from the prequel article that a proper optimization of statistical biases in numerical estimators can achieve lower mean-squared errors than analytical ones (namely, estimators derived from the unscaled parameter-shift rule). The key point is that sampling is required in variational algorithms, and the additional parameter degree of freedom in numerical estimators permits us to achieve optimal mean-squared errors that drop exponentially with the number of qubits, commensurate with the exponentially decaying gradient and Hessian magnitudes as a result of barren plateaus occurring in deep universal quantum circuits. This possibility is absent in analytical estimators. Such optimized numerical estimators are as easily computable as the commonly accepted analytical ones and should therefore be the preferred estimators if one considers the mean-squared error as the figure of merit for estimation accuracy.

With realistic noisy quantum circuits, we have shown here that, when one uses numerical estimators that are optimized for noiseless circuits to estimate gradient and Hessian components of noisy circuit functions, there exist nonzero sampling-copy-number regimes where these so-called naively optimized numerical estimators can still achieve lower mean-squared errors than analytical estimators. Under two physically realistic assumptions on both the quantum-circuit structure and noise channel as stated in Sec. IV A, we

explicitly show that the scaled parameter-shift estimators outperform the corresponding unscaled analytical ones within a sampling-copy-number range that increases exponentially with qubit number and also increases reciprocally with the total noise-channel error rate. These properties are carried over to practical channels modeling noisy two-qubit gates on a layered circuit *ansatz*.

These naively optimized numerical estimators have innate noise biases that do not result in faithful gradient and Hessian estimation even when the sampling-copy number becomes infinity, because they are strictly not meant for noisy quantum circuits. To resolve this problem, we proposed an experimentally operational error-mitigation technique that does not require the precise knowledge concerning the type of noise channel acting on the circuit; only the error rate is needed. This technique employs, again, the two assumptions about the circuit and noise channel and seeks to minimize the mean-squared error upper bounds of numerical schemes to obtain heuristically optimized estimators that are more compatible with noisy circuits. Indeed, we showed that the heuristically optimal scaled parameter-shift estimators not only completely eliminate noise biases under noise channels with constant error terms but also significantly reduce these noise biases when physically realistic circuit noise models are considered. The heuristically optimized finite-difference estimators, unfortunately, are just as noisily biased as their naively optimized counterparts and should be avoided.

The heuristic nature of the error-mitigation procedure introduced in this work originates from the two assumptions about the circuit unitary properties and noise channels. We emphasize, however, that since these assumptions are approximately aligned with moderately deep quantum circuits and realistic circuit noise channels, the corresponding heuristically optimal scaled parameter-shift estimators are consequently also relevant and interesting in practical situations.

From Lemma 2, the asymptotic relative performance between heuristic scaled parameter-shift and unscaled parameter-shift schemes drops exponentially with the number of qubits. This exponential decrease originates from the onset of barren plateaus, which is the limiting factor for a significant performance gap for large circuit sizes and depths. In relation to this, which is also the end message of the prequel article, we reiterate here that having a statistically accurate estimation scheme is *only the first* of many important steps towards the goal of trainable quantum circuits, and that this quest is by no means finished with this work. Much more efforts are required in seeking new initialization strategies and more expressive circuit *ansätze* to circumvent the barren plateau problem, or, more generally, the concentration-of-measure phenomenon in variational quantum algorithms.

ACKNOWLEDGMENTS

This work is supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (Grants No. RS-2023-00237959 and No. NRF-2022M3E4A1076099) via the Institute of Applied Physics at Seoul National University, and the Brain Korea 21 FOUR Project grant funded by the Korean Ministry of Education.

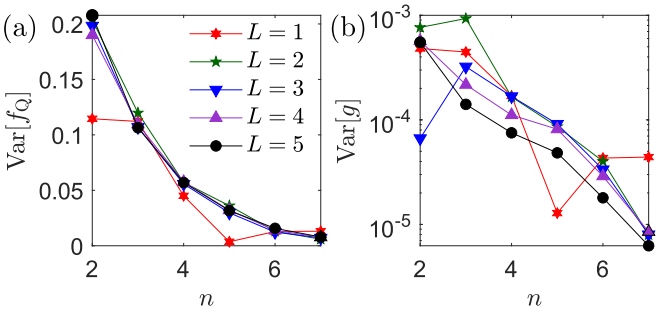

FIG. 6. Plots of (a) $\mathrm{Var}_\theta[f_\mathrm{Q}]$ and (b) $\mathrm{Var}_\theta[g]$ against n for $\eta_0 = 0.05$ and various L values under the CNOT-gate depolarizing channel. All curves are averaged over 1000 random sets of PEPQC parameters (Haar-distributed single-qubit unitary rotations).

APPENDIX A: REMARKS ON CONSTANT NOISE-CHANNEL ERROR TERMS

Figure 6 illustrates the behaviors of $\mathrm{Var}_\theta[f_\mathrm{Q}]$ and $\mathrm{Var}_\theta[g]$ with respect to n for the CNOT-gate depolarizing channel defined in (7) with a uniform error rate η_0. The decreasing variances over n is an indication of the concentration of measure phenomenon that occurs when the number of free parameters or system dimension tends to infinity [88–90]. The ratio r_var tends to increase with n, especially when L is large.

To show that the constant-g assumption is not bad even for general Pauli channels, we also look at the distribution of g for the case in which the noise-channel map $\mathcal{E}$ corresponds to a general Pauli channel where the 15-dimensional column $\eta_0 \widehat{=} (\eta_{0,12}, \eta_{0,13}, \ldots, \eta_{0,44})^\top$ per noisy CNOT gate has all entries that sum to some fixed η_0. The action is then given by

$$\begin{aligned} &U_{\mathrm{CNOT},jk}\rho_0 U_{\mathrm{CNOT},jk}^\dagger \\ &\quad\mapsto (1-\eta_0) U_{\mathrm{CNOT},jk}\rho_0 U_{\mathrm{CNOT},jk}^\dagger \\ &\qquad + \sum_{1\neq P_{jk}\in\mathcal{P}_2^{(j,k)}} \eta_{0,jk} P_{jk} U_{\mathrm{CNOT},jk}\rho_0 U_{\mathrm{CNOT},jk}^\dagger P_{jk}. \end{aligned} \tag{A1}$$

Figure 7 shows a very similar characteristic for g in which and that r_var increases with increasing n and L.

APPENDIX B: DERIVATIONS OF EQS. (11) AND (12)

We present the derivations of $\mathcal{D}_\mathrm{FD}(\partial f_\mathrm{Q})$ and $\mathcal{D}_\mathrm{SPS}(\partial f_\mathrm{Q})$. All other expressions may be obtained in a similar fashion. Starting with the general definition

$$\mathcal{D}_\cdot(\partial f_\mathrm{Q}) = \overline{\langle(\widehat{[\partial_\cdot] f_\mathrm{Q}}_\eta - \partial f_\mathrm{Q})^2\rangle}, \tag{B1}$$

where $\widehat{[\partial_\cdot] f_\mathrm{Q}}_\eta$ is the gradient estimator subjected to noise of error rate η (the subscripts μ and l will be dropped in this discussion). Note that

$$\mathcal{D}_\cdot(\partial f_\mathrm{Q}) = \underbrace{\overline{\langle(\widehat{[\partial_\cdot] f_\mathrm{Q}}_\eta - [\partial_\cdot] f_{\mathrm{Q}_\eta})^2\rangle}}_{\text{finite-copy error}} + \underbrace{\langle([\partial_\cdot] f_{\mathrm{Q}_\eta} - \partial f_\mathrm{Q})^2\rangle}_{\text{approximation error}} \tag{B2}$$

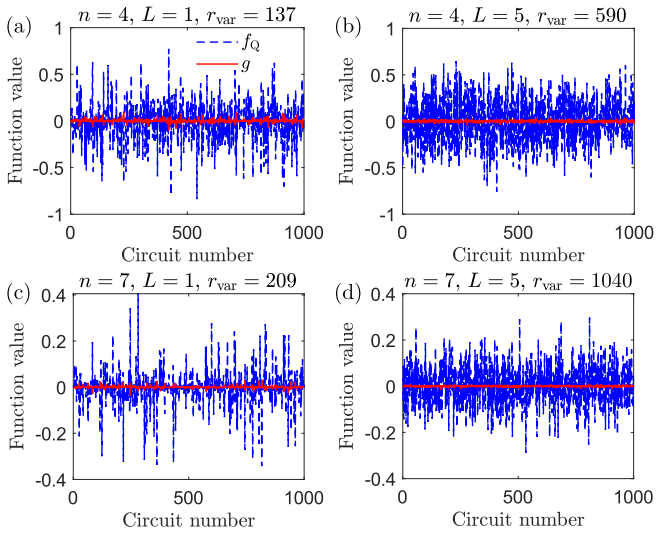

FIG. 7. Distributions of f_Q and g in $f_{Q,\eta}$ for PEPQCs with (a), (b) $n = 4$ and (c), (d) $n = 7$ qubits over 1000 sets of randomly generated PEPQC parameters (Haar-distributed single-qubit unitary rotations) in each figure panel. The ratio $r_{\text{var}} = \text{Var}_\theta[f_Q]/\text{Var}_\theta[g]$ is given in every panel. The Pauli-channel error rate is set at $\eta_0 = 0.05$ and η_0 is randomly chosen for each PEPQC parameter set. The overall error rates $\eta = 1 - (1 - \eta_0)^{nL}$ are (a) 0.185, (b) 0.642, (c) 0.302, and (d) 0.834. The respective observables are $O = X_1Y_2Z_3X_4$ and $O = X_1Y_2Z_3X_4Y_5Z_6X_7$ for $n = 4$ and 7.

is a sum of the finite-copy and approximation errors. Gradient estimators are formed by taking the difference between two translated circuit functions and dividing it by a scalar.

For FD, this scalar is twice the translation according to (2), so that the independence of the data collected for each translated function results in

$$\overline{\langle([\widehat{\partial_{\text{FD}}}]f_{Q_\eta} - [\partial_{\text{FD}}]f_{Q_\eta})^2\rangle} = \frac{1}{\epsilon^2}\left\{\left\langle\overline{\left[f_{Q_\eta}\widehat{\left(\theta + \frac{\epsilon}{2}\right)} - f_{Q_\eta}\left(\theta + \frac{\epsilon}{2}\right)\right]^2}\right\rangle + \left\langle\overline{\left[f_{Q_\eta}\widehat{\left(\theta - \frac{\epsilon}{2}\right)} - f_{Q_\eta}\left(\theta - \frac{\epsilon}{2}\right)\right]^2}\right\rangle\right\}. \tag{B3}$$

Since $f_{Q_\eta}(\theta) = \text{tr}\{\rho_{\theta,\eta}O\} = \sum_{k=0}^{d-1} o_k\langle k|\rho_{\theta,\eta}|k\rangle$, where $|k\rangle$ is an eigenket of O with eigenvalue o_k, $\widehat{f_{Q_\eta}(\theta)}$ can be defined as an unbiased estimator of $f_{Q_\eta}(\theta)$ inasmuch as

$$\widehat{f_{Q_\eta}(\theta)} = \sum_{k=0}^{d-1} o_k\nu_{k,\theta,\eta} = \frac{1}{N}\sum_{k=0}^{d-1} o_k n_{k,\theta,\eta}, \tag{B4}$$

with $\nu_{k,\theta,\eta} \to p_{k,\theta,\eta} = \langle k|\rho_{\theta,\eta}|k\rangle$ in the limit of large N, such that $\overline{\widehat{f_{Q_\eta}(\theta)}} = f_{Q_\eta}(\theta)$ as $\overline{\nu_{k,\theta,\eta}} = p_{k,\theta,\eta}$. So, using the identity

$$\overline{\nu_{k,\theta,\eta}\nu_{k',\theta,\eta}} - p_{k,\theta,\eta}p_{k',\theta,\eta} = \frac{1}{N}(\delta_{k,k'}p_{k,\theta,\eta} - p_{k,\theta,\eta}p_{k',\theta,\eta}) \tag{B5}$$

for the multinomial distribution, a traceless Pauli observable O implies that

$$\begin{aligned}
&\left\langle\overline{\left[f_{Q_\eta}\widehat{\left(\theta + \frac{\epsilon}{2}\right)} - f_{Q_\eta}\left(\theta + \frac{\epsilon}{2}\right)\right]^2}\right\rangle \\
&= \frac{1}{N}\left\langle\sum_{k,k'=0}^{d-1} o_k o_{k'}(\delta_{k,k'}p_{k,\theta+\epsilon/2,\eta} - p_{k,\theta+\epsilon/2,\eta}p_{k',\theta+\epsilon/2,\eta})\right\rangle \\
&= \frac{1}{N}\left[1 - \left\langle f_{Q_\eta}\left(\theta + \frac{\epsilon}{2}\right)^2\right\rangle\right] \\
&= \frac{1}{N}[1 - (1-\eta)^2\langle f_Q^2\rangle - \eta^2 g^2],
\end{aligned} \tag{B6}$$

where we made use of the constant-g and two-design assumptions—$\langle f_Q g\rangle = \langle f_Q\rangle g = 0$. The finite-copy error hence reads

$$\overline{\langle([\widehat{\partial_{\text{FD}}}]f_{Q_\eta} - [\partial_{\text{FD}}]f_{Q_\eta})^2\rangle} = \frac{2}{N\epsilon^2}[1 - (1-\eta)^2\langle f_Q^2\rangle - \eta^2 g^2]. \tag{B7}$$

The approximation error is straightforward to cope with using the same two assumptions:

$$\begin{aligned}
&\langle([\partial_{\text{FD}}]f_{Q_\eta} - \partial f_Q)^2\rangle \\
&= \left\langle\left\{\frac{1}{\epsilon}\left[f_{Q_\eta}\left(\theta + \frac{\epsilon}{2}\right) - f_{Q_\eta}\left(\theta - \frac{\epsilon}{2}\right)\right] - \partial f_Q\right\}^2\right\rangle \\
&= \left\langle\left\{\frac{1-\eta}{\epsilon}\left[f_Q\left(\theta + \frac{\epsilon}{2}\right) - f_Q\left(\theta - \frac{\epsilon}{2}\right)\right] - \partial f_Q\right\}^2\right\rangle \\
&= \left\langle\left[(1-\eta)\text{sinc}\left(\frac{\epsilon}{2}\right)\partial f_Q - \partial f_Q\right]^2\right\rangle \\
&= \left[1 - (1-\eta)\text{sinc}\left(\frac{\epsilon}{2}\right)\right]^2\langle|\partial f_Q|^2\rangle.
\end{aligned} \tag{B8}$$

What is left the assignment $N_T = 2N$.

By the same token, the SPS finite-copy error is given by

$$\begin{aligned}
&\overline{\langle([\widehat{\partial_{\text{SPS}}}]f_{Q_\eta} - [\partial_{\text{SPS}}]f_{Q_\eta})^2\rangle} \\
&= \frac{\lambda^2}{4}\left\{\left\langle\overline{\left[f_{Q_\eta}\widehat{\left(\theta + \frac{\pi}{2}\right)} - f_{Q_\eta}\left(\theta + \frac{\pi}{2}\right)\right]^2}\right\rangle \right. \\
&\quad \left. + \left\langle\overline{\left[f_{Q_\eta}\widehat{\left(\theta - \frac{\pi}{2}\right)} - f_{Q_\eta}\left(\theta - \frac{\pi}{2}\right)\right]^2}\right\rangle\right\} \\
&= \frac{\lambda^2}{2N}[1 - (1-\eta)^2\langle f_Q^2\rangle - \eta^2 g^2].
\end{aligned} \tag{B9}$$

Likewise, its approximation error is straightforwardly acquired through these steps:

$$\begin{aligned}
&\langle([\partial_{\text{SPS}}]f_{Q_\eta} - \partial f_Q)^2\rangle \\
&= \left\langle\left\{\frac{\lambda}{2}\left[f_{Q_\eta}\left(\theta + \frac{\pi}{2}\right) - f_{Q_\eta}\left(\theta - \frac{\pi}{2}\right)\right] - \partial f_Q\right\}^2\right\rangle \\
&= \left\langle\left\{\frac{(1-\eta)\lambda}{2}\left[f_Q\left(\theta + \frac{\pi}{2}\right) - f_Q\left(\theta - \frac{\pi}{2}\right)\right] - \partial f_Q\right\}^2\right\rangle
\end{aligned}$$

$$= \langle[(1-\eta)\lambda\partial f_{\rm Q} - \partial f_{\rm Q}]^2\rangle$$
$$= [1-(1-\eta)\lambda]^2\langle|\partial f_{\rm Q}|^2\rangle. \tag{B10}$$

The quadratic dependence in λ for the SPS MSEs in (12) permits the acquisition of explicit $N_*^{\rm NSPS}$ closed forms for $g=0$. When the total error rate of the noise channel is η, these *exact* formulas are

$$N_*^{\rm NSPS} = \frac{(d^2-1)h(d,\eta)}{2d^2\eta(1-\eta)}\ (\partial f_{\rm Q}\ \text{estimation}),$$
$$N_*^{\rm NSPS} = \frac{9(d^2-1)h(d,\eta)}{16d^2\eta(1-\eta)}\ (\partial\partial f_{\rm Q}\ \text{estimation}),$$
$$N_*^{\rm NSPS} = \frac{(d^2-1)^2h(d,\eta)}{d^4\eta(1-\eta)}\ (\partial\partial' f_{\rm Q}\ \text{estimation}),$$
$$h(d,\eta) = d + 4\eta + \eta^2(d-2) + \{4\eta(2-\eta)^2 + 4d\eta[2+\eta(1-\eta)(3-\eta)] + d^2[1+\eta(8-6\eta+\eta^3)]\}^{1/2}. \tag{B11}$$

Clearly, if $\eta\to 0$, $h(d,\eta)\to 2d$ and we once again arrive at (14).

APPENDIX C: BIASES IN NSPS, NFD, AND HFD ESTIMATORS

It is clear that using numerical estimators such as those of NSPS and NFD, which are optimized for noiseless quantum circuits, introduce noise biases, which are permanent systematic errors when used to estimate gradient and Hessian components for noisy circuit functions.

To quantify these noise biases for NSPS and HSPS, we take the constant-g and two-design (with the TDS condition) approximations and investigate things in the regime of large $N_{\rm T}$, where $\lambda_{\rm opt}\cong 1$ for all NSPS schemes. Hence, while the finite-copy errors all go as $1/N_{\rm T}$, the approximation errors respectively approach $\langle|\partial f_{\rm Q}|^2\rangle\eta^2$, $\langle|\partial\partial f_{\rm Q}|^2\rangle\eta^2$, and $\langle|\partial\partial' f_{\rm Q}|^2\rangle\eta^2$, which are the noise biases. On the other hand, from the results of (15), we have the asymptotic answer $\lambda_{{\rm opt},\eta}\cong 1/(1-\eta)>1$ for all HSPS schemes. It is then trivial to see that the approximation error approaches zero in the large-$N_{\rm T}$ limit. This reasoning works well so long as g is approximately constant.

A similar argument may be invoked to study the noise biases of NFD estimators defined by the optimal parameters $\epsilon_{\rm opt}$ meant for noiseless circuits. In (C8) of Ref. [59], these parameters, in the $N_{\rm T}\gg d$ limit are found to be

$$\epsilon_{\rm opt}(\partial f_{\rm Q}) \cong \left[\frac{1152d}{\langle(\partial f_{\rm Q})^2\rangle N_{\rm T}(d+1)}\right]^{1/6},$$
$$\epsilon_{\rm opt}(\partial\partial f_{\rm Q}) \cong \left[\frac{2592d}{\langle(\partial\partial f_{\rm Q})^2\rangle N_{\rm T}(d+1)}\right]^{1/8},$$
$$\epsilon_{\rm opt}(\partial\partial' f_{\rm Q}) \cong \left[\frac{2304d}{\langle(\partial\partial' f_{\rm Q})^2\rangle N_{\rm T}(d+1)}\right]^{1/8}. \tag{C1}$$

For such an astronomical $N_{\rm T}$, we return to the textbook optimality requirement that $\epsilon_{\rm opt}\to 0$, so that $\mathrm{sinc}(\epsilon_{\rm opt}/2)\to 1$. Notice that $\epsilon_{\rm opt}$ tends to zero much slower than $1/N_{\rm T}$, so that the finite-copy errors of (11) (evaluated with $\epsilon_{\rm opt}$ for NFD) still approach zero in a well-defined manner as $N_{\rm T}\to\infty$. Thus, the noise biases of NFD estimators are precisely those of NSPS estimators.

For the HFD estimators, it turns out that the approximation error can never be completely eliminated even in the limit of large $N_{\rm T}$. The straightforward reason is that when $\eta<1$, unlike $\lambda_{{\rm opt},\eta}$, which is to be greater than one for noise biases to vanish, the sinc functions entering the approximation errors of the FD estimators are all never greater than one. Therefore, the best these HFD estimators can do is to achieve noise biases equal to those of the NFD estimators in the asymptotic limit.

[1] I. Chuang and M. Nielsen, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, 2000).

[2] T. D. Ladd, F. Jelezko, R. Laflamme, Y. Nakamura, C. Monroe, and J. L. O'Brien, Quantum computers, Nature (London) **464**, 45 (2010).

[3] E. T. Campbell, B. M. Terhal, and C. Vuillot, Roads towards fault-tolerant universal quantum computation, Nature (London) **549**, 172 (2017).

[4] B. Lekitsch, S. Weidt, A. G. Fowler, K. Mølmer, S. J. Devitt, C. Wunderlich, and W. K. Hensinger, Blueprint for a microwave trapped ion quantum computer, Sci. Adv. **3**, e1601540 (2017).

[5] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell, B. Burkett, Y. Chen, Z. Chen, B. Chiaro, R. Collins, W. Courtney, A. Dunsworth, E. Farhi, B. Foxen, A. Fowler *et al.*, Quantum supremacy using a programmable superconducting processor, Nature (London) **574**, 505 (2019).

[6] H.-S. Zhong, H. Wang, Y.-H. Deng, M.-C. Chen, L.-C. Peng, Y.-H. Luo, J. Qin, D. Wu, X. Ding, Y. Hu, P. Hu, X.-Y. Yang, W.-J. Zhang, H. Li, Y. Li, X. Jiang, L. Gan, G. Yang, L. You, Z. Wang *et al.*, Quantum computational advantage using photons, Science **370**, 1460 (2020).

[7] Y. Wu, W.-S. Bao, S. Cao, F. Chen, M.-C. Chen, X. Chen, T.-H. Chung, H. Deng, Y. Du, D. Fan, M. Gong, C. Guo, C. Guo, S. Guo, L. Han, L. Hong, H.-L. Huang, Y.-H. Huo, L. Li, N. Li *et al.*, Strong quantum computational advantage using a superconducting quantum processor, Phys. Rev. Lett. **127**, 180501 (2021).

[8] L. K. Grover, A fast quantum mechanical algorithm for database search, *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '96 (Association for Computing Machinery, New York, 1996) pp. 212–219.

[9] P. W. Shor, Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer, SIAM J. Comput. **26**, 1484 (1997).

[10] R. Raussendorf and H. J. Briegel, A one-way quantum computer, Phys. Rev. Lett. **86**, 5188 (2001).

[11] A. Kitaev, Fault-tolerant quantum computation by anyons, Ann. Phys. (NY) **303**, 2 (2003).

[12] R. Raussendorf, J. Harrington, and K. Goyal, Topological fault-tolerance in cluster state quantum computation, New J. Phys. **9**, 199 (2007).

[13] A. Sehrawat, L. H. Nguyen, and B.-G. Englert, Test-state approach to the quantum search problem, Phys. Rev. A **83**, 052311 (2011).

[14] A. Montanaro, Quantum algorithms: An overview, npj Quantum Inf. **2**, 15023 (2016).

[15] J. Preskill, Quantum Computing in the NISQ era and beyond, Quantum **2**, 79 (2018).

[16] T. R. Bromley, J. M. Arrazola, S. Jahangiri, J. Izaac, N. Quesada, A. D. Gran, M. Schuld, J. Swinarton, Z. Zabaneh, and N. Killoran, Applications of near-term photonic quantum computers: Software and algorithms, Quantum Sci. Technol. **5**, 034010 (2020).

[17] 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).

[18] A. Finnila, M. Gomez, C. Sebenik, C. Stenson, and J. Doll, Quantum annealing: A new method for minimizing multidimensional functions, Chem. Phys. Lett. **219**, 343 (1994).

[19] T. Kadowaki and H. Nishimori, Quantum annealing in the transverse Ising model, Phys. Rev. E **58**, 5355 (1998).

[20] S. Aaronson and A. Arkhipov, The computational complexity of linear optics, *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, STOC '11 (Association for Computing Machinery, New York, 2011), pp. 333–342.

[21] S. Aaronson, A linear-optical proof that the permanent is #P-hard, Proc. R. Soc. London A **467**, 3393 (2011).

[22] C. S. Hamilton, R. Kruse, L. Sansoni, S. Barkhofen, C. Silberhorn, and I. Jex, Gaussian boson sampling, Phys. Rev. Lett. **119**, 170501 (2017).

[23] A. Trabesinger, Quantum simulation, Nat. Phys. **8**, 263 (2012).

[24] I. M. Georgescu, S. Ashhab, and F. Nori, Quantum simulation, Rev. Mod. Phys. **86**, 153 (2014).

[25] J. Biamonte, Universal variational quantum computation, Phys. Rev. A **103**, L030401 (2021).

[26] 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, Nat. Rev. Phys. **3**, 625 (2021).

[27] Y. Cao, J. Romero, J. P. Olson, M. Degroote, P. D. Johnson, M. Kieferová, I. D. Kivlichan, T. Menke, B. Peropadre, N. P. D. Sawaya, S. Sim, L. Veis, and A. Aspuru-Guzik, Quantum chemistry in the age of quantum computing, Chem. Rev. (Washington, DC, U. S.) **119**, 10856 (2019).

[28] S. Endo, Z. Cai, S. C. Benjamin, and X. Yuan, Hybrid quantum-classical algorithms and quantum error mitigation, J. Phys. Soc. Jpn. **90**, 032001 (2021).

[29] S. McArdle, S. Endo, A. Aspuru-Guzik, S. C. Benjamin, and X. Yuan, Quantum computational chemistry, Rev. Mod. Phys. **92**, 015003 (2020).

[30] 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, Nat. Commun. **5**, 4213 (2014).

[31] D. Wecker, M. B. Hastings, and M. Troyer, Progress towards practical quantum variational algorithms, Phys. Rev. A **92**, 042303 (2015).

[32] 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).

[33] E. Farhi, J. Goldstone, and S. Gutmann, A quantum approximate optimization algorithm, arXiv:1411.4028.

[34] 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).

[35] M. Schuld, I. Sinayskiy, and F. Petruccione, An introduction to quantum machine learning, Contemp. Phys. **56**, 172 (2015).

[36] M. Schuld and N. Killoran, Quantum machine learning in feature Hilbert spaces, Phys. Rev. Lett. **122**, 040504 (2019).

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

[38] P. Date and W. Smith, Quantum discriminator for binary classification, Sci. Rep. **14**, 1328 (2024).

[39] 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).

[40] T. Dutta, A. Pérez-Salinas, J. P. S. Cheng, J. I. Latorre, and M. Mukherjee, Single-qubit universal classifier implemented on an ion-trap quantum device, Phys. Rev. A **106**, 012411 (2022).

[41] 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).

[42] S. Shin, Y. S. Teo, and H. Jeong, Exponential data encoding for quantum supervised learning, Phys. Rev. A **107**, 012422 (2023).

[43] S. Shin, Y. S. Teo, and H. Jeong, Dequantizing quantum machine learning models using tensor networks, arXiv:2307.06937.

[44] S. E. Smart and D. A. Mazziotti, Quantum-classical hybrid algorithm using an error-mitigating n-representability condition to compute the Mott metal-insulator transition, Phys. Rev. A **100**, 022517 (2019).

[45] B. van Straaten and B. Koczor, Measurement cost of metric-aware variational quantum algorithms, PRX Quantum **2**, 030324 (2021).

[46] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Quantum circuit learning, Phys. Rev. A **98**, 032309 (2018).

[47] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, Evaluating analytic gradients on quantum hardware, Phys. Rev. A **99**, 032331 (2019).

[48] A. Mari, T. R. Bromley, and N. Killoran, Estimating the gradient and higher-order derivatives on quantum hardware, Phys. Rev. A **103**, 012405 (2021).

[49] D. Wierichs, J. Izaac, C. Wang, and C. Y.-Y. Lin, General parameter-shift rules for quantum gradients, Quantum **6**, 677 (2022).

[50] S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, 2009).

[51] J. Fiurášek, Maximum-likelihood estimation of quantum measurement, Phys. Rev. A **64**, 024102 (2001).

[52] J. Řeháček, Z. Hradil, E. Knill, and A. I. Lvovsky, Diluted maximum-likelihood algorithm for quantum tomography, Phys. Rev. A **75**, 042108 (2007).

[53] Y. S. Teo, H. Zhu, B.-G. Englert, J. Řeháček, and Z. Hradil, Quantum-state reconstruction by maximizing likelihood and entropy, Phys. Rev. Lett. **107**, 020404 (2011).

[54] S. Amari and S. Douglas, Why natural gradient? in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No. 98CH36181)* (1998), Vol. 2, pp. 1213–1216.

[55] S. Amari, Natural gradient works efficiently in learning, Neural Comput. **10**, 251 (1998).

[56] B. Koczor and S. C. Benjamin, Quantum natural gradient generalized to noisy and nonunitary circuits, Phys. Rev. A **106**, 062416 (2022).

[57] J. Stokes, J. Izaac, N. Killoran, and G. Carleo, Quantum natural gradient, Quantum **4**, 269 (2020).

[58] D. Wierichs, C. Gogolin, and M. Kastoryano, Avoiding local minima in variational quantum eigensolvers with the natural gradient optimizer, Phys. Rev. Res. **2**, 043246 (2020).

[59] Y. S. Teo, Optimized numerical gradient and hessian estimation for variational quantum algorithms, Phys. Rev. A **107**, 042421 (2023).

[60] P. Schonfeld, Best linear minimum bias estimation in linear regression, Econometrica **39**, 531 (1971).

[61] J. P. Romano and A. F. Siegel, *Counterexamples in Probability and Statistics* (Wadsworth & Brooks/Cole, Monterey, 1986).

[62] M. Hardy, An illuminating counterexample, arXiv:math/0206006.

[63] Y. Eldar, Minimum variance in biased estimation: Bounds and asymptotically optimal estimators, IEEE Trans. Signal Process. **52**, 1915 (2004).

[64] J. Shang, H. K. Ng, and B.-G. Englert, Quantum state tomography: Mean squared error matters, bias does not, arXiv:1405.5350.

[65] A. W. Harrow and R. A. Low, Random quantum circuits are approximate 2-designs, Commun. Math. Phys. **291**, 257 (2009).

[66] C. Dankert, R. Cleve, J. Emerson, and E. Livine, Exact and approximate unitary 2-designs and their application to fidelity estimation, Phys. Rev. A **80**, 012304 (2009).

[67] R. Cleve, D. W. Leung, L. Liu, and C. Wang, Near-linear constructions of exact unitary 2-designs, Quantum Inf. Comput. **16**, 721 (2016).

[68] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, Nat. Commun. **9**, 4812 (2018).

[69] A. Arrasmith, M. Cerezo, P. Czarnik, L. Cincio, and P. J. Coles, Effect of barren plateaus on gradient-free optimization, Quantum **5**, 558 (2021).

[70] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, Cost function dependent barren plateaus in shallow parametrized quantum circuits, Nat. Commun. **12**, 1791 (2021).

[71] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, Connecting ansatz expressibility to gradient magnitudes and barren plateaus, PRX Quantum **3**, 010313 (2022).

[72] C. Oh, L. Jiang, and B. Fefferman, On classical simulation algorithms for noisy boson sampling, arXiv:2301.11532.

[73] C. Oh, M. Liu, Y. Alexeev, B. Fefferman, and L. Jiang, Tensor network algorithm for simulating experimental Gaussian boson sampling, arXiv:2306.03709.

[74] Y. S. Teo, S. Shin, H. Kwon, S.-H. Lee, and H. Jeong, Virtual distillation with noise dilution, Phys. Rev. A **107**, 022608 (2023).

[75] D. Hangleiter and J. Eisert, Computational advantage of quantum random sampling, Rev. Mod. Phys. **95**, 035001 (2023).

[76] D. Aharonov, X. Gao, Z. Landau, Y. Liu, and U. Vazirani, A polynomial-time classical algorithm for noisy random circuit sampling, in *Proceedings of the 55th Annual ACM Symposium on Theory of Computing, STOC 2023* (Association for Computing Machinery, New York, 2023), pp. 945–957.

[77] A. Deshpande, P. Niroula, O. Shtanko, A. V. Gorshkov, B. Fefferman, and M. J. Gullans, Tight bounds on the convergence of noisy random circuits to the uniform distribution, PRX Quantum **3**, 040329 (2022).

[78] C. Oh, K. Noh, B. Fefferman, and L. Jiang, Classical simulation of lossy boson sampling using matrix product operators, Phys. Rev. A **104**, 022407 (2021).

[79] H. Qi, D. J. Brod, N. Quesada, and R. García-Patrón, Regimes of classical simulability for noisy Gaussian boson sampling, Phys. Rev. Lett. **124**, 100502 (2020).

[80] K. Noh, L. Jiang, and B. Fefferman, Efficient classical simulation of noisy random quantum circuits in one dimension, Quantum **4**, 318 (2020).

[81] R. García-Patrón, J. J. Renema, and V. Shchesnovich, Simulating boson sampling in lossy architectures, Quantum **3**, 169 (2019).

[82] M. Oszmaniec and D. J. Brod, Classical simulation of photonic linear optics with lost particles, New J. Phys. **20**, 092002 (2018).

[83] M. J. Bremner, A. Montanaro, and D. J. Shepherd, Achieving quantum supremacy with sparse and noisy commuting quantum computations, Quantum **1**, 8 (2017).

[84] G. Kalai and G. Kindler, Gaussian noise sensitivity and boson sampling, arXiv:1409.3093.

[85] D. Aharonov, M. Ben-Or, R. Impagliazzo, and N. Nisan, Limitations of noisy reversible computation, arXiv:quant-ph/9611028.

[86] Z. Puchała and J. Miszczak, Symbolic integration with respect to the haar measure on the unitary groups, Bull. Pol. Acad. Sci.: Tech. Sci. **65**, 21 (2017).

[87] A. A. Mele, Introduction to Haar measure tools in quantum information: A beginner's tutorial, arXiv:2307.08956.

[88] A. Barvinok, Measure concentration in optimization, Math. Program. **79**, 33 (1997).

[89] M. Ledoux, *The Concentration of Measure Phenomenon*, Mathematical Surveys and Monographs (Amer. Math. Soc., Providence, 2001), Vol. 89.

[90] M. P. Müller, D. Gross, and J. Eisert, Concentration of measure for quantum states with a fixed expectation value, Commun. Math. Phys. **303**, 785 (2011).