


## Efficient Estimation of Trainability for Variational Quantum Circuits

Valentin Heyraud<sup>1</sup>, Zejian Li<sup>1,2</sup>, Kaelan Donatella<sup>1</sup>, Alexandre Le Boité<sup>1</sup>, and Cristiano Ciuti<sup>1,\*</sup>

<sup>1</sup>Université Paris Cité, CNRS, Matériaux et Phénomènes Quantiques (MPQ), Paris F-75013, France

<sup>2</sup>The Abdus Salam International Center for Theoretical Physics (ICTP), Strada Costiera 11, Trieste I-34151, Italy

 (Received 20 February 2023; revised 7 September 2023; accepted 3 November 2023; published 4 December 2023)

Parameterized quantum circuits used as variational ansatzes are emerging as promising tools to tackle complex problems ranging from quantum chemistry to combinatorial optimization. These variational quantum circuits can suffer from the well-known curse of barren plateaus, which is characterized by an exponential vanishing of the cost-function gradient with the system size, making training unfeasible for practical applications. Since a generic quantum circuit cannot be simulated efficiently, the determination of its trainability is an important problem. Here we find an efficient method to compute the gradient of the cost function and its variance for a wide class of variational quantum circuits. Our scheme relies on our proof of an exact mapping from randomly initialized circuits to a set of Clifford circuits that can be efficiently simulated on a classical computer by virtue of the celebrated Gottesman-Knill theorem. This method is scalable and can be used to certify trainability for variational quantum circuits and explore design strategies that can overcome the barren-plateau problem. As illustrative examples, we show results with up to 100 qubits.

DOI: [10.1103/PRXQuantum.4.040335](https://doi.org/10.1103/PRXQuantum.4.040335)

### I. INTRODUCTION

Inspired by the success of machine-learning methods, variational quantum algorithms [1–3] have emerged as a promising way to harness the power of quantum computing in various domains ranging from quantum chemistry [4–6] to combinatorial optimization problems [7–9]. These algorithms use the output of parameterized quantum circuits as variational ansatzes, whose parameters are classically optimized through gradient-based methods.

Variational quantum circuits can suffer from trainability issues caused by the existence of barren plateaus [10], a limitation that has been extensively studied in the recent literature [11–38]. It is characterized by an exponential vanishing of the cost function’s gradient with the system size that makes training variational quantum circuits impossible for a large number of qubits. Barren plateaus can originate from various and fundamentally different phenomena. Their emergence was first shown in Ref. [10] for 2-designs (a random unitary transformation matching the Haar distribution up to the second moment). Recent works linked barren plateaus to the expressibility of the

ansatz [11] as well as noise [12] and entanglement. In particular, Ortiz Marrero *et al.* [13] showed that, for architectures that can be split into a hidden and a visible subsystem, such as quantum Boltzmann machines or feed-forward quantum neural networks, an excess of entanglement between the two subsystems would result in a highly mixed state for the visible subsystem. This can lead to a flat landscape for the cost function. The effect of the structure of the cost function on the appearance of barren plateaus was also investigated in other works [14,15], and it was shown that global cost functions are more prone to exhibit barren plateaus. Note that shallow models such as quantum kernel machines [39–43] and reservoir computing models [44–47], while often easier to train than variational quantum algorithms, might also suffer from trainability issues of a similar nature [48].

Numerous investigations have proposed strategies to address the barren-plateau issue. In the context of entanglement-induced barren plateaus, most strategies rely on limiting the amount of entanglement [16–20]. Other methods make use of tailored distributions of the initial circuit parameters and carefully designed circuit architectures [21–29]. Yet, only a handful of configurations offer trainability guarantees and robustness against barren plateaus [30,31]. Tackling this fundamental issue remains an important theoretical challenge.

In this paper, we propose an alternative approach to the problem by providing an efficient method to estimate the average gradients and their variance for a wide class of

\*cristiano.ciuti@u-paris.fr

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI.

variational quantum algorithms. By studying the quantum channel associated with a random single-qubit rotation, we prove that, under some simple conditions, the first and second moments can be expressed as mixed-unitary channels [49] composed of Clifford gates [50]. Upon some additional general assumptions for the random angle distribution, we demonstrate that this allows us to exactly map randomly initialized circuits composed of Clifford gates and parametrized rotations to an ensemble of Clifford circuits. Moreover, we prove that the obtained ensemble can be efficiently sampled to compute quantities of interest, such as the variance of the gradient or the average of the cost function over the initial random parameters. Making use of the celebrated Gottesman-Knill theorem [51,52], we analytically prove the efficiency of our method that can be implemented on a classical computer with a complexity scaling polynomially in both the number of variational parameters and the system size. In addition, we show some numerical experiments to illustrate our method on examples of random circuits and faithfully reproduce the exponential suppression of the variance first found in Ref. [10] with polynomial resources and for circuits acting on up to 100 qubits.

## II. THEORETICAL FRAMEWORK

### A. Variational problem

In variational quantum algorithms, a parameterized unitary transformation  $\hat{U}(\boldsymbol{\theta})$  acting on  $n$  qubits is used as a variational ansatz to achieve a task expressed as the minimization of a cost function

$$C(\boldsymbol{\theta}) = \text{Tr}[\hat{U}(\boldsymbol{\theta})\hat{\rho}\hat{U}^\dagger(\boldsymbol{\theta})\hat{O}] \quad (1)$$

for some observable  $\hat{O}$  and some initial  $n$ -qubit state  $\hat{\rho}$ . This formulation is general and encompasses typical tasks, such as the preparation of a target state  $|\psi\rangle$  (setting  $\hat{O} = -|\psi\rangle\langle\psi|$ ) or a ground-state search for some Hamiltonian  $\hat{\mathcal{H}}$  (setting  $\hat{O} = \hat{\mathcal{H}}$ ). The considered parameterized unitaries are typically composed of a succession of parameterized gates and fixed layers. Here we consider a generic ansatz of the form

$$\hat{U}(\boldsymbol{\theta}) = \prod_{i=1}^M \hat{U}_i(\theta_i)\hat{W}_i, \quad (2)$$

where each unitary  $\hat{U}_i(\theta_i) = e^{-i\theta_i\hat{P}_i/2}$  is a single-qubit rotation associated with a given Pauli operator  $\hat{P}_i \in \{\hat{X}, \hat{Y}, \hat{Z}\}$ , while the  $\hat{W}_k$  are fixed layers composed of a sequence of unparameterized gates that can act on multiple qubits. Upon absorbing Clifford gates in the fixed layers, we consider the rather general class of circuits based on  $Z$  rotations [53]. The unitary transformation  $\hat{U}(\boldsymbol{\theta})$  depends on  $M$  continuous parameters gathered in the vector  $\boldsymbol{\theta} =$

$(\theta_0, \dots, \theta_{M-1})$ . These rotation parameters can be optimized using classical gradient-descent techniques. The gradient of the cost function with respect to the  $k$ th parameter can be conveniently estimated using the parameter-shift rule [54,55]:

$$\partial_k C(\boldsymbol{\theta}) = \frac{1}{2} \left( C\left(\boldsymbol{\theta} + \frac{\pi}{2}\mathbf{e}_k\right) - C\left(\boldsymbol{\theta} - \frac{\pi}{2}\mathbf{e}_k\right) \right). \quad (3)$$

Here  $\mathbf{e}_k$  is the canonical vector along component  $k$ . It is worth noting that the  $\pm\pi/2$  shifts in parameter  $\theta_k$  can be factored out and seen as an extra Clifford gate added to the fixed layer  $\hat{W}_k$ . In fact, remarking that the phase gate  $\hat{S}$  can be written  $\hat{S} = e^{i\pi/4}e^{-i\pi\hat{Z}/2}$  and assuming that  $\hat{P}_k = \hat{Z}$ , we have

$$\begin{aligned} \hat{U}_k(\theta_k + \pi/2)\hat{W}_k &= e^{-i\theta_k\hat{Z}/2}e^{-i\pi\hat{Z}/2}\hat{W}_k \\ &= e^{-i\pi/4}\hat{U}_k(\theta_k)\hat{S}\hat{W}_k. \end{aligned} \quad (4)$$

We define  $\hat{W}_{k,\pm} = e^{\mp i\pi/4}\hat{S}\hat{W}_k$  such that we can write

$$\hat{U}_k(\theta_k \pm \pi/2)\hat{W}_k = \hat{U}_k(\theta_k)\hat{W}_{k,\pm}. \quad (5)$$

We denote by

$$\hat{U}_\pm(\boldsymbol{\theta}) = \hat{U}\left(\boldsymbol{\theta} \pm \frac{\pi}{2}\mathbf{e}_k\right) \quad (6)$$

the shifted unitaries appearing in the parameter-shift rule. From what precedes we have

$$\hat{U}_\pm(\boldsymbol{\theta}) = \prod_{i=1}^M \hat{U}_i(\theta_i)\hat{V}_{i,\pm} \quad (7)$$

with

$$\hat{V}_{i,\pm} = \begin{cases} \hat{W}_{k,\pm} & \text{if } i = k, \\ \hat{W}_i & \text{otherwise.} \end{cases} \quad (8)$$

### B. Unitary ensembles and $t$ -fold channels

To start the optimization process, the rotation angles are randomly initialized according to some probability distribution  $p(\boldsymbol{\theta})$ . The initialized circuit can then be represented by a unitary ensemble  $\mathbb{U} = \{\hat{U}, \mathbb{P}(\hat{U})\}$ , where  $\mathbb{P}$  is a probability measure on  $\mathbb{U}$ . One is often interested in computing averages of quantities that are polynomial of a given order  $t$  in the entries of  $\hat{U}$ . Such quantities can be completely determined by the knowledge of the  $t$ -fold channel [56]

$$\Phi_{\mathbb{U}}^{(t)}(\hat{\rho}) = \int_{\mathbb{U}} \hat{U}^{\otimes t} \hat{\rho} \hat{U}^{\dagger \otimes t} d\mathbb{P}(\hat{U}), \quad (9)$$

where  $\hat{\rho}$  is an initial state of  $t$  copies of the original  $n$ -qubit system. As an example, the expected value of the square of

the cost function at the initialization can be evaluated using  $\mathbb{E}_\theta[C(\theta)^2] = \text{Tr}[\Phi_\theta^{(2)}(\hat{\rho}^{\otimes 2})\hat{O}^{\otimes 2}]$ , where we denote by  $\Phi_\theta^{(t)}$  the  $t$ -fold channel associated with the unitary ensemble  $\{\hat{U}(\theta), \mathbb{R}^M \ni \theta \sim p(\theta)\}$ . In Appendix B we define *second-order quantities* (respectively *first-order quantities*) as quantities that can be obtained from the knowledge of the 2-fold (respectively 1-fold) channel. To give a concrete example,  $\mathbb{E}_\theta[C(\theta)^2]$  is a second-order quantity.

More generally, one can characterize the expressivity of a given ansatz by comparing its  $t$ -fold channels to those obtained for a Haar (uniform) distribution over the whole unitary group [11, 57, 58]. Unitary ensembles whose  $t$ -fold channels match the  $t$ -fold channels for the Haar measure, the so-called  $t$ -designs [59, 60], have played a crucial role in the original discovery of the barren-plateau phenomenon [10]. Moreover, in multiple cases random quantum circuits are approximate  $t$ -designs [61–63]. For instance, Harrow and Low [61] showed that quantum circuit composed of a polynomial number of gates randomly drawn from a universal set of two-qubit gates and applied to random pairs of qubits are approximate 2-designs. This result has been extended to cases where the gates are applied to nearest-neighbor qubits in Refs. [62, 63].

### C. Barren plateaus

For a unitary ensemble  $\mathbb{U}$  that describes parameterized ansatz  $\hat{U}(\theta)$  with random continuous parameters  $\theta$  and a possibly random architecture, a cost function  $C(\hat{U}(\theta))$  is said to exhibit a barren plateau if the probability of obtaining a gradient that deviates from zero by some  $\epsilon > 0$  vanishes exponentially with the system size  $n$ . More precisely,  $\mathbb{P}_{\mathbb{U}}(|\partial_k C| > \epsilon) \leq \mathcal{O}(\exp(-\alpha n))$  for some  $\alpha > 0$  [11]. As mentioned earlier, barren plateaus were first found for unitary ensembles forming 2-designs [10] and connections to expressivity [11], noise [12], entanglement [13, 16–20], and the degree of locality of the cost function [14, 15] were later discovered. In many cases, the average value of the gradient vanishes exactly, for instance when the rotation parameters are initialized uniformly in  $[-\pi, \pi]$ . However, this does not imply a vanishing of the gradient amplitude on average, and thus does not tell us much about the trainability of the model. In this unbiased case, the variance is a relevant quantity. Because of the Chebyshev inequality, one has  $\mathbb{P}_{\mathbb{U}}(|\partial_k C| > \epsilon) \leq \text{Var}[\partial_k C]/\epsilon^2$ , so that a vanishing variance implies the existence of a barren plateau. On the other hand, a nonvanishing variance guarantees large fluctuations of the initial gradient and thus a good initial trainability, independently of the gradient bias.

For variational quantum algorithms, the gradient is to be estimated through measurements realized on a hardware platform. As argued in Ref. [64], probing an exponentially small gradient requires an exponential precision on the measurements and thus an exponential number of experiments, which is prohibitive. Barren plateaus are often

seen as a quantum version of the vanishing gradient phenomenon in classical machine learning. The impact of vanishing gradients on the trainability of a classical deep neural network is discussed in Ref. [65]. The exact nature of barren plateaus was investigated further by Liu *et al.* [66] in a quantum machine-learning context, using a least-square loss function and relying on a neural tangent kernel formalism. In particular, the authors discussed the fundamental differences between barren plateaus and the classical vanishing gradient, and they showed that in a certain overparameterized regime the training procedure might be robust to noise. Let us note that variational quantum algorithms can suffer from other issues beyond barren plateaus, related for instance to the number of local minima of the loss landscape [67] or to the complexity associated with the classical optimization procedure [68]. Also, higher-order moments may help diagnose the trainability of variational quantum algorithms [69].

## III. RESULTS AND DISCUSSION

The main finding of this theoretical work is that, under some rather general assumptions on the distribution of the rotation parameter  $\theta$ , it is possible to map the 1-fold and 2-fold channels of a random rotation  $\hat{R}_Z(\theta)$  to a finite unitary ensemble of Clifford gates. Moreover, we prove that such a mapping allows us to estimate quantities of interest such as the gradient variance using only Clifford circuits. Finally, we illustrate our rigorous proofs through numerical experiments. The detailed mathematical proofs are presented in the Appendices A–C.

### A. Exact mapping and efficient sampling

As mentioned earlier, we focus on the class of variational quantum circuits composed of fixed Clifford gates alternated with single-qubit parameterized rotations along the  $X$ ,  $Y$ , or  $Z$  directions, such as that depicted in Fig. 1. As explained in Sec. II A, we restrict our study to rotations along  $Z$ , as we can obtain the cases of rotations along  $Y$  and  $X$  by adding extra Clifford gates to the different fixed layers of the considered ansatz. Let us consider a rotation along the  $Z$  axis with a distribution that is symmetric about the  $\theta = 0$  angle [70]. We show in Appendix A that the 1-fold channel corresponding to a first-order average can be written as a convex sum of the unitary channels associated with the identity and the Pauli- $Z$  gates, as schematically represented on the upper part of Fig. 2. Note that this result has been derived and used in Ref. [71] in the case of a uniform probability distribution for  $\theta$  in order to analyze a variational ansatz through the lens of ZX calculus.

To compute the 1-fold channel for the randomly initialized ansatz of the form given in Eq. (2) with independent rotation parameters, one can simply compose the 1-fold channels associated with each rotation, intertwined with the unitary channels associated with the fixed gates  $\hat{W}_k$ .

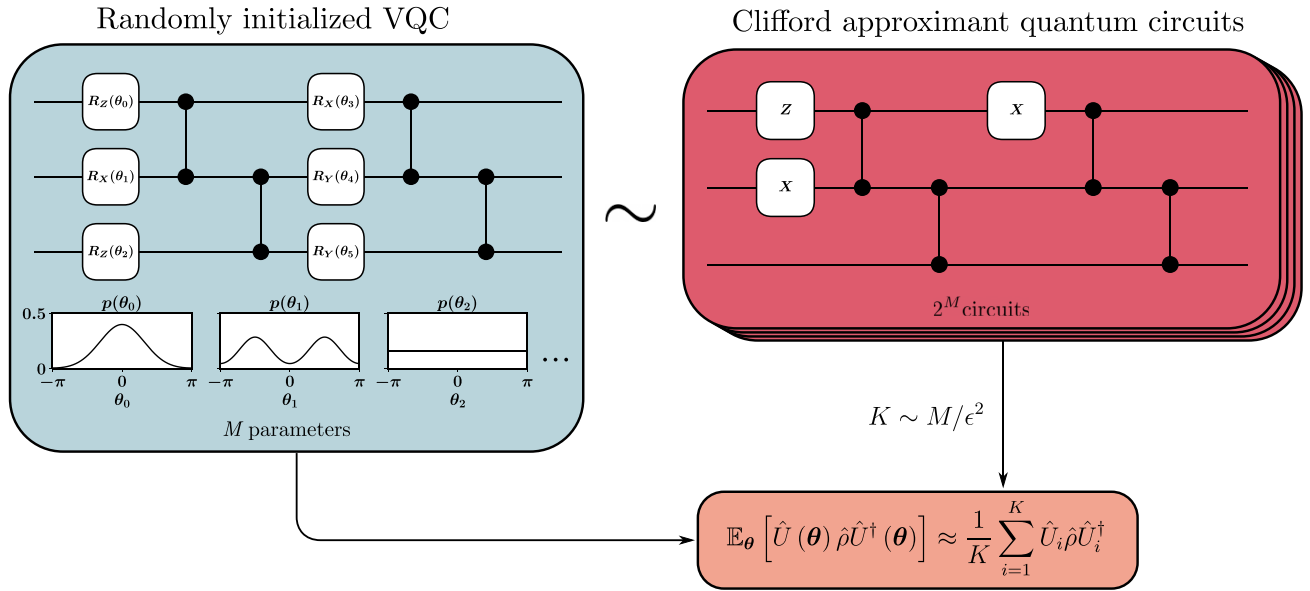


FIG. 1. Schematic representation of the mapping from a variational quantum circuit (VQC) with random parameters to Clifford approximant circuits for first-order quantities (quantities that require only knowledge of the rotation 1-fold channels to be computed; see Appendix B). For a circuit with  $M$  parameters, a sample size of the order of  $M/\epsilon^2$  is enough to get an approximation of the average on the initial circuit with a precision  $\epsilon$  on the observable mean values (see Appendix C).

We find that the 1-fold channel of the ansatz is a convex sum of  $2^M$  Clifford unitary channels, where  $M$  is the number of rotations. One can view this convex sum as an average over a finite ensemble of Clifford approximant circuits. Examples of such circuits are provided in Appendix F for a simple architecture similar to that in Fig. 1. Although the number of Clifford approximant circuits in this ensemble is exponential in the number of parameters, we show in Appendix C 2 that a number of samples polynomial in  $M/\epsilon^2$  are sufficient to approximate the average of an observable expectation value (or, more generally, of any first-order quantity) to any desired precision  $\epsilon$ . This result relies on a classical concentration argument, and is schematically represented in Fig. 1 for a simple circuit at the first order. From this, one can estimate the expectation value of the gradient, as it suffices to replace  $\hat{U}(\theta)$  by  $\hat{U}_{\pm}(\theta)$  (as defined in Sec. II A) in the 1-fold channel definition to obtain the expectation of  $C(\theta \pm \pi/2)$ . This gives the expectation of the gradient thanks to the parameter-shift rule.

We prove in Appendix A that the 2-fold channel associated with a random  $Z$  rotation is also a linear combination of Clifford channels, provided that the probability distribution is an even function of  $\theta$ . This result is depicted in Fig. 2(b). To obtain this mapping, we use the Choi representation of quantum channels [49]. The Choi operators representing unitary quantum channels given by tensor products of  $Z$ -rotation gates are diagonal. Hence, one

can represent these channels by the diagonal coefficients of their associated Choi operators. Using this representation and the linearity of the expectation, we obtain a tractable representation of the average 2-fold channel of a  $Z$  rotation. The decomposition is then derived by solving a linear system of equations obtained by identifying the coefficients of the previous representation for the different channels involved. When the inequalities  $\mathbb{E}_{\theta}[f_{+}(\theta)] \geq 0$  and  $\mathbb{E}_{\theta}[f_{-}(\theta)] \geq 0$  with  $f_{\pm}(\theta) = \cos \theta (\cos \theta \pm 1)$  are satisfied, the previous linear combination is in fact a convex sum. Equivalently, the 2-fold channel of a  $Z$  rotation with an even angular probability distribution is a Clifford mixed-unitary channel if

$$\mathbb{E}_{\theta}[\cos^2 \theta] \geq |\mathbb{E}_{\theta}[\cos \theta]|. \quad (10)$$

The zeros of  $f_{+}, f_{-}$  are the angles  $\{k\pi/2, k \in \mathbb{Z}\}$  for which  $\hat{R}_Z(\theta)$  matches a Clifford gate (see Appendix A). Indeed, if the distribution of  $\theta$  is a convex sum of Dirac distributions at these angles, the average over  $\theta$  becomes a discrete average over the corresponding Clifford unitaries. Hence, the associated 2-fold channel is indeed a convex sum of Clifford channels. One can also verify that the previous conditions are satisfied for distributions that are both even with respect to angle  $\theta$  and  $\pi$  periodic. For example, the uniform distribution is included. In the case of a centered Gaussian distribution, the previous conditions are satisfied if and only if the corresponding width is large enough.

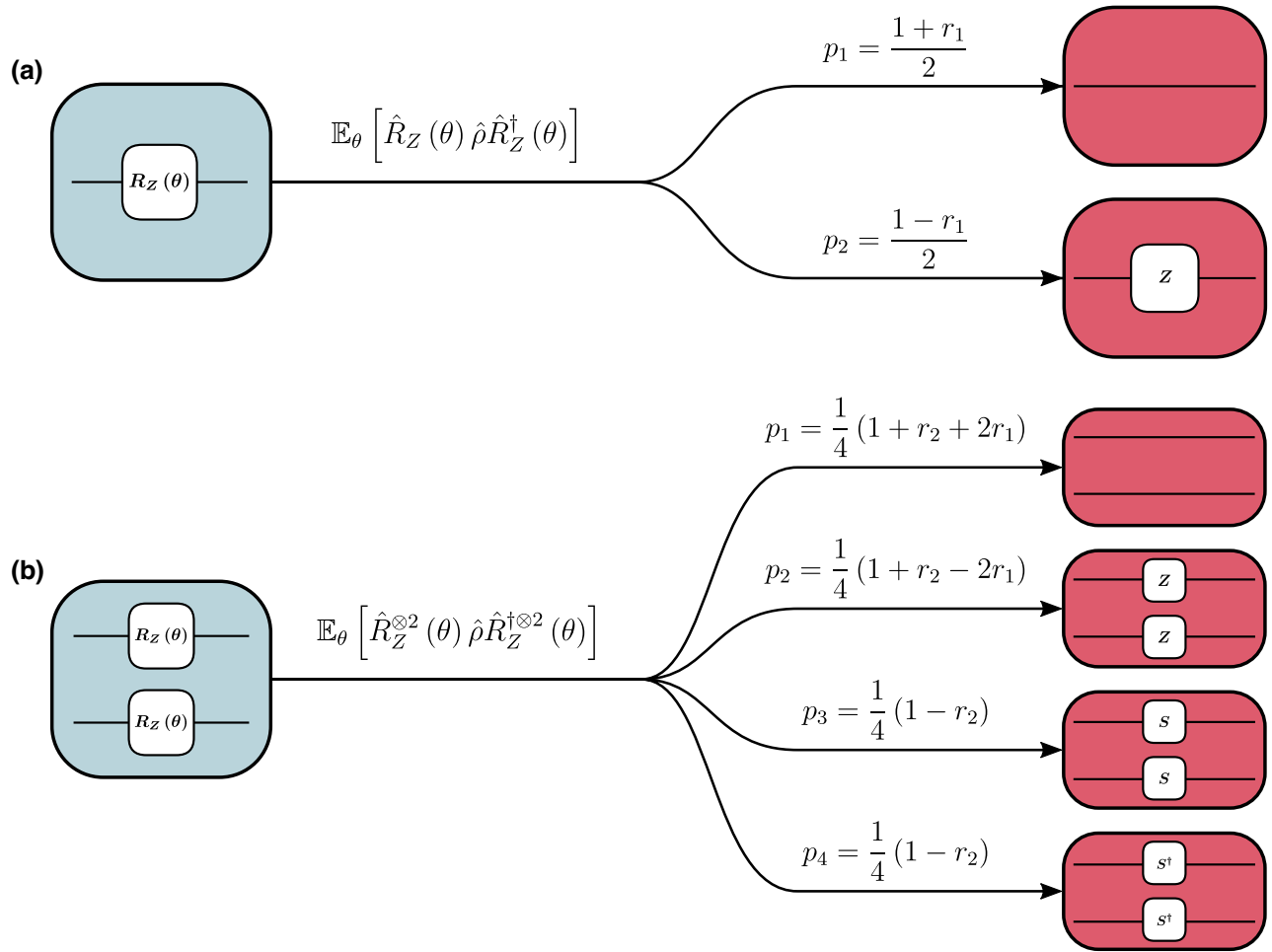


FIG. 2. Schematic representation of the mapping rules from  $Z$  rotations with a random parameter to unitary ensembles composed of Clifford gates. Panels (a) and (b) are respectively for first- and second-order averages. The mapping here is for probability distributions that are even with respect to  $\theta$ : we define  $r_1 = \mathbb{E}_\theta[\cos(\theta)]$  and  $r_2 = \mathbb{E}_\theta[\cos(2\theta)]$ . The coefficients  $p_i$  are the probabilities dictating how the corresponding Clifford circuits are sampled.

Provided the distributions of the rotation angles satisfy the conditions discussed above, the scheme can be extended to the second order, allowing one to approximate second-order quantities, such as the average of the squared cost function  $\mathbb{E}_\theta[C(\boldsymbol{\theta})^2]$ , using a set of Clifford approximant circuits. By the parameter-shift rule, the expectation of the squared gradient can be calculated from knowledge of four quantities of the form  $\mathbb{E}_\theta[C(\boldsymbol{\theta} \pm (\pi/2)\mathbf{e}_k)C(\boldsymbol{\theta} \pm (\pi/2)\mathbf{e}_k)]$ . The latter can be estimated with Clifford approximants by replacing the  $\hat{U}^{\otimes 2}$  term in the definition of the 2-fold channel by  $\hat{U}_\pm \otimes \hat{U}_\pm$ . Hence, the scheme covers the estimation of the gradient variance. Note that at the second order, the approximant circuits are obtained by replacing the rotation 2-fold channels by one of the four two-qubit Clifford gates depicted in Fig. 2, yielding an ensemble of  $4^M$  possible Clifford circuits. As for first-order quantities, a number of samples scaling linearly in  $M$  are enough to guarantee convergence. These rigorous results are summarized in the following

theorem, whose detailed proof is shown in Appendices A and C.

*Theorem 1.*—For a variational ansatz composed of fixed Clifford gates and of  $M$  parameterized rotations along the X, Y, or Z direction, if the random variational parameters  $(\theta_1, \dots, \theta_M)$  are independent and symmetric with respect to one of the Clifford angles, i.e.,  $\in \{0, \pi/2, \pi, 3\pi/2\}$ , then, for any error  $\epsilon > 0$  and a probability  $1 - \delta$  to meet such accuracy, any first-order quantity can be computed using

$$K \geq O\left(\frac{M}{\epsilon^2} \log\left(\frac{2}{\delta}\right)\right)$$

Clifford approximant circuits. Moreover, if the distribution of  $\theta_i$  satisfies the inequality

$$\begin{aligned} & \mathbb{E}_{\theta_i}[\cos^2(\theta_i - \mathbb{E}_{\theta_i}[\theta_i])] \\ & \geq |\mathbb{E}_{\theta_i}[\cos(\theta_i - \mathbb{E}_{\theta_i}[\theta_i])]| \quad \text{for all } i \in \{1, \dots, M\} \end{aligned}$$

then the same holds for any second-order quantity.

Finally one makes use of the Gottesman-Knill theorem, which states that, for a Clifford unitary  $\hat{U}$  and an observable  $\hat{O}$  acting nontrivially on  $N_O$  qubits, the expectation value  $\text{Tr}[|0\rangle\langle 0|^{\otimes n} \hat{U}^\dagger \hat{O} \hat{U}]$  can be classically computed with a polynomial complexity in both  $n$  and  $N_O$ . Our method inherits this complexity, and, in particular, we can classically estimate the gradient expectation and variance with a polynomial complexity in  $n$ ,  $N_O$ , and  $M$ .

In Appendix D we extend the scheme presented above for the 2-fold channels to the case where the distribution of the random angle does not satisfy the convex condition of Eq. (10). In that case, it is still possible to use Clifford approximant circuits to estimate second-order quantities, but this comes at the price of an exponential complexity in the number of variational parameters  $M$ . This result is based on a sampling scheme proposed by Piveteau *et al.* [72]. In that context, the method allows one to trade an exponential complexity in the system size for an exponential complexity in the number of variational parameters  $M$ .

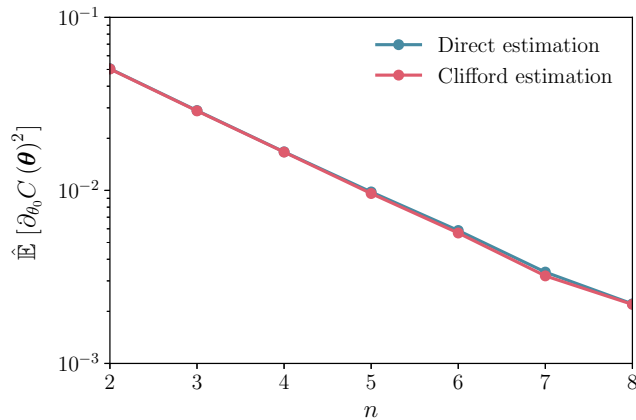


FIG. 3. Estimated average of the squared gradient of the cost function with respect to the first variational parameter versus the number of qubits  $n$ . We emphasize that derivatives with respect to the other angles  $\theta_k$  give similar results (not shown). The results are for random circuits composed of a single layer of gates, with one rotation per qubit. Such rotations are randomly chosen among  $R_X, R_Y, R_Z$ . The rotation layer is followed by a layer of alternated CZ gates (note that this is the same type of architecture as that represented in Fig. 1). The random rotation angles are independent and follow the uniform probability distribution on the interval  $[0, 2\pi]$ . In order to get the estimation, we randomly sampled 500 different circuit architectures. For each gate architecture, we computed the average of the squared gradient, assuming a uniform distribution of the rotation angles, using both a direct estimation and our method based on the mapping to Clifford approximant circuits. In particular, we sampled 500 vectors of angle parameters for the direct estimation and 500 Clifford circuits for our method. Note that, for the uniform distribution, the average gradient vanishes, and thus estimating the squared gradient is equivalent to estimating the gradient variance.

In Appendix E, we also present the extension of our scheme to the general case of  $N$ -fold channels. We prove that the  $N$ -fold channel associated with random  $Z$  rotations can be decomposed as a real sum of Clifford unitary channels. From this decomposition, we derive a condition for the  $N$ -fold channel to be a convex sum of Clifford unitaries by imposing the coefficients of the combination to be positive. However we show that the obtained decomposition is not unique, so that the derived condition is sufficient but not necessary. Finding a sufficient and necessary condition on the distribution of a random angle that guarantees that the corresponding  $N$ -fold channels are Clifford mixed unitary channels remains an open problem.

## B. Numerical simulations

To illustrate the applications of our exact mapping and the ensuing estimation method, we have performed numerical experiments on concrete examples. Let us consider a simple variational quantum circuit composed of layers of single-qubit rotations along either the  $X$ ,  $Y$ , or  $Z$  axes, alternated with fixed layers of control- $Z$  gates. Such an ansatz is shown for three qubits in Fig. 1. We further assume that the rotation angles are independent and identically distributed according to the uniform law over  $[0, 2\pi]$ . Moreover, we assume that the cost function is of the form in Eq. (1) with  $\hat{O} = |0\rangle\langle 0|^{\otimes n}$ .

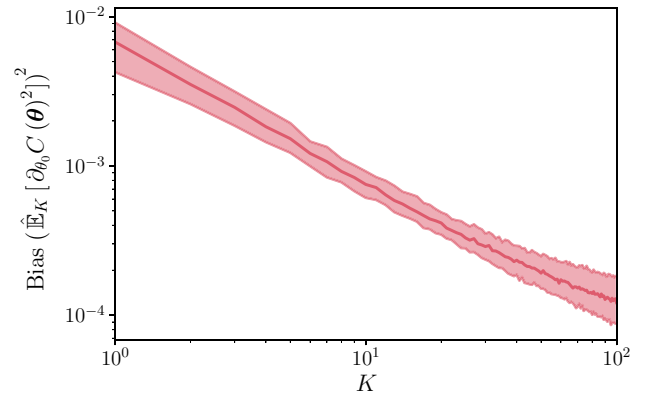


FIG. 4. Squared statistical bias of the estimator considered in Fig. 3 for random circuits with  $n = 5$  qubits versus the number of Clifford approximant circuits  $K$ . The results have been obtained with 500 randomly drawn circuit architectures. For each sample size  $K$ , we consider a bootstrap batch of 100 estimators (each estimator is obtained by sampling  $K$  circuits from a set of 2000 Clifford approximant circuits for each choice of the rotation direction). Then, for each  $K$ , the statistical bias is derived from the bootstrap batch. The estimator's true expected value is provided by the direct estimation of the average squared gradient with 4000 samples. The shaded area corresponds to the interval between the 20th and 80th percentiles of the estimated biases for the 500 random architectures.

We consider these architectures with random directions of the rotation gates. Up to a different fixed first layer, such random circuits have been showed to exhibit barren plateaus in Ref. [64]. Note that in this particular case the averaging was done on both the rotation angles and the rotation directions. Here we reproduce this result using Clifford approximants. To do so, we sample both the exact circuit architecture by randomly selecting the rotation directions uniformly from  $\{X, Y, Z\}$ , and then we either sample the rotation angles directly or we sample a Clifford approximant circuit. For a uniform distribution, we have, for all  $k \in \mathbb{Z}$ ,  $\mathbb{E}_\theta[\cos k\theta] = 0$ , so the sampling of the replacement Clifford gates is uniform (as represented in Fig. 2 for  $r_1 = r_2 = 0$ ). Moreover, by the parameter-shift rule [Eq. (3)], it is clear that, for uniformly distributed rotations, the average gradient is analytically zero; thus, it suffices to estimate the average of the squared gradient as  $\text{Var}_\theta [\partial_k C(\theta)] = \mathbb{E}_\theta[\partial_k C(\theta)^2]$ .

In Fig. 3 the estimations of the average squared gradient using either direct evaluations or by sampling Clifford approximants are shown. Note that the average is taken over both the random rotation angles and the variable architecture (i.e., the random direction of the rotation gates). The estimation obtained from Clifford approximants accurately matches the direct estimation and the average squared gradient vanishes exponentially with the number of qubits, as expected. In addition, the evolution of the bias of the Clifford estimation with the number of approximant circuits  $K$  is shown in Fig. 4. The bias decreases polynomially with  $K$ . As appears in Fig. 5, the same trend holds for the variance of the Clifford estimators. These results are in agreement with the analytical scaling derived in Appendix C.

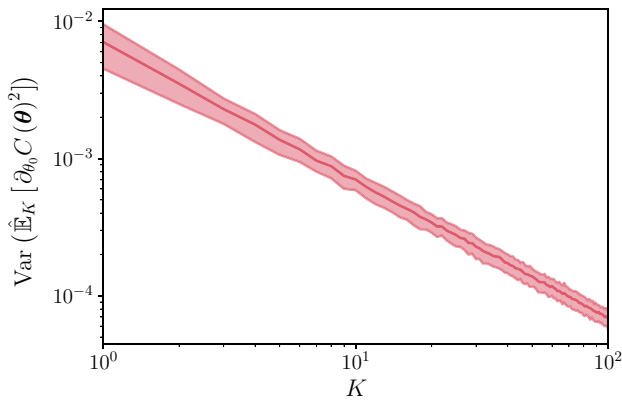


FIG. 5. Variance of the estimator of the expected squared gradient with respect to the first parameter  $\theta_0$  versus the number of Clifford approximant circuits  $K$ . Same type of random circuits as in Fig. 3 with  $n = 5$  qubits. We have used the same bootstrap procedure as in Fig. 4. The shaded area corresponds to the interval between the 20th and 80th percentiles of the estimated biases for 500 random architectures.

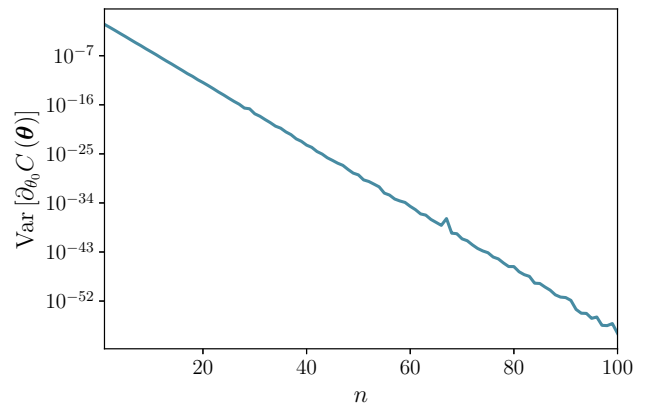


FIG. 6. Estimated variance of the gradient of the cost function with respect to the first variational parameter versus the number of qubits  $n$ . Each variance is estimated using  $10^5$  Clifford circuits. The circuits and cost function used are the same as those used for Fig. 3.

Using our method, we also reproduced the results showing the exponential suppression of the gradient variance presented in Ref. [64] for up to 100 qubits. The results are presented in Fig. 6.

Finally, we illustrate the impact of an ansatz architecture on its trainability by evaluating the gradient variances

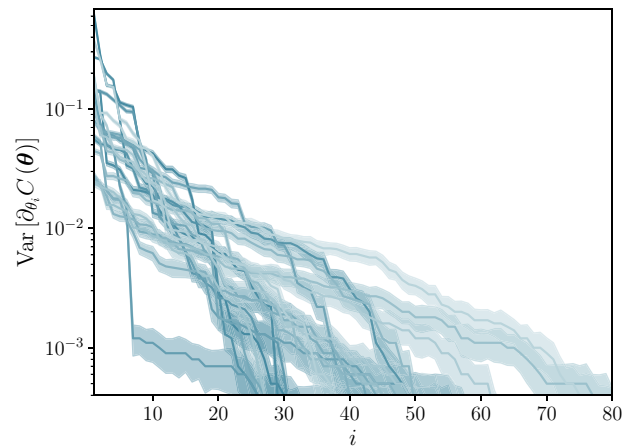


FIG. 7. Variance of the cost-function gradient with respect to the variational parameters in decreasing order and for 20 random circuits. The circuits are acting on  $n = 40$  qubits. Each circuit is composed of ten layers. For each layer  $l$ , a number  $m_l$  is randomly chosen in the interval  $[0, n]$ , and the layer is then built by first applying  $m_l$  random single-qubit rotation gates to randomly chosen qubits, and then applying a series of entangling gates of the same type and with the same arrangement as those considered in Fig. 3. For each circuit and each parameter  $\theta_i$ , the gradients are estimated using  $10^4$  Clifford approximant circuits. The Hamiltonian of the cost function is a sum of ten random Pauli strings. Note that each curve corresponds to a different architecture.

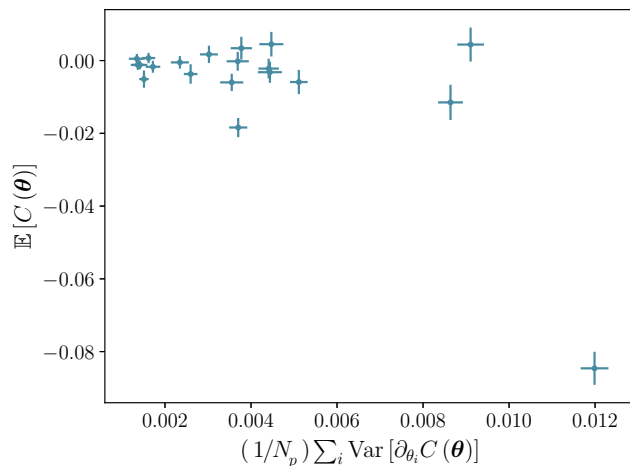


FIG. 8. Mean value of the cost function versus variance of the cost-function gradient for 20 random circuits and 40 qubits. Each point represents a circuit, and the circuits considered are the same as those of Fig. 7. The  $x$  axis is the average of the variances of the gradients with respect to the  $N_p$  circuit rotation parameters  $\theta_i$ ,  $i = 1, \dots, N_p$ .

for a set of randomly drawn ansatzes and for a given cost-function Hamiltonian. We consider random circuits acting on 40 qubits, and a Hamiltonian composed of a sum of ten randomly chosen Pauli strings. As for the results presented in Fig. 3, the considered variational circuits are composed of layers of rotations alternated with fixed entangling layers. Here we considered circuits with ten layers. For each rotation layer, a random subset of rotations is replaced by identity gates (see Fig. 7 for details). The variances of the cost-function partial derivatives with respect to the ansatz parameters are shown in Fig. 7, and Fig. 8 shows the average variance of the gradients versus the mean value of the cost function for different random circuits. These results show that, for a given cost function, modifying the ansatz architecture has a strong effect on the trainability. We therefore believe that our method may be used to systematically examine such effects for large systems with a reasonable cost, hence guiding the design of better variational ansatzes.

#### IV. CONCLUSIONS AND PERSPECTIVES

In this paper we presented a classically efficient method to estimate first- and second-order expectation values for a large class of randomly initialized variational quantum circuits. This includes estimating the average gradient of the cost function and its variance, which can be used to estimate the trainability. Our method applies to the large class of circuits whose architecture is composed of fixed Clifford gates and single-qubit parameterized rotations, provided that the rotation angles are independent and that their distributions are symmetric with respect to an

angle  $\theta_0 \in \{k\pi/2, k \in \mathbb{Z}\}$  and satisfy  $\mathbb{E}_\theta[\cos^2(\theta - \theta_0)] \geq |\mathbb{E}_\theta[\cos(\theta - \theta_0)]|$ . The method relies on an exact mapping of randomly initialized variational quantum circuits to ensembles of Clifford circuits and on the Gottesman-Knill theorem. We provide rigorous convergence guarantees, and, in particular, we show that the complexity of the method scales polynomially in both the system size and the number of parameters of the considered ansatz. We investigated the generalization of the proposed scheme to the case of  $N$ -fold channels, and showed that the  $N$ -fold average of random  $Z$  rotations can be expressed as a real combination of Clifford unitaries. However, such a decomposition is not unique, and finding a sufficient and necessary condition for the considered  $N$ -fold channel to be a Clifford mixed-unitary channel remains an open problem. Solving this problem is of great interest as it could allow us to generalize the scheme presented in this work to ansatzes with correlated variational parameters.

We believe that such a tool will prove very useful in future applications, as it could be employed to conduct classical optimization of architectures and initialization of large-scale variational quantum circuits. As the absence of barren plateaus can be guaranteed by a large enough variance of the gradient, regardless of the exact origin of the potential barren plateaus, this method could be used to certify trainability for a system with a very large number of qubits.

#### ACKNOWLEDGMENTS

This work was supported by Region Île-de-France in the framework of the Domaine d'Intérêt Majeur (DIM) Science et Ingénierie en Région Île-de-France pour les Technologies Quantiques (SIRTEQ). This work was granted access to the High Performance Computing (HPC) resources of Très Grand Centre de Calcul (TGCC) under Allocation No. 2022-A0120512462 made by Grand Equipement National de Calcul Intensif (GENCI). We would like to thank Zakari Denis for helpful discussions during the early stages of this work.

#### APPENDIX A: 1-FOLD AND 2-FOLD CHANNELS OF A RANDOM $Z$ ROTATION

##### 1. 1-fold channel

Here we give the expression of the 1-fold channel for a single-qubit rotation around the  $Z$  axis. The rotations around the  $X$  and  $Y$  axes can then be obtained by combination with Hadamard and phase gates. Let us define  $\hat{\Pi}_0 := |0\rangle\langle 0|$  and  $\hat{\Pi}_1 := |1\rangle\langle 1|$ :

$$\hat{R}_Z(\theta) = e^{-i\theta/2} \hat{\Pi}_0 + e^{i\theta/2} \hat{\Pi}_1, \quad (\text{A1})$$

$$\begin{aligned} \hat{R}_Z(\theta) \hat{\rho} \hat{R}_Z^\dagger(\theta) &= \hat{\Pi}_0 \hat{\rho} \hat{\Pi}_0 + \hat{\Pi}_1 \hat{\rho} \hat{\Pi}_1 + e^{i\theta} \hat{\Pi}_1 \hat{\rho} \hat{\Pi}_0 \\ &\quad + e^{-i\theta} \hat{\Pi}_0 \hat{\rho} \hat{\Pi}_1. \end{aligned} \quad (\text{A2})$$



Thus,

$$\begin{aligned} \mathbb{E}_\theta[\hat{R}_Z(\theta)\hat{\rho}\hat{R}_Z^\dagger(\theta)] &= \hat{\Pi}_0\hat{\rho}\hat{\Pi}_0 + \hat{\Pi}_1\hat{\rho}\hat{\Pi}_1 + \mathbb{E}_\theta[e^{i\theta}]\hat{\Pi}_1\hat{\rho}\hat{\Pi}_0 \\ &\quad + \mathbb{E}_\theta[e^{-i\theta}]\hat{\Pi}_0\hat{\rho}\hat{\Pi}_1. \end{aligned} \quad (\text{A3})$$

We recognize the characteristic function of the distribution of  $\theta$ , namely,

$$\phi(t) := \mathbb{E}_\theta[e^{it\theta}].$$

Assuming that this probability distribution is even in  $\theta$ , we have  $\phi(t) \in \mathbb{R}$  for all  $t$  and we can define  $r_1 = \phi(1) = \phi(1)^* = \phi(-1)$ . As we have  $\mathbb{1} = \hat{\Pi}_0 + \hat{\Pi}_1$  and  $\hat{Z} = \hat{\Pi}_0 - \hat{\Pi}_1$ , we get

$$\begin{aligned} \hat{\rho} &= (\hat{\Pi}_0\hat{\rho}\hat{\Pi}_0 + \hat{\Pi}_1\hat{\rho}\hat{\Pi}_1) + (\hat{\Pi}_1\hat{\rho}\hat{\Pi}_0 + \hat{\Pi}_0\hat{\rho}\hat{\Pi}_1), \\ \hat{Z}\hat{\rho}\hat{Z} &= (\hat{\Pi}_0\hat{\rho}\hat{\Pi}_0 + \hat{\Pi}_1\hat{\rho}\hat{\Pi}_1) - (\hat{\Pi}_1\hat{\rho}\hat{\Pi}_0 + \hat{\Pi}_0\hat{\rho}\hat{\Pi}_1), \end{aligned} \quad (\text{A4})$$

and, hence,

$$\frac{1+r_1}{2}\hat{\rho} + \frac{1-r_1}{2}\hat{Z}\hat{\rho}\hat{Z} = \mathbb{E}_\theta[R_Z(\theta)\hat{\rho}R_Z^\dagger(\theta)]. \quad (\text{A5})$$

This is indeed a convex sum of Clifford channels under the condition that  $r_1 \in [-1, 1]$ , which is always satisfied. For distributions that are symmetric with respect to a Clifford angle  $\in \{k\pi/2, k \in \{0, 1, 2, 3\}\}$ , we can factor out the corresponding rotation, which is (up to a phase) a Clifford gate. This way we can fall back to the case of an unbiased even distribution, i.e., symmetric with respect to the zero angle. Note that in the particular case of the uniform distribution over  $[0, 2\pi]$ , we have  $r_1 = 0$ .

## 2. 2-fold channel

In this section we make use of the Choi representation of quantum channels, which allows us to represent channels acting on two-qubit states by  $16 \times 16$  matrices. For a quantum channel (i.e., a completely positive trace-preserving map)  $\mathcal{E}$ , the Choi operator is defined by

$$\Lambda(\mathcal{E}) = \sum_{i,j,k,l=0}^1 |ij\rangle\langle kl| \otimes \mathcal{E}(|ij\rangle\langle kl|). \quad (\text{A6})$$

Its corresponding matrix entries are

$$\begin{aligned} \Lambda(\mathcal{E})_{(ijkl),(mnpq)} &= \text{Tr}[\Lambda(\mathcal{E})^\dagger(|ij\rangle\langle kl| \otimes |mn\rangle\langle pq|)] \\ &= \text{Tr}[\mathcal{E}(|ij\rangle\langle kl|)^\dagger |mn\rangle\langle pq|]. \end{aligned} \quad (\text{A7})$$

In the following, we write

$$\mathcal{E}[\hat{U}](\hat{\rho}) := \hat{U}\hat{\rho}\hat{U}^\dagger \quad (\text{A8})$$

for the quantum channel associated with a unitary transformation  $\hat{U}$ . We assume that  $\hat{U}$  is diagonal in the computational basis, so that we can write

$$\hat{U} = \sum_{i,j=0}^1 \lambda_{ij} \hat{\Pi}_{ij}, \quad (\text{A9})$$

where we define the projectors  $\hat{\Pi}_{ij} := \hat{\Pi}_i \otimes \hat{\Pi}_j$ . For  $\hat{U}$  unitary, we have  $\hat{U}\hat{U}^\dagger = \mathbb{1} = \sum_{i,j} \lambda_{ij} \lambda_{ij}^* \hat{\Pi}_{ij}$ , and hence  $\lambda_{ij} = e^{i\theta_{ij}}$  for all  $i, j$ . Therefore we have

$$\begin{aligned} \mathcal{E}[\hat{U}](|ij\rangle\langle kl|) &= \sum_{m,n,p,q} \lambda_{mn} \lambda_{pq}^* \hat{\Pi}_{mn} |ij\rangle\langle kl| \hat{\Pi}_{pq} \\ &= \lambda_{ij} \lambda_{kl}^* |ij\rangle\langle kl| \\ &= e^{i(\theta_{ij} - \theta_{kl})} |ij\rangle\langle kl|. \end{aligned} \quad (\text{A10})$$

Thus, the Choi matrix of  $\mathcal{E}[\hat{U}]$  is diagonal whenever  $\hat{U}$  is of the form given in Eq. (A9). We can represent it by a  $4 \times 4$  matrix  $M$ , whose entries are defined by

$$M_{(ij),(kl)} := \Lambda_{(ijkl),(ijkl)}. \quad (\text{A11})$$

Note that matrix  $M$  is Hermitian and that its diagonal entries are always equal to one, due to Eq. (A10). In the following we represent each channel by its associated matrix  $M$  in the basis  $(00), (01), (10), (11)$ .

As earlier, we focus on rotations around the  $Z$  axis. We have

$$\Phi_Z^{(2)}(\hat{\rho}) := \mathbb{E}_\theta[(\hat{R}_Z(\theta) \otimes \hat{R}_Z(\theta))\hat{\rho}(\hat{R}_Z^\dagger(\theta) \otimes \hat{R}_Z^\dagger(\theta))] \quad (\text{A12})$$

and

$$\begin{aligned} \hat{R}_Z(\theta) \otimes \hat{R}_Z(\theta) &= (e^{-i\theta} \hat{\Pi}_0 \otimes \hat{\Pi}_0 + e^{i\theta} \hat{\Pi}_1 \otimes \hat{\Pi}_1) \\ &\quad + (\hat{\Pi}_0 \otimes \hat{\Pi}_1 + \hat{\Pi}_1 \otimes \hat{\Pi}_0). \end{aligned} \quad (\text{A13})$$

Defining

$$\Gamma_\theta = (e^{-i\theta} \hat{\Pi}_{00} + e^{i\theta} \hat{\Pi}_{11}), \quad (\text{A14})$$

$$\Xi = \hat{\Pi}_{01} + \hat{\Pi}_{10},$$

we can write

$$\begin{aligned} \Phi_Z^{(2)}(\hat{\rho}) &= \mathbb{E}_\theta[\Xi\hat{\rho}\Xi^\dagger] + \mathbb{E}_\theta[\Gamma_\theta\hat{\rho}\Gamma_\theta^\dagger] \\ &\quad + \mathbb{E}_\theta[\Gamma_\theta\hat{\rho}\Xi^\dagger] + \mathbb{E}_\theta[\Xi\hat{\rho}\Gamma_\theta^\dagger]. \end{aligned} \quad (\text{A15})$$

### a. Uniform distribution

For the uniform distribution of  $\theta$  in  $[0, 2\pi]$ , we have  $\mathbb{E}_\theta[e^{\pm i\theta}] = \mathbb{E}_\theta[e^{\pm 2i\theta}] = 0$ , and, thus,

$$\begin{aligned}\mathbb{E}_\theta[\Gamma_\theta \hat{\rho} \Xi^\dagger] &= 0, \\ \mathbb{E}_\theta[\Gamma_\theta \hat{\rho} \Gamma_\theta^\dagger] &= \hat{\Pi}_{00} \hat{\rho} \hat{\Pi}_{00} + \hat{\Pi}_{11} \hat{\rho} \hat{\Pi}_{11}, \\ \mathbb{E}_\theta[\Xi \hat{\rho} \Xi^\dagger] &= \hat{\Pi}_{01} \hat{\rho} \hat{\Pi}_{01} + \hat{\Pi}_{10} \hat{\rho} \hat{\Pi}_{10}, \\ &\quad + \hat{\Pi}_{01} \hat{\rho} \hat{\Pi}_{10} + \hat{\Pi}_{10} \hat{\rho} \hat{\Pi}_{01}.\end{aligned}$$

Finally, we get

$$\begin{aligned}\Phi_Z^{(2)}(\hat{\rho}) &= \hat{\Pi}_{00} \hat{\rho} \hat{\Pi}_{00} + \hat{\Pi}_{11} \hat{\rho} \hat{\Pi}_{11} + \hat{\Pi}_{01} \hat{\rho} \hat{\Pi}_{01} \\ &\quad + \hat{\Pi}_{10} \hat{\rho} \hat{\Pi}_{10} + \hat{\Pi}_{01} \hat{\rho} \hat{\Pi}_{10} + \hat{\Pi}_{10} \hat{\rho} \hat{\Pi}_{01}.\end{aligned}\quad (\text{A16})$$

We can represent  $\Phi_Z^{(2)}$  by its associated matrix

$$M(\Phi_Z^{(2)}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.\quad (\text{A17})$$

One can verify that the following channels also have a diagonal Choi matrix, and we can use the same representation of their diagonals, giving

$$M(\mathcal{E}[\mathbb{1}]) = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix},\quad (\text{A18a})$$

$$M(\mathcal{E}[\hat{Z} \otimes \hat{Z}]) = \begin{pmatrix} 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix},\quad (\text{A18b})$$

$$M(\mathcal{E}[\hat{S} \otimes \hat{S}]) = \begin{pmatrix} 1 & i & i & -1 \\ -i & 1 & 1 & i \\ -i & 1 & 1 & i \\ -1 & -i & -i & 1 \end{pmatrix},\quad (\text{A18c})$$

$$M(\mathcal{E}[\hat{S}^\dagger \otimes \hat{S}^\dagger]) = \begin{pmatrix} 1 & -i & -i & -1 \\ i & 1 & 1 & -i \\ i & 1 & 1 & -i \\ -1 & i & i & 1 \end{pmatrix},\quad (\text{A18d})$$

with  $\hat{S} = \Pi_0 + i\Pi_1$  the phase gate. Gathering all together, we have

$$\begin{aligned}4M(\Phi_Z^{(2)}) &= M(\mathcal{E}[\mathbb{1}]) + M(\mathcal{E}[\hat{Z} \otimes \hat{Z}]) \\ &\quad + M(\mathcal{E}[\hat{S} \otimes \hat{S}]) + M(\mathcal{E}[\hat{S}^\dagger \otimes \hat{S}^\dagger]).\end{aligned}\quad (\text{A19})$$

The final result in the main text then follows by linearity and uniqueness of the Choi matrix.

### b. Even distribution

Let us consider an even probability distribution of  $\theta$  (i.e., a distribution for which  $\theta$  has the same law as  $-\theta$ ). For such distributions, we again have  $\phi_\theta(t) = \phi_\theta(-t) \in [-1, 1] \subset \mathbb{R}$  for all  $t \in \mathbb{R}$  and, thus,

$$\phi_\theta(t) = \frac{1}{2}(\phi_\theta(t) + \phi_\theta(-t)) = \mathbb{E}_\theta[\cos(t\theta)].$$

Defining  $r_1 = \phi_\theta(1)$  and  $r_2 = \phi_\theta(2)$ , we can write

$$\begin{aligned}\mathbb{E}_\theta[\Gamma_\theta \hat{\rho} \Xi^\dagger] &= r_1(\hat{\Pi}_{00} + \hat{\Pi}_{11})\hat{\rho}(\hat{\Pi}_{01} + \hat{\Pi}_{10}), \\ \mathbb{E}_\theta[\Gamma_\theta \hat{\rho} \Gamma_\theta^\dagger] &= \hat{\Pi}_{00} \hat{\rho} \hat{\Pi}_{00} + \hat{\Pi}_{11} \hat{\rho} \hat{\Pi}_{11} \\ &\quad + r_2(\hat{\Pi}_{00} \hat{\rho} \hat{\Pi}_{11} + \hat{\Pi}_{11} \hat{\rho} \hat{\Pi}_{00}), \\ \mathbb{E}_\theta[\Xi \hat{\rho} \Xi^\dagger] &= \hat{\Pi}_{01} \hat{\rho} \hat{\Pi}_{01} + \hat{\Pi}_{10} \hat{\rho} \hat{\Pi}_{10} \\ &\quad + \hat{\Pi}_{01} \hat{\rho} \hat{\Pi}_{10} + \hat{\Pi}_{10} \hat{\rho} \hat{\Pi}_{01}.\end{aligned}$$

Hence we obtain

$$M(\Phi_Z^{(2)}) = \begin{pmatrix} 1 & r_1 & r_1 & r_2 \\ r_1 & 1 & 1 & r_1 \\ r_1 & 1 & 1 & r_1 \\ r_2 & r_1 & r_1 & 1 \end{pmatrix}.\quad (\text{A20})$$

We can express  $M(\Phi_Z^{(2)})$  as a linear combination of matrices (A18), giving

$$\begin{aligned}M(\Phi_Z^{(2)}) &= aM(\mathcal{E}[\mathbb{1}]) + bM(\mathcal{E}[\hat{Z} \otimes \hat{Z}]) \\ &\quad + \frac{c}{2}(M(\mathcal{E}[\hat{S} \otimes \hat{S}]) + M(\mathcal{E}[\hat{S}^\dagger \otimes \hat{S}^\dagger])).\end{aligned}\quad (\text{A21})$$

The coefficients  $a, b, c$  can be found by solving the linear system

$$\begin{aligned}a + b + c &= 1, \\ a - b &= r_1, \\ a + b - c &= r_2,\end{aligned}$$

and one finds that

$$\begin{aligned}M(\Phi_Z^{(2)}) &= \frac{1}{4}(1 + r_2 + 2r_1)M(\mathcal{E}[\mathbb{1}]) \\ &\quad + \frac{1}{4}(1 + r_2 - 2r_1)M(\mathcal{E}[\hat{Z} \otimes \hat{Z}]) \\ &\quad + \frac{1}{4}(1 - r_2)M(\mathcal{E}[\hat{S} \otimes \hat{S}]) \\ &\quad + \frac{1}{4}(1 - r_2)M(\mathcal{E}[\hat{S}^\dagger \otimes \hat{S}^\dagger]).\end{aligned}\quad (\text{A22})$$

Therefore, the associated channel is

$$\begin{aligned}\Phi_Z^{(2)}(\hat{\rho}) &= \frac{1}{4}(1 + r_2 + 2r_1)\hat{\rho} \\ &\quad + \frac{1}{4}(1 + r_2 - 2r_1)(\hat{Z} \otimes \hat{Z})\hat{\rho}(\hat{Z} \otimes \hat{Z}) \\ &\quad + \frac{1}{4}(1 - r_2)(\hat{S} \otimes \hat{S})\hat{\rho}(\hat{S} \otimes \hat{S}) \\ &\quad + \frac{1}{4}(1 - r_2)(\hat{S}^\dagger \otimes \hat{S}^\dagger)\hat{\rho}(\hat{S} \otimes \hat{S}).\end{aligned}\quad (\text{A23})$$

*Remark.*—Defining  $CZ = \hat{\Pi}_0 \otimes 1 + \hat{\Pi}_1 \otimes \hat{Z}$ , the control-Z gate, and  $CZ_X = (\hat{X} \otimes \hat{X})CZ(\hat{X} \otimes \hat{X})$ , we have

$$M(\mathcal{E}[CZ]) = \begin{pmatrix} 1 & 1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ -1 & -1 & -1 & 1 \end{pmatrix}, \quad (\text{A24})$$

$$M(\mathcal{E}[CZ_X]) = \begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix},$$

and, thus,

$$\mathcal{E}[\hat{S} \otimes \hat{S}] + \mathcal{E}[\hat{S}^\dagger \otimes \hat{S}^\dagger] = \mathcal{E}[CZ] + \mathcal{E}[CZ_X]. \quad (\text{A25})$$

Therefore, the decomposition of  $\Phi_Z^{(2)}$  into a convex sum of Clifford channels in Eq. (A23) is not unique.

The decomposition obtained in Eq. (A23) is a convex sum if one assumes that  $(1 + r_2 - 2r_1) \geq 0$  and  $(1 + r_2 + 2r_1) \geq 0$ . This condition holds if and only if

$$\mathbb{E}_\theta \left[ \frac{1}{2}(1 + \cos 2\theta) \pm \cos \theta \right] \geq 0,$$

namely, if and only if

$$\mathbb{E}_\theta[\cos^2 \theta] \geq |\mathbb{E}_\theta[\cos \theta]|. \quad (\text{A26})$$

This condition can be reformulated as a positivity condition on the expectations of two functions of the random angle, as represented in Fig. 9. It is fulfilled for the distributions that are  $\pi$  periodic, as in that case we have  $\mathbb{E}_\theta[\cos \theta] = 0$ . Another example of a distribution that satisfies this constraint is a Gaussian distribution with a large enough variance. In fact, for a centered Gaussian distribution of variance  $\sigma^2$ , we have  $r_1 = e^{-\sigma^2/2}$  and  $r_2 = e^{-2\sigma^2}$ ,

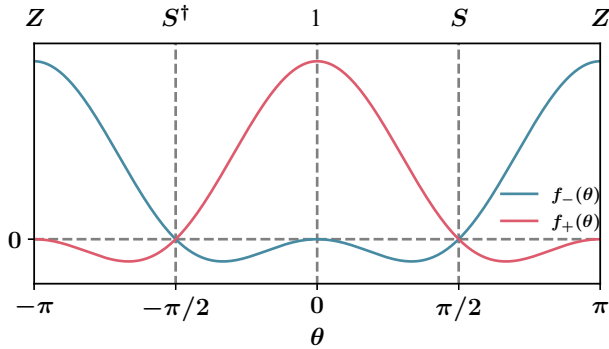


FIG. 9. Plot of  $f_{\pm}(\theta) = \cos \theta(\cos \theta \pm 1)$  versus  $\theta$ . The condition in Eq. (A26) is fulfilled if and only if  $\mathbb{E}_\theta[f_+(\theta)] \geq 0$  and  $\mathbb{E}_\theta[f_-(\theta)] \geq 0$ .  $\hat{S}$  is the phase gate.

so the condition becomes

$$1 + e^{-2\sigma^2} - 2e^{-\sigma^2/2} \geq 0. \quad (\text{A27})$$

One can show that this condition is equivalent to  $\sigma^2 \geq \sigma_0^2$  for some specific  $\sigma_0 \in \mathbb{R}$ , yielding a requirement on the width of the Gaussian.

## APPENDIX B: FIRST- AND SECOND-ORDER QUANTITIES

In this appendix, we define the notion of first- and second-order quantities as quantities that can be obtained from knowledge of the 1-fold and 2-fold channels, respectively, for each of the random rotations appearing in a given ansatz. We also show that the average cost function and the average gradient are first-order quantities, while the average of the squared cost function and of the squared gradient are second-order quantities.

### 1. First-order quantities

Let us consider the ansatz defined by

$$\hat{U}(\boldsymbol{\theta}) = \prod_{i=1}^M \hat{U}_i(\theta_i) \hat{W}_i, \quad (\text{B1})$$

and denote by

$$\begin{aligned} \mathcal{U}_i(\theta_i)(\hat{\rho}) &= \hat{U}_i(\theta_i) \hat{\rho} \hat{U}_i^\dagger(\theta_i), \\ \mathcal{W}_i(\hat{\rho}) &= \hat{W}_i \hat{\rho} \hat{W}_i^\dagger \end{aligned} \quad (\text{B2})$$

the unitary channels associated with different layers of the circuit. The whole circuit unitary transformation then reads

$$\begin{aligned} \mathcal{U}(\boldsymbol{\theta})(\hat{\rho}) &= \mathcal{U}_M(\theta_M) \circ \cdots \circ \mathcal{W}_1(\hat{\rho}) \\ &= \bigcirc_{i=1}^M (\mathcal{U}_i(\theta_i) \circ \mathcal{W}_i)(\hat{\rho}). \end{aligned} \quad (\text{B3})$$

The cost function is then given by

$$C(\boldsymbol{\theta}) = \text{Tr}[\mathcal{U}(\boldsymbol{\theta})(\hat{\rho}) \hat{O}] \quad (\text{B4})$$

and its expectation with respect to  $\boldsymbol{\theta}$  is

$$\begin{aligned} \mathbb{E}_\theta[C(\boldsymbol{\theta})] &= \mathbb{E}_\theta[\text{Tr}[\mathcal{U}(\boldsymbol{\theta})(\hat{\rho}) \hat{O}]] \\ &= \text{Tr}[\mathbb{E}_\theta[\mathcal{U}(\boldsymbol{\theta})(\hat{\rho})] \hat{O}] \\ &= \text{Tr}[\mathbb{E}_\theta[\hat{U}(\boldsymbol{\theta}) \hat{\rho} \hat{U}^\dagger(\boldsymbol{\theta})] \hat{O}] \\ &= \text{Tr} \left[ \int_{\mathbb{R}^M} \hat{U}(\boldsymbol{\theta}) \hat{\rho} \hat{U}^\dagger(\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \hat{O} \right] \\ &= \text{Tr}[\Phi_\theta^{(1)}(\hat{\rho}) \hat{O}]. \end{aligned} \quad (\text{B5})$$

Here, we used both the linearity of the expectation and the definition of the 1-fold channel from Eq. (9). The cost-function expectation can thus be obtained from knowledge

of the complete 1-fold channel  $\Phi_\theta^{(1)}$ . Assuming that the angles  $\{\theta_i\}$  are independent from each other, the expectation against  $\theta$  can be factored in expectations against the  $\theta_i$ , which allows us to write

$$\begin{aligned}\Phi_\theta^{(1)}(\hat{\rho}) &= \mathbb{E}_\theta[\mathcal{U}(\theta)(\hat{\rho})] \\ &= \bigcirc_{i=1}^M (\mathbb{E}_{\theta_i}[\mathcal{U}_i(\theta_i)] \circ \mathcal{W}_i)(\hat{\rho}).\end{aligned}\quad (\text{B6})$$

As explained in the main text, we can consider, without loss of generality, all the rotations to be  $Z$  rotations. Then the channels  $\mathbb{E}_{\theta_i}[\mathcal{U}_i(\theta_i)]$  are exactly 1-fold channels associated with a  $Z$  rotation acting on a single qubit, and hence can be computed from the results of Appendix A. As stated earlier, we refer to quantities that can be obtained from knowledge of the 1-fold channels associated with each rotation of the ansatz as *first-order quantities*. Hence, the average cost function  $\mathbb{E}_\theta[C(\theta)]$  is a first-order quantity.

Another example of an interesting first-order quantity is the average of the gradient. From the parameter-shift rule and using the linearity of the expectation, we have

$$\begin{aligned}\mathbb{E}_\theta[\partial_k C(\theta)] &= \frac{1}{2} \mathbb{E}_\theta \left[ C\left(\theta + \frac{\pi}{2} \mathbf{e}_k\right) \right] \\ &\quad - \frac{1}{2} \mathbb{E}_\theta \left[ C\left(\theta - \frac{\pi}{2} \mathbf{e}_k\right) \right].\end{aligned}\quad (\text{B7})$$

Here  $\mathbf{e}_k$  is the unit vector along the  $k$ th component. The  $\pm\pi/2$  shifts in parameter  $\theta_k$  can be factored out and seen as an extra Clifford gate added to the fixed layer  $\hat{W}_k$ . In fact, assuming that  $\hat{P}_k = \hat{Z}$  and denoting by  $\hat{S}$  the phase gate, we have  $\hat{U}_k(\theta_k + \pi/2) \hat{W}_k = e^{-i\theta_k \hat{P}_k/2} e^{-i\pi \hat{Z}/2} \hat{W}_k = e^{-i\pi/4} \hat{U}_k(\theta_k) \hat{S} \hat{W}_k$ . Defining  $\hat{W}_{k,\pm} = e^{\mp i\pi/4} \hat{S} \hat{W}_k$ , we get  $\hat{U}_k(\theta_k + \pi/2) \hat{W}_k = \hat{U}_k(\theta_k) \hat{W}_{k,+}$ . We can proceed likewise to define  $\hat{W}_{k,-}$ . In the following, we can write, for all  $i \neq k$ ,  $\hat{V}_{i,\pm} = \hat{W}_i$  and  $\hat{V}_{k,\pm} = \hat{W}_{k,\pm}$  for the modified fixed layers that include the considered shift. We have

$$\begin{aligned}\mathbb{E}_\theta[\hat{U}_\pm(\theta) \hat{\rho} \hat{U}_\pm^\dagger(\theta)] &= \mathbb{E}_\theta \left[ \bigcirc_{i=1}^M (\mathcal{U}_i(\theta_i) \circ \mathcal{V}_{i,\pm})(\hat{\rho}) \right] \\ &= \bigcirc_{i=1}^M (\mathbb{E}_{\theta_i}[\mathcal{U}_i(\theta_i)] \circ \mathcal{V}_{i,\pm})(\hat{\rho}),\end{aligned}\quad (\text{B8})$$

where  $\mathcal{V}_{i,\pm}(\hat{\rho}) = \hat{V}_{i,\pm} \hat{\rho} \hat{V}_{i,\pm}^\dagger$ . The average gradient is therefore a first-order quantity, namely, depending on 1-fold channels only.

## 2. Second-order quantities

We now turn our attention to the mean value of the squared cost function. This is given by

$$\begin{aligned}\mathbb{E}_\theta[C(\theta)^2] &= \mathbb{E}_\theta[\text{Tr}[\mathcal{U}(\theta)(\hat{\rho}) \hat{O}]^2] \\ &= \mathbb{E}_\theta[\text{Tr}[(\mathcal{U}(\theta)(\hat{\rho}) \hat{O})^{\otimes 2}]] \\ &= \mathbb{E}_\theta[\text{Tr}[\mathcal{U}^{(2)}(\theta)(\hat{\rho}^{\otimes 2}) \hat{O}^{\otimes 2}]].\end{aligned}\quad (\text{B9})$$

For every state  $\hat{\rho}$  of a system of  $2n$  qubits (i.e., a doubled version of the original system where the copy is not connected by gates to the original circuit), we define

$$\mathcal{U}^{(2)}(\theta)(\hat{\rho}) = \hat{U}^{\otimes 2}(\theta) \hat{\rho} \hat{U}^{\dagger \otimes 2}(\theta).\quad (\text{B10})$$

Likewise, we can define the doubled version of the circuit layers as

$$\begin{aligned}\mathcal{U}_i^{(2)}(\theta_i)(\hat{\rho}) &= \hat{U}_i^{\otimes 2}(\theta_i) \hat{\rho} \hat{U}_i^{\dagger \otimes 2}(\theta_i), \\ \mathcal{W}_i^{(2)}(\hat{\rho}) &= \hat{W}_i^{\otimes 2} \hat{\rho} \hat{W}_i^{\otimes 2},\end{aligned}\quad (\text{B11})$$

giving

$$\mathcal{U}^{(2)}(\theta)(\hat{\rho}) = \bigcirc_{i=1}^M (\mathcal{U}_i^{(2)}(\theta_i) \circ \mathcal{W}_i^{(2)})(\hat{\rho}).\quad (\text{B12})$$

Thus, for independent rotations, we have

$$\begin{aligned}\Phi_\theta^{(2)}(\hat{\rho}) &= \mathbb{E}_\theta[\mathcal{U}^{(2)}(\theta)(\hat{\rho})] \\ &= \bigcirc_{i=1}^M (\mathbb{E}_{\theta_i}[\mathcal{U}_i^{(2)}(\theta_i)] \circ \mathcal{W}_i^{(2)})(\hat{\rho}).\end{aligned}\quad (\text{B13})$$

As for first-order quantities, we refer to quantities that can be obtained from knowledge of the average 2-fold channels of the rotations layers  $\mathbb{E}_{\theta_i}[\mathcal{U}_i^{(2)}(\theta_i)]$  as second-order quantities.

The average of the squared cost function is thus a second-order quantity, and as for the first-order case, we can show that the squared gradient is also a second-order quantity. In fact, by making use of the parameter-shift rule, we see that, to obtain the average of the squared gradient, we have to compute the following four terms:

$$\mathbb{E}_\theta[C(\theta + a_1 \mathbf{e}_k) C(\theta + a_2 \mathbf{e}_k)]\quad (\text{B14})$$

with  $a_1, a_2 \in \{\pi/2, -\pi/2\}$ . As before, it suffices to replace the  $\mathcal{W}_i^{(2)}$  in Eq. (B13) with

$$\mathcal{V}_{i,a_1,a_2}^{(2)}(\hat{\rho}) = (\hat{V}_{i,a_1} \otimes \hat{V}_{i,a_2}) \hat{\rho} (\hat{V}_{i,a_1}^\dagger \otimes \hat{V}_{i,a_2}^\dagger).\quad (\text{B15})$$

Finally, the gradient variance can be computed as

$$\text{Var}_\theta[\partial_k C(\theta)] = \mathbb{E}_\theta[\partial_k C(\theta)^2] - \mathbb{E}_\theta[\partial_k C(\theta)]^2,\quad (\text{B16})$$

which is the sum of a first- and a second-order quantity.

## APPENDIX C: PROOF OF THE SAMPLING EFFICIENCY

In this appendix we prove that, to obtain an estimation of any first- or second-order quantity for a given ansatz up to a precision  $\epsilon$  and probability  $\delta \in [0, 1]$  to meet this precision, it suffices to sample a number of Clifford approximant circuits  $K \sim \log(2\delta)M/\epsilon^2$ . By invoking the Gottesman-Knill theorem, we obtain an estimation of any of the previous quantities with a complexity polynomial in both the size of the system and the number of variational parameters of the considered ansatz.

### 1. Details on the mapping

Here we give details on the mapping of the randomly initialized parameterized circuit to Clifford approximants.

*Remark.*—We use the notation adapted to first-order quantities. The generalization to the second order and the shifted versions is straightforward as it suffices to replace each channel by its doubled and/or shifted version, as done in Appendix B.

Assuming that the  $\theta_i$  are independent from each other, averaging  $\mathcal{U}(\boldsymbol{\theta})$  over  $\boldsymbol{\theta}$  amounts to replacing each rotation channel  $\mathcal{U}_i(\theta_i)$  by a convex sum of  $m$  Clifford unitary channels  $\mathcal{U}_{ij}$  with associated weights  $p_{ij}$ . Thus,  $\mathbb{E}_{\boldsymbol{\theta}}[\mathcal{U}(\boldsymbol{\theta})(\hat{\rho})]$  is replaced by a discrete average over  $m^M$  Clifford unitary channels (with  $m = 2$  for the 1-fold channel and  $m = 4$  for the 2-fold channel):

$$\mathbb{E}_{\boldsymbol{\theta}}[\mathcal{U}(\boldsymbol{\theta})(\hat{\rho})] = \bigcirc_{i=1}^M \left( \sum_{j=1}^m p_{ij} \mathcal{U}_{ij} \circ \mathcal{W}_i \right) (\hat{\rho}). \quad (\text{C1})$$

As we want to sample from that sum, we can define, for each  $i$ , a discrete random variable  $X_i$  taking values in  $\{1, \dots, m\}$  such that  $\mathbb{P}(X_i = j) = p_{ij}$ . This represents a choice of a given unitary in the previous convex sum. Gathering these for all  $k$  we get a random vector  $\mathbf{X} = (X_1, \dots, X_M) \in \{1, \dots, m\}^M$  that completely defines a unique unitary  $\mathcal{U}(\mathbf{X})$  through

$$\mathcal{U}(j_1, \dots, j_M) = \bigcirc_{i=1}^M \mathcal{U}_{ij_i} \circ \mathcal{W}_i. \quad (\text{C2})$$

Thus we have

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}[\mathcal{U}(\boldsymbol{\theta})(\hat{\rho})] &= \mathbb{E}_{\mathbf{X}}[\mathcal{U}(\mathbf{X})(\hat{\rho})] \\ &= \bigcirc_{i=1}^M \left( \sum_{j=1}^m p_{ij} \mathcal{U}_{ij} \circ \mathcal{W}_i \right) (\hat{\rho}). \end{aligned} \quad (\text{C3})$$

The main idea is now to approximate the  $k$ -fold channels by an empirical average over  $K$  samples of the previous

Clifford circuits, namely,

$$\hat{\Phi}(\hat{\rho}) := \frac{1}{K} \sum_{i=1}^K \mathcal{U}(\mathbf{X}_i)(\hat{\rho}). \quad (\text{C4})$$

### 2. Sampling efficiency

Our result relies on classical arguments for the sampling of bounded functions depending on a set of random variables using McDiarmid's concentration inequality [73,74], which we state below.

*Definition 1 (Bounded difference property).*—A function  $f : \mathcal{X}^M \rightarrow \mathbb{R}$  satisfies the bounded difference property if and only if there exist bounds  $\{c_1, \dots, c_M\}$  such that, for all  $i \in \{1, \dots, M\}$  and all  $(x_1, \dots, x_M)$ ,

$$\sup_{x'_i \in \mathcal{X}} |f(x_1, \dots, x_i, \dots, x_M) - f(x_1, \dots, x'_i, \dots, x_M)| < c_i.$$

*Theorem 2 (McDiarmid's inequality).*—Let  $f : \mathcal{X}^M \rightarrow \mathbb{R}$  satisfy the bounded difference property with bounds  $\{c_1, \dots, c_M\}$  and a random vector  $\mathbf{X} = (X_1, \dots, X_M)$  taking values in  $\mathcal{X}^M$ . Then, for all  $\epsilon > 0$ ,

$$\mathbb{P}(|f(\mathbf{X}) - \mathbb{E}_{\mathbf{X}}[f(\mathbf{X})]| \geq \epsilon) \leq 2 \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^M c_i^2}\right).$$

We show that the quantities we want to estimate satisfy the bounded difference property and apply McDiarmid's inequality to prove that our previous sampling is efficient. In the following we define

$$f(\mathbf{x}) = \text{Tr}[\mathcal{U}(\mathbf{x})(\hat{\rho})\hat{O}], \quad (\text{C5})$$

where  $\hat{O}$  is the cost-function observable defined in the main text and, as in the previous section,  $\mathcal{U}(\mathbf{x})$  the unitary channel associated with a given Clifford approximant circuit that is completely specified by a discrete vector  $\mathbf{x} = (x_1, \dots, x_i, \dots, x_M) \in \{1, \dots, m\}^M$ . By the Hölder inequality [49,75],  $f$  is upper bounded:

$$|f(\mathbf{x})| \leq \|\hat{\rho}\|_1 \|\hat{O}\|_{\infty} \quad (\text{C6})$$

with  $\|A\|_1, \|A\|_{\infty}$  the Schatten-1 norm and spectral norm, respectively [49]. We note that  $\|\hat{\rho}\|_1 = 1$  for  $\hat{\rho}$  a density operator. Defining a second vector for which only the  $i$ th component is changed,  $\mathbf{x}' = (x_1, \dots, x'_i, \dots, x_M)$ , we get, using the triangle inequality,

$$\begin{aligned} |f(\mathbf{x}) - f(\mathbf{x}')| &\leq |f(\mathbf{x})| + |f(\mathbf{x}')| \\ &\leq 2\|\hat{O}\|_{\infty}. \end{aligned} \quad (\text{C7})$$

Hence,  $f$  satisfies the bounded difference property with  $c_i = c = 2\|\hat{O}\|_{\infty}$ , and we can apply McDiarmid's inequality, which gives almost the desired result. To go further, we

define

$$\begin{aligned} f_K(\mathbf{x}_1, \dots, \mathbf{x}_K) &= \sum_{j=1}^K f(x_{j1}, \dots, x_{jM}) \\ &= \sum_{j=1}^K \text{Tr}[\mathcal{U}(\mathbf{x}_j)(\hat{\rho})\hat{O}] \\ &= K\text{Tr}[\hat{\Phi}(\hat{\rho})\hat{O}]. \end{aligned} \quad (\text{C8})$$

Clearly,  $f_K$  satisfies the bounded difference property with the same bound  $c$  [to see this, we take all  $x_{ij}$  equal except for  $x_{kl}$ , and it follows that the difference  $f_K(\mathbf{x}_1, \dots, \mathbf{x}_K) - f_K(\mathbf{x}'_1, \dots, \mathbf{x}'_K)$  is simply  $f(\mathbf{x}_k) - f(\mathbf{x}'_k)$ ]. Thus, McDiarmid's inequality applies to  $f_K$ , which is a function of  $KM$  parameters:

$$\begin{aligned} \mathbb{P}(|f_K(\mathbf{X}) - \mathbb{E}_{\mathbf{X}}[f_K(\mathbf{X})]| \geq K\epsilon) &= \mathbb{P}\left(\left|\frac{1}{K}f_K(\mathbf{X}) - \mathbb{E}_{\mathbf{X}}\left[\frac{1}{K}f_K(\mathbf{X})\right]\right| \geq \epsilon\right) \\ &= \mathbb{P}(|\text{Tr}[\hat{\Phi}(\hat{\rho})\hat{O}] - \mathbb{E}_{\theta}[\text{Tr}[\mathcal{U}(\theta)(\hat{\rho})\hat{O}]]| \geq \epsilon) \\ &\leq 2\exp\left(-\frac{2K^2\epsilon^2}{KM c^2}\right) \\ &= 2\exp\left(-\frac{K\epsilon^2}{2M\|\hat{O}\|_{\infty}^2}\right). \end{aligned} \quad (\text{C9})$$

Therefore, choosing a precision  $\epsilon > 0$  and a probability  $1 - \delta \in [0, 1]$  to meet this precision, we get

$$\begin{aligned} \mathbb{P}(|\text{Tr}[\hat{\Phi}(\hat{\rho})\hat{O}] - \mathbb{E}_{\theta}[\text{Tr}[\mathcal{U}(\theta)(\hat{\rho})\hat{O}]]| \leq \epsilon) \\ \geq 1 - \delta \end{aligned} \quad (\text{C10})$$

whenever the number of sampled Clifford circuits  $K$  is

$$K \geq \frac{2}{\epsilon^2} \log\left(\frac{2}{\delta}\right) M \|\hat{O}\|_{\infty}^2 = O(M). \quad (\text{C11})$$

Note that in Eq. (C9), replacing the observable  $\hat{O}$  by its normalized counterpart  $\hat{O}/\|\hat{O}\|$  with an associated precision  $\tilde{\epsilon}$  gives the same scaling for  $K$ , as in the case  $\tilde{\epsilon} = \epsilon/\|\hat{O}\|$ . Hence we can always work with a normalized observable. However, if one is interested in the scaling with the system size  $n$ , we have to consider a sequence of observables  $\hat{O}_n$ , whose norms can present a particular scaling in  $n$ , so the presence of the norm of  $\hat{O}$  in Eq. (C11) allows us to keep track of this effect. In many situations of interest, the observables considered scale polynomially in the system size, and so does  $K$ . Finally, one can use the Gottesman-Knill theorem, which states that, for a Clifford unitary  $\hat{U}$  and an observable  $\hat{O}$  acting nontrivially on  $N_O$

qubits, the expectation value  $\text{Tr}[|0\rangle\langle 0|^{\otimes n} \hat{U}^{\dagger} \hat{O} \hat{U}]$  can be classically computed with a complexity polynomial in both  $N_O$  and the number of qubits  $n$  [24]. Our scheme inherits this scaling and we can estimate the gradient variance  $\text{Var}_{\theta}[\partial_k C(\theta)]$  for each  $k$  on a classical computer with complexity  $O(n^p N_O^q M)$ , where  $M$  is the number of parameters in the variational quantum circuit.

## APPENDIX D: SAMPLING EFFICIENCY IN THE GENERAL CASE

In this section we extend the previous scheme to more general distributions. We first discuss in Appendix D 1 the scaling of the sampling complexity with the convexity condition relaxed, i.e., where we no longer require the decomposition of the 2-fold channel [Eq. (A23)] to be a convex sum and only assume that the distribution of  $\theta$  is even. Then, in Appendix D 2 we study the case of an arbitrary distribution of the rotation angles, which is not necessarily symmetrically distributed. Finally, we show that our previous scheme still applies in this general case, but that it requires to sample a number of Clifford approximant circuits scaling exponentially in the number of variational parameters  $M$ . Compared to a brute-force simulation, this method can be used to trade an exponential complexity in the system size for an exponential complexity in the number of variational parameters.

### 1. Sampling efficiency in the nonconvex case

Here we consider distributions of rotation angle  $\theta$  that are even, but do not satisfy the convexity condition of Eq. (A26). In this case, our decomposition of the 1-fold channel remains convex, while the 2-fold channel becomes a nonconvex sum; hence, the coefficients for the Clifford channels can no longer be interpreted as probabilities. We first show how one can still estimate such nonconvex sums via probabilistic sampling [72]. Letting

$$\mathbb{E}_{\theta}[\mathcal{U}(\theta)(\hat{\rho})] = \bigcirc_{k=1}^M \left( \sum_{j=1}^m q_{kj} \mathcal{U}_{kj} \circ \mathcal{W}_k \right) (\hat{\rho}), \quad (\text{D1})$$

we hereby assume that

$$q_{kj} \in \mathbb{R}, \quad \sum_{j=1}^m q_{kj} = 1, \quad \text{for all } k. \quad (\text{D2})$$

Defining

$$\gamma_k := \sum_{j=1}^m |q_{kj}|, \quad \tilde{p}_{kj} := |q_{kj}|/\gamma_k, \quad (\text{D3})$$

Eq. (D1) can be rewritten in terms of convex sums:

$$\mathbb{E}_\theta[\mathcal{U}(\boldsymbol{\theta})] = \bigcirc_{k=1}^M \sum_{j=1}^m \tilde{p}_{kj} [\gamma_k \operatorname{sgn}(q_{kj})] \mathcal{U}_{kj} \circ \mathcal{W}_k. \quad (\text{D4})$$

Similar to Appendix C1, we now define the random vector  $\tilde{\mathbf{X}} = (\tilde{X}_1, \dots, \tilde{X}_M) \in \{1, \dots, m\}^M$ , with probabilities  $\mathbb{P}(\tilde{X}_k = j) = \tilde{p}_{kj}$ , and the rescaled random unitary channel  $\tilde{\mathcal{U}}(\tilde{\mathbf{X}})$  through

$$\tilde{\mathcal{U}}(j_1, \dots, j_M) = \bigcirc_{k=1}^M [\gamma_k \operatorname{sgn}(q_{kj_k})] \mathcal{U}_{kj_k} \circ \mathcal{W}_k. \quad (\text{D5})$$

Therefore, we recover the form of an expectation value similar to Eq. (C3):

$$\mathbb{E}_\theta[\mathcal{U}(\boldsymbol{\theta})(\hat{\rho})] = \mathbb{E}_{\tilde{\mathbf{X}}}[\tilde{\mathcal{U}}(\tilde{\mathbf{X}})(\hat{\rho})]. \quad (\text{D6})$$

This allows us to apply the same arguments as in Appendix C2 by considering the function

$$\tilde{f}(\mathbf{x}) = \operatorname{Tr}[\tilde{\mathcal{U}}(\mathbf{x})(\hat{\rho})\hat{O}] \quad (\text{D7})$$

instead of  $f(\mathbf{x})$  defined in Eq. (C5). The function bound (C6) should be rescaled accordingly:

$$|\tilde{f}(\mathbf{x})| \leq \gamma \|\hat{O}\|_\infty \quad (\text{D8})$$

with the scaling factor defined as

$$\gamma := \prod_{k=1}^M \gamma_k. \quad (\text{D9})$$

The number of sampled Clifford circuits previously derived in Eq. (C11) should therefore be scaled with the same factor:

$$K \geq \frac{2}{\epsilon^2} \log\left(\frac{2}{\delta}\right) \gamma M \|\hat{O}\|_\infty^2. \quad (\text{D10})$$

Note that the factor  $\gamma_k \geq 1$  can be regarded as a measure of “nonconvexity” in the decomposition of the  $k$ th channel. In the case of a convex sum, where  $q_{kj} > 0$  for all  $k, j$ , the scaling factor is simply  $\gamma = 1^M = 1$  and we recover the previous results.

We now show that  $\gamma_k$  is upper bounded. Following our discussion in Appendix A, it suffices to consider the 2-fold channel for a single-qubit  $Z$  rotation, where the decomposition can be possibly nonconvex. Without loss of generality,

let us rewrite Eq. (A23) as

$$\begin{aligned} \Phi_Z^{(2)}(\hat{\rho}) &= q_{k1} \hat{\rho} + q_{k2} (Z \otimes Z) \hat{\rho} (Z \otimes Z) \\ &\quad + q_{k3} CZ \hat{\rho} CZ + q_{k4} CZ_X \hat{\rho} CZ_X \end{aligned} \quad (\text{D11})$$

for some  $k$ , where

$$\begin{aligned} q_{k1} &= \mathbb{E}_\theta \left[ \frac{1}{4} (1 + \cos 2\theta + 2 \cos \theta) \right], \\ q_{k2} &= \mathbb{E}_\theta \left[ \frac{1}{4} (1 + \cos 2\theta - 2 \cos \theta) \right], \\ q_{k3} &= q_{k4} = \mathbb{E}_\theta \left[ \frac{1}{4} (1 - \cos 2\theta) \right]. \end{aligned}$$

Define the non-negative function

$$\begin{aligned} \varphi(\theta) &:= \left| \frac{1}{4} (1 + \cos 2\theta + 2 \cos \theta) \right| \\ &\quad + \left| \frac{1}{4} (1 + \cos 2\theta - 2 \cos \theta) \right| \\ &\quad + 2 \times \left| \frac{1}{4} (1 - \cos 2\theta) \right|. \end{aligned} \quad (\text{D12})$$

We then get

$$\begin{aligned} \gamma_k &= |q_{k1}| + |q_{k2}| + |q_{k3}| + |q_{k4}| \\ &\leq \mathbb{E}_\theta[\varphi(\theta)] \\ &\leq \mathbb{E}_\theta \left[ \sup_{\theta'} \varphi(\theta') \right] \\ &= \sup_{\theta'} \varphi(\theta') \\ &= \frac{5}{4}. \end{aligned} \quad (\text{D13})$$

Here the function  $\varphi(\theta)$  reaches its maximum for  $\theta = \pm\pi/3, \pm 2\pi/3$ . Therefore, the factor  $\gamma_k$  reaches its upper bound  $5/4$  if the distribution of  $\theta$  is a sum of Dirac-delta distributions peaked at  $\theta = \pm\pi/3$  and/or  $\theta = \pm 2\pi/3$ , in which case we obtain the worst-case scaling of the number of sampled Clifford circuits (D10):

$$\begin{aligned} K &\geq \frac{2}{\epsilon^2} \log\left(\frac{2}{\delta}\right) \gamma M \|\hat{O}\|_\infty^2 = O(\gamma M), \\ \gamma &\leq \left(\frac{5}{4}\right)^M. \end{aligned} \quad (\text{D14})$$

Combining the above result with the Gottesman-Knill theorem, for a cost-function observable  $\hat{O}$  acting nontrivially on  $N_O$  qubits, our scheme implies a complexity of at most  $O(n^p N_O^q (5/4)^M M)$  for the estimation of the gradient variance  $\operatorname{Var}_\theta [\partial_k C(\boldsymbol{\theta})]$  for each  $k$  on a classical computer in the general scenario, where  $n$  is the number of qubits,  $M$  is the number of parameters in the variational quantum circuit, and  $p, q$  are some constants inherited from the Gottesman-Knill theorem.

## 2. Sampling efficiency for the general case

In this section, we extend our scheme to the most generic case, by considering an arbitrary probability distribution for the rotation angles  $\theta$ , and derive the corresponding sampling complexity. As before, one only needs to consider the 1- and 2-fold channels for a single-qubit  $Z$ -rotation gate. In what follows, let us define  $\mathbb{E}_\theta[e^{i\theta}] := r_1 + is_1$  and  $\mathbb{E}_\theta[e^{2i\theta}] := r_2 + is_2$ . Note that  $r_1 = \mathbb{E}_\theta[\cos \theta]$  and  $r_2 = \mathbb{E}_\theta[\cos 2\theta]$  are defined in the same way as for the symmetric case before, while  $s_1 = \mathbb{E}_\theta[\sin \theta]$  and  $s_2 = \mathbb{E}_\theta[\sin 2\theta]$  are in general nonzero since we no longer assume the distribution of  $\theta$  to be even.

### a. 1-fold channel

The expression of the 1-fold channel for a single-qubit  $Z$  rotation is given by Eq. (A3), which we develop below *without* assuming an even distribution in  $\theta$ . We get

$$\begin{aligned} & \mathbb{E}_\theta[\hat{R}_Z(\theta)\hat{\rho}\hat{R}_Z^\dagger(\theta)] \\ &= \hat{\Pi}_0\hat{\rho}\hat{\Pi}_0 + \hat{\Pi}_1\hat{\rho}\hat{\Pi}_1 + \mathbb{E}_\theta[e^{i\theta}]\hat{\Pi}_1\hat{\rho}\hat{\Pi}_0 \\ & \quad + \mathbb{E}_\theta[e^{-i\theta}]\hat{\Pi}_0\hat{\rho}\hat{\Pi}_1 \\ &= \hat{\Pi}_0\hat{\rho}\hat{\Pi}_0 + \hat{\Pi}_1\hat{\rho}\hat{\Pi}_1 + (r_1 + is_1)\hat{\Pi}_1\hat{\rho}\hat{\Pi}_0 \\ & \quad + (r_1 - is_1)\hat{\Pi}_0\hat{\rho}\hat{\Pi}_1 \\ &= \frac{1+r_1}{2}\mathcal{E}[\mathbb{1}](\hat{\rho}) + \frac{1-r_1}{2}\mathcal{E}[\hat{Z}](\hat{\rho}) \\ & \quad + \frac{s_1}{2}\mathcal{E}[\hat{S}^\dagger](\hat{\rho}) - \frac{s_1}{2}\mathcal{E}[\hat{S}](\hat{\rho}), \end{aligned} \quad (\text{D15})$$

where  $\hat{S} = \hat{\Pi}_0 + i\hat{\Pi}_1$  is the phase gate, and one can use this definition together with Eq. (A4) to verify the equation above.

Here, parameter  $s_1$  can be understood as a measure of asymmetry in the probability distribution of  $\theta$ . In the symmetric case, we have  $s_1 = 0$  and the sum reduces to the convex one given by Eq. (A5). Following the same procedure as in Appendix D 1, this (possibly nonconvex) linear combination of Clifford channels can be estimated via sampling, and the number of required samples should be scaled, according to the nonconvexity of the sum, by a factor  $\gamma = \prod_{k=1}^M \gamma_k$  [see the definitions in Eqs. (D1)–(D3) and (D9)]. We now derive an upper bound for  $\gamma_k^{(1)}$ , the scaling factor associated with a single (the  $k$ th) 1-fold  $Z$ -rotation channel that can be decomposed in the form of Eq. (D15) in general. We proceed by applying the same argument as in Eqs. (D11)–(D13):

$$\begin{aligned} \gamma_k^{(1)} &= \left| \frac{1+r_1}{2} \right| + \left| \frac{1-r_1}{2} \right| + \left| \frac{s_1}{2} \right| + \left| -\frac{s_1}{2} \right| \\ &= \left| \mathbb{E}_\theta \left[ \frac{1+\cos\theta}{2} \right] \right| + \left| \mathbb{E}_\theta \left[ \frac{1-\cos\theta}{2} \right] \right| + |\mathbb{E}_\theta[\sin\theta]| \end{aligned}$$

$$\begin{aligned} & \leq \mathbb{E}_\theta \left[ \left| \frac{1+\cos\theta}{2} \right| + \left| \frac{1-\cos\theta}{2} \right| + |\sin\theta| \right] \\ & \leq \sup_\theta \left\{ \left| \frac{1+\cos\theta}{2} \right| + \left| \frac{1-\cos\theta}{2} \right| + |\sin\theta| \right\} \\ & = 2. \end{aligned} \quad (\text{D16})$$

This implies that the number of samples  $K^{(1)}$  required for the estimation of the generic 1-fold channel [see Eq. (D14)] scales as

$$\begin{aligned} K^{(1)} &\sim O(\gamma^{(1)}M), \\ \gamma^{(1)} &= \prod_{k=1}^M \gamma_k^{(1)} \leq 2^M. \end{aligned} \quad (\text{D17})$$

Note that the bound derived above depends on the specific choice of the Clifford channels in the decomposition. As the Clifford group does not form a linearly independent set, it should be possible to find a different decomposition that yields a different upper bound and further optimize the complexity.

### b. 2-fold channel

The 2-fold channel for a single-qubit  $Z$  rotation is given by Eq. (A15):

$$\begin{aligned} \Phi_Z^{(2)}(\hat{\rho}) &= \mathbb{E}_\theta[\Xi\hat{\rho}\Xi^\dagger] + \mathbb{E}_\theta[\Gamma_\theta\hat{\rho}\Gamma_\theta^\dagger] \\ & \quad + \mathbb{E}_\theta[\Gamma_\theta\hat{\rho}\Xi^\dagger] + \mathbb{E}_\theta[\Xi\hat{\rho}\Gamma_\theta^\dagger]. \end{aligned} \quad (\text{D18})$$

For a generic probability distribution of  $\theta$ , we have

$$\begin{aligned} \mathbb{E}_\theta[\Xi\hat{\rho}\Xi^\dagger] &= \hat{\Pi}_{01}\hat{\rho}\hat{\Pi}_{01} + \hat{\Pi}_{10}\hat{\rho}\hat{\Pi}_{10} \\ & \quad + \hat{\Pi}_{01}\hat{\rho}\hat{\Pi}_{10} + \hat{\Pi}_{10}\hat{\rho}\hat{\Pi}_{01}, \\ \mathbb{E}_\theta[\Gamma_\theta\hat{\rho}\Gamma_\theta^\dagger] &= \hat{\Pi}_{00}\hat{\rho}\hat{\Pi}_{00} + \hat{\Pi}_{11}\hat{\rho}\hat{\Pi}_{11} \\ & \quad + r_2(\hat{\Pi}_{00}\hat{\rho}\hat{\Pi}_{11} + \hat{\Pi}_{11}\hat{\rho}\hat{\Pi}_{00}) \\ & \quad + is_2(\hat{\Pi}_{11}\hat{\rho}\hat{\Pi}_{00} - \hat{\Pi}_{00}\hat{\rho}\hat{\Pi}_{11}), \\ \mathbb{E}_\theta[\Gamma_\theta\hat{\rho}\Xi^\dagger] &= r_1(\hat{\Pi}_{00} + \hat{\Pi}_{11})\hat{\rho}(\hat{\Pi}_{01} + \hat{\Pi}_{10}) \\ & \quad + is_1(\hat{\Pi}_{11} - \hat{\Pi}_{00})\hat{\rho}(\hat{\Pi}_{01} + \hat{\Pi}_{10}), \\ \mathbb{E}_\theta[\Xi\hat{\rho}\Gamma_\theta^\dagger] &= \mathbb{E}_\theta[\Gamma_\theta\hat{\rho}\Xi^\dagger]^\dagger. \end{aligned}$$

As one can verify, the Choi representation of the above terms are all diagonal, so their sum can be represented via the  $M$  matrix as before:

$$M(\Phi_Z^{(2)}) = \begin{pmatrix} 1 & r_1 + is_1 & r_1 + is_1 & r_2 + is_2 \\ r_1 - is_1 & 1 & 1 & r_1 + is_1 \\ r_1 - is_1 & 1 & 1 & r_1 + is_1 \\ r_2 - is_2 & r_1 - is_1 & r_1 - is_1 & 1 \end{pmatrix}. \quad (\text{D19})$$



This can again be decomposed as a weighted sum of the channels  $\mathcal{E}[\mathbb{1}]$ ,  $\mathcal{E}[\hat{Z} \otimes \hat{Z}]$ ,  $\mathcal{E}[\hat{S} \otimes \hat{S}]$ , and  $\mathcal{E}[\hat{S}^\dagger \otimes \hat{S}^\dagger]$  given in Eq. (A18) and of the following Clifford channels:

$$\begin{aligned}
M(\mathcal{E}[\mathbb{1} \otimes \hat{S}]) &= \begin{pmatrix} 1 & i & 1 & i \\ -i & 1 & -i & 1 \\ 1 & i & 1 & i \\ -i & 1 & -i & 1 \end{pmatrix}, \\
M(\mathcal{E}[\hat{S} \otimes \mathbb{1}]) &= \begin{pmatrix} 1 & 1 & i & i \\ 1 & 1 & i & i \\ -i & -i & 1 & 1 \\ -i & -i & 1 & 1 \end{pmatrix}, \\
M(\mathcal{E}[\mathbb{1} \otimes \hat{S}^\dagger]) &= \begin{pmatrix} 1 & -i & 1 & -i \\ i & 1 & i & 1 \\ 1 & -i & 1 & -i \\ i & 1 & i & 1 \end{pmatrix}, \\
M(\mathcal{E}[\hat{S}^\dagger \otimes \mathbb{1}]) &= \begin{pmatrix} 1 & 1 & -i & -i \\ 1 & 1 & -i & -i \\ i & i & 1 & 1 \\ i & i & 1 & 1 \end{pmatrix}, \\
M(\mathcal{E}[\hat{Z} \otimes \hat{S}]) &= \begin{pmatrix} 1 & i & -1 & -i \\ -i & 1 & i & -1 \\ -1 & -i & 1 & i \\ i & -1 & -i & 1 \end{pmatrix}, \\
M(\mathcal{E}[\hat{S} \otimes \hat{Z}]) &= \begin{pmatrix} 1 & -1 & i & -i \\ -1 & 1 & -i & i \\ -i & i & 1 & -1 \\ i & -i & -1 & 1 \end{pmatrix}, \\
M(\mathcal{E}[\hat{Z} \otimes \hat{S}^\dagger]) &= \begin{pmatrix} 1 & -i & -1 & i \\ i & 1 & -i & -1 \\ -1 & i & 1 & -i \\ -i & -1 & i & 1 \end{pmatrix}, \\
M(\mathcal{E}[\hat{S}^\dagger \otimes \hat{Z}]) &= \begin{pmatrix} 1 & -1 & -i & i \\ -1 & 1 & i & -i \\ i & -i & 1 & -1 \\ -i & i & -1 & 1 \end{pmatrix}.
\end{aligned}$$

Note that the channels listed above are all diagonal in the Choi representation and hence the  $M$  matrices capture all their nonzero entries. Following the same reasoning as in Appendix A, we solve a linear system to obtain the following decomposition:

$$\begin{aligned}
\Phi_Z^{(2)}(\hat{\rho}) &= \frac{s_2}{8} (\mathcal{E}[\hat{S} \otimes \mathbb{1}](\hat{\rho}) + \mathcal{E}[\mathbb{1} \otimes \hat{S}](\hat{\rho})) \\
&\quad + \frac{s_2}{8} (\mathcal{E}[\hat{Z} \otimes \hat{S}^\dagger](\hat{\rho}) + \mathcal{E}[\hat{S}^\dagger \otimes \hat{Z}](\hat{\rho})) \\
&\quad - \frac{s_2}{8} (\mathcal{E}[\hat{S}^\dagger \otimes \mathbb{1}](\hat{\rho}) + \mathcal{E}[\mathbb{1} \otimes \hat{S}^\dagger](\hat{\rho})) \\
&\quad - \frac{s_2}{8} (\mathcal{E}[\hat{Z} \otimes \hat{S}](\hat{\rho}) + \mathcal{E}[\hat{S} \otimes \hat{Z}](\hat{\rho}))
\end{aligned}$$

$$\begin{aligned}
&+ \frac{1+r_2+2r_1}{4} \mathcal{E}[\mathbb{1}](\hat{\rho}) \\
&+ \frac{1+r_2-2r_1}{4} \mathcal{E}[\hat{Z} \otimes \hat{Z}](\hat{\rho}) \\
&+ \frac{1-r_2+2s_1}{4} \mathcal{E}[\hat{S} \otimes \hat{S}](\hat{\rho}) \\
&+ \frac{1-r_2-2s_1}{4} \mathcal{E}[\hat{S}^\dagger \otimes \hat{S}^\dagger](\hat{\rho}). \quad (\text{D20})
\end{aligned}$$

*Remark.*—Denoting by  $\widehat{\text{CNOT}} = \hat{\Pi}_0 \otimes \mathbb{1} + \hat{\Pi}_1 \otimes \hat{X}$  the controlled-NOT (CNOT) gate and by  $\widehat{\text{CNOT}}_X := (\hat{X} \otimes \hat{X}) \widehat{\text{CNOT}} (\hat{X} \otimes \hat{X})$  its conjugation by the  $\hat{X} \otimes \hat{X}$  gate, we have

$$\begin{aligned}
M(\mathcal{E}[\widehat{\text{CNOT}}(\hat{S} \otimes \hat{S})\widehat{\text{CNOT}}]) &= \begin{pmatrix} 1 & i & -1 & i \\ -i & 1 & i & 1 \\ -1 & -i & 1 & -i \\ -i & 1 & i & 1 \end{pmatrix}, \\
M(\mathcal{E}[\widehat{\text{CNOT}}_X(\hat{S} \otimes \hat{S})\widehat{\text{CNOT}}_X]) &= \begin{pmatrix} 1 & -i & 1 & i \\ i & 1 & i & -1 \\ 1 & -i & 1 & i \\ -i & -1 & -i & 1 \end{pmatrix}.
\end{aligned}$$

Again, by solving a linear system one finds another decomposition of the 2-fold channel that involves the above channels, namely,

$$\begin{aligned}
\Phi_Z^{(2)}(\hat{\rho}) &= \frac{s_2}{4} \mathcal{E}[\widehat{\text{CNOT}}(\hat{S} \otimes \hat{S})\widehat{\text{CNOT}}](\hat{\rho}) \\
&\quad - \frac{s_2}{4} \mathcal{E}[\hat{Z} \otimes \hat{S}](\hat{\rho}) \\
&\quad + \frac{s_2}{4} \mathcal{E}[\widehat{\text{CNOT}}_X(\hat{S} \otimes \hat{S})\widehat{\text{CNOT}}_X](\hat{\rho}) \\
&\quad - \frac{s_2}{4} \mathcal{E}[\mathbb{1} \otimes \hat{S}^\dagger](\hat{\rho}) \\
&\quad + \frac{1+r_2+2r_1}{4} \mathcal{E}[\mathbb{1}](\hat{\rho}) \\
&\quad + \frac{1+r_2-2r_1}{4} \mathcal{E}[\hat{Z} \otimes \hat{Z}](\hat{\rho}) \\
&\quad + \frac{1-r_2+2s_1}{4} \mathcal{E}[\hat{S} \otimes \hat{S}](\hat{\rho}) \\
&\quad + \frac{1-r_2-2s_1}{4} \mathcal{E}[\hat{S}^\dagger \otimes \hat{S}^\dagger](\hat{\rho}). \quad (\text{D21})
\end{aligned}$$

Similar to our treatment with the 1-fold channel, let us derive an upper bound for  $\gamma_k^{(2)}$ , the scaling factor for the number of samples required for the estimation of the

generic 2-fold  $k$ th  $Z$ -rotation channel:

$$\begin{aligned}
\gamma_k^{(2)} &= 4 \left| \frac{s_2}{8} \right| + 4 \left| -\frac{s_2}{8} \right| \\
&\quad + \left| \frac{1+r_2+2r_1}{4} \right| + \left| \frac{1+r_2-2r_1}{4} \right| \\
&\quad + \left| \frac{1-r_2+2s_1}{4} \right| + \left| \frac{1-r_2-2s_1}{4} \right| \\
&= |\mathbb{E}_\theta[\sin 2\theta]| \\
&\quad + \left| \mathbb{E}_\theta \left[ \frac{1 + \cos 2\theta + 2 \cos \theta}{4} \right] \right| \\
&\quad + \left| \mathbb{E}_\theta \left[ \frac{1 + \cos 2\theta - 2 \cos \theta}{4} \right] \right| \\
&\quad + \left| \mathbb{E}_\theta \left[ \frac{1 - \cos 2\theta + 2 \sin \theta}{4} \right] \right| \\
&\quad + \left| \mathbb{E}_\theta \left[ \frac{1 - \cos 2\theta - 2 \sin \theta}{4} \right] \right|. \\
&\leq \sup_\theta \left\{ |\sin 2\theta| \right. \\
&\quad + \left| \frac{1 + \cos 2\theta + 2 \cos \theta}{4} \right| \\
&\quad + \left| \frac{1 + \cos 2\theta - 2 \cos \theta}{4} \right| \\
&\quad + \left| \frac{1 - \cos 2\theta + 2 \sin \theta}{4} \right| \\
&\quad \left. + \left| \frac{1 - \cos 2\theta - 2 \sin \theta}{4} \right| \right\} \\
&= 1 + \sqrt{2}. \tag{D22}
\end{aligned}$$

This implies that the number of samples  $K^{(2)}$  required for the estimation of the generic 2-fold channel scales as

$$\begin{aligned}
K^{(2)} &\sim O(\gamma^{(2)} M), \\
\gamma^{(2)} &= \prod_{k=1}^M \gamma_k^{(2)} \leq (1 + \sqrt{2})^M, \tag{D23}
\end{aligned}$$

which is dominant over the complexity of the estimation of the 1-fold channel [Eq. (D17)] since  $1 + \sqrt{2} > 2$ .

Again, combining the above result with the Gottesman-Knill theorem, for a cost-function observable  $\hat{O}$  acting nontrivially on  $N_O$  qubits, our scheme implies a complexity of no more than  $O(n^p N_O^q (1 + \sqrt{2})^M M)$  for the estimation of the gradient variance  $\text{Var}_\theta [\partial_k C(\theta)]$  for each  $k$  on a classical computer in the most generic case, where  $n$  is the number of qubits,  $M$  is the number of parameters in the variational ansatz, and  $p, q$  are some constants inherited from the Gottesman-Knill theorem.

## APPENDIX E: $N$ -FOLD CHANNEL FOR A RANDOM $Z$ ROTATION

In this section we give a decomposition of the  $N$ -fold channel as a real sum of Clifford unitary channels. This allows us to extend our scheme to the estimation of  $N$ th-order quantities with a complexity scaling polynomially in both the number of variational parameters  $M$  and the system size  $n$  when the decomposition is convex, and exponential in  $M$  otherwise. We give a *sufficient* condition on the distribution of the random angle  $\theta$  for the decomposition to be a convex one.

Recall that, for any unitary  $\hat{U}$ , we defined  $\mathcal{E}[\hat{U}](\hat{\rho}) := \hat{U}\hat{\rho}\hat{U}^\dagger$ . In Eq. (D15) we obtained a decomposition of the 1-fold channel of a  $Z$  rotation in terms of Clifford unitary channels for a generic distribution of the random angle, namely,

$$\begin{aligned}
\mathbb{E}_\theta[\hat{R}_Z(\theta)\hat{\rho}\hat{R}_Z^\dagger(\theta)] &= \frac{1+r_1}{2}\mathcal{E}[\mathbb{1}](\hat{\rho}) + \frac{1-r_1}{2}\mathcal{E}[\hat{Z}](\hat{\rho}) \\
&\quad + \frac{s_1}{2}\mathcal{E}[\hat{S}^\dagger](\hat{\rho}) - \frac{s_1}{2}\mathcal{E}[\hat{S}](\hat{\rho}). \tag{E1}
\end{aligned}$$

More generally, we have

$$\begin{aligned}
\hat{R}_Z(\theta)\hat{\rho}\hat{R}_Z^\dagger(\theta) &= \frac{1 + \cos \theta}{2}\mathcal{E}[\mathbb{1}](\hat{\rho}) \\
&\quad + \frac{1 - \cos \theta}{2}\mathcal{E}[\hat{Z}](\hat{\rho}) \\
&\quad + \frac{\sin \theta}{2}\mathcal{E}[\hat{S}^\dagger](\hat{\rho}) \\
&\quad - \frac{\sin \theta}{2}\mathcal{E}[\hat{S}](\hat{\rho}) \tag{E2}
\end{aligned}$$

for any  $\theta \in \mathbb{R}$ . This can be seen as a consequence of Eq. (E1) for a Dirac probability measure centered at  $\theta$ . One can directly generalize this equation to obtain an expression of the  $N$ -fold channel as a real sum of Clifford unitary channels, as

$$\hat{R}_Z^{\otimes N}(\theta)\hat{\rho}\hat{R}_Z^{\otimes N\dagger}(\theta) = \sum_{I=(i_1, \dots, i_N)} \lambda_I(\theta) \mathcal{E} \left[ \bigotimes_{j=1}^N \hat{U}_{i_j} \right] (\hat{\rho}), \tag{E3}$$

where the sum goes over all the multi-indices  $I = (i_1, \dots, i_N) \in \{0, 1, 2, 3\}$ , and  $\hat{U}_0 = \mathbb{1}$ ,  $\hat{U}_1 = \hat{Z}$ ,  $\hat{U}_2 = \hat{S}$ , and  $\hat{U}_3 = \hat{S}^\dagger$ . The coefficient  $\lambda_I(\theta)$  for a multi-index  $I$  representing a product of numbers  $m_i$  of the  $\hat{U}_i$  gates is given by

$$\begin{aligned}
\lambda_I(\theta) &= \frac{1}{2^N} (1 + \cos \theta)^{m_0} (1 - \cos \theta)^{m_1} \\
&\quad \times \sin^{m_2}(-\theta) \sin^{m_3}(\theta) \tag{E4}
\end{aligned}$$





## APPENDIX F: EXAMPLE OF FIRST- AND SECOND-ORDER CLIFFORD APPROXIMANT CIRCUITS FOR A SIMPLE ANSATZ

In this appendix we provide a sample of Clifford approximant circuits for the estimation of  $\mathbb{E}_\theta[C(\theta)]$  and  $\mathbb{E}_\theta[C(\theta)^2]$  for the simple circuit depicted in Fig. 10. The generalization to Clifford approximants for other quantities, such as the expectation of the squared gradient, can be derived from that example as it suffices to introduce the adequate Clifford gates to the fixed layers to obtain the right estimators (see Secs. II A and B). This circuit acts on three qubits and is composed of two layers of rotations that are alternated with fixed two-qubit control- $Z$  gates. To obtain a first-order approximant for these circuits, it suffices to randomly replace each rotation by either the identity gate (a wire) or the Pauli gate corresponding to the direction of the concerned rotation gate. Three examples of the first-order Clifford approximant are represented in Fig. 11. The second-order approximants are derived by first mapping each rotation along  $X$  or  $Y$  to a rotation along  $Z$ , making use of the identities  $\hat{X} = \hat{H}^\dagger \hat{Z} \hat{H}$  and  $\hat{Y} = (\hat{S} \hat{H}) \hat{Z} (\hat{S} \hat{H})^\dagger$ , where  $\hat{H}, \hat{S}$  are respectively the Hadamard and phase gates. As a result, we get the ansatz with layers of  $Z$  rotations alternated with fixed layers composed of Clifford gates represented in Fig. 12. This circuit is then doubled vertically to give a circuit acting on six qubits. Finally, each pair of rotations sharing the same angle is randomly replaced by two single-qubit gates according to the scheme of Fig. 2. Examples of the resulting Clifford approximant circuits are provided in Fig. 13.

- 
- [1] 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).
- [2] 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).
- [3] M. Cerezo, G. Verdon, H.-Y. Huang, L. Cincio, and P. J. Coles, Challenges and opportunities in quantum machine learning, *Nat. Comput. Sci.* **2**, 567 (2022).
- [4] 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).
- [5] 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).
- [6] GOOGLE AI QUANTUM AND COLLABORATORS, F. Arute *et al.*, Hartree-Fock on a superconducting qubit quantum computer, *Science* **369**, 1084 (2020).
- [7] E. Farhi, J. Goldstone, and S. Gutmann, A quantum approximate optimization algorithm, *arXiv:1411.4028* (2014).
- [8] N. Lacroix, C. Hellings, C. K. Andersen, A. Di Paolo, A. Remm, S. Lazar, S. Krinner, G. J. Norris, M. Gabureac, J. Heinsoo, A. Blais, C. Eichler, and A. Wallraff, Improving the performance of deep quantum optimization algorithms with continuous gate sets, *PRX Quantum* **1**, 020304 (2020).
- [9] M. P. Harrigan *et al.*, Quantum approximate optimization of non-planar graph problems on a planar superconducting processor, *Nat. Phys.* **17**, 332 (2021).
- [10] 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).
- [11] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, Connecting ansatz expressibility to gradient magnitudes and barren plateaus, *PRX Quantum* **3**, 010313 (2022).
- [12] S. Wang, E. Fontana, M. Cerezo, K. Sharma, A. Sone, L. Cincio, and P. J. Coles, Noise-induced barren plateaus in variational quantum algorithms, *Nat. Commun.* **12**, 6961 (2021).
- [13] C. Ortiz Marrero, M. Kieferová, and N. Wiebe, Entanglement-induced barren plateaus, *PRX Quantum* **2**, 040316 (2021).
- [14] A. V. Uvarov and J. D. Biamonte, On barren plateaus and cost function locality in variational quantum algorithms, *J. Phys. A: Math. Theor.* **54**, 245301 (2021).
- [15] 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).
- [16] T. L. Patti, K. Najafi, X. Gao, and S. F. Yelin, Entanglement devised barren plateau mitigation, *Phys. Rev. Res.* **3**, 033090 (2021).
- [17] R. Wiersema, C. Zhou, J. F. Carrasquilla, and Y. B. Kim, Measurement-induced entanglement phase transitions in variational quantum circuits, *arXiv:2111.08035* (2021).
- [18] J. Kim and Y. Oz, Entanglement diagnostics for efficient quantum computation, *J. Stat. Mech.: Theory Exp.* **2022**, 073101 (2022).
- [19] J. Kim and Y. Oz, Quantum energy landscape and circuit optimization, *Phys. Rev. A* **106**, 052424 (2022).
- [20] S. H. Sack, R. A. Medina, A. A. Michailidis, R. Kueng, and M. Serbyn, Avoiding barren plateaus using classical shadows, *PRX Quantum* **3**, 020365 (2022).
- [21] L. Friedrich and J. Maziero, Avoiding barren plateaus with classical deep neural networks, *Phys. Rev. A* **106**, 042433 (2022).
- [22] E. Grant, L. Wossnig, M. Ostaszewski, and M. Benedetti, An initialization strategy for addressing barren plateaus in parametrized quantum circuits, *Quantum* **3**, 214 (2019).
- [23] H.-Y. Liu, Z.-Y. Chen, T.-P. Sun, Y.-C. Wu, Y.-J. Han, and G.-P. Guo, Mitigating barren plateaus with transfer-learning-inspired parameter initializations, *arXiv:2112.10952* (2022).
- [24] K. Mitarai, Y. Suzuki, W. Mizukami, Y. O. Nakagawa, and K. Fujii, Quadratic Clifford expansion for efficient benchmarking and initialization of variational quantum algorithms, *Phys. Rev. Res.* **4**, 033012 (2022).
- [25] G. S. Ravi, P. Gokhale, Y. Ding, W. M. Kirby, K. N. Smith, J. M. Baker, P. J. Love, H. Hoffmann, K. R. Brown, and F. T. Chong, CAFQA: A classical simulation bootstrap for variational quantum algorithms, *arXiv:2202.12924* (2022).

- [26] J. Kim, J. Kim, and D. Rosa, Universal effectiveness of high-depth circuits in variational eigenproblems, *Phys. Rev. Res.* **3**, 023203 (2021).
- [27] J. Kim, Y. Oz, and D. Rosa, Quantum chaos and circuit parameter optimization, [arXiv:2201.01452](https://arxiv.org/abs/2201.01452) (2022).
- [28] M. H. Cheng, K. E. Khosla, C. N. Self, M. Lin, B. X. Li, A. C. Medina, and M. S. Kim, Clifford circuit initialisation for variational quantum algorithms, [arXiv:2207.01539](https://arxiv.org/abs/2207.01539) (2022).
- [29] J. Dborin, F. Barratt, V. Wimalaweera, L. Wright, and A. G. Green, Matrix product state pre-training for quantum machine learning, *Quantum Sci. Technol.* **7**, 035014 (2022).
- [30] A. Pesah, M. Cerezo, S. Wang, T. Volkoff, A. T. Sornborger, and P. J. Coles, Absence of barren plateaus in quantum convolutional neural networks, *Phys. Rev. X* **11**, 041011 (2021).
- [31] L. Schatzki, M. Larocca, Q. T. Nguyen, F. Sauvage, and M. Cerezo, Theoretical guarantees for permutation-equivariant quantum neural networks, [arXiv:2210.09974](https://arxiv.org/abs/2210.09974) (2022).
- [32] Z. Holmes, A. Arrasmith, B. Yan, P. J. Coles, A. Albrecht, and A. T. Sornborger, Barren plateaus preclude learning scramblers, *Phys. Rev. Lett.* **126**, 190501 (2021).
- [33] A. Arrasmith, M. Cerezo, P. Czarnik, L. Cincio, and P. J. Coles, Effect of barren plateaus on gradient-free optimization, *Quantum* **5**, 558 (2021).
- [34] A. Arrasmith, Z. Holmes, M. Cerezo, and P. J. Coles, Equivalence of quantum barren plateaus to cost concentration and narrow gorges, *Quantum Sci. Technol.* **7**, 045015 (2022).
- [35] S. Wang, P. Czarnik, A. Arrasmith, M. Cerezo, L. Cincio, and P. J. Coles, Can error mitigation improve trainability of noisy variational quantum algorithms?, [arXiv:2109.01051](https://arxiv.org/abs/2109.01051) (2021).
- [36] Y. Du, T. Huang, S. You, M.-H. Hsieh, and D. Tao, Quantum circuit architecture search for variational quantum algorithms, *Npj Quantum Inf.* **8**, 1 (2022).
- [37] K. Sharma, M. Cerezo, L. Cincio, and P. J. Coles, Trainability of dissipative perceptron-based quantum neural networks, *Phys. Rev. Lett.* **128**, 180505 (2022).
- [38] G. De Palma, M. Marvian, C. Rouzé, and D. S. França, Limitations of variational quantum algorithms: A quantum optimal transport approach, *PRX Quantum* **4**, 010309 (2023).
- [39] V. Heyraud, Z. Li, Z. Denis, A. Le Boité, and C. Ciuti, Noisy quantum kernel machines, *Phys. Rev. A* **106**, 052421 (2022).
- [40] S. Jerbi, L. J. Fiderer, H. Poulsen Nautrup, J. M. Kübler, H. J. Briegel, and V. Dunjko, Quantum machine learning beyond kernel methods, *Nat. Commun.* **14**, 517 (2023).
- [41] Z. Li, V. Heyraud, K. Donatella, Z. Denis, and C. Ciuti, Machine learning via relativity-inspired quantum dynamics, *Phys. Rev. A* **106**, 032413 (2022).
- [42] M. Schuld, Supervised quantum machine learning models are kernel methods, [arXiv:2101.11020](https://arxiv.org/abs/2101.11020) (2021).
- [43] P. Rebentrost, M. Mohseni, and S. Lloyd, Quantum support vector machine for big data classification, *Phys. Rev. Lett.* **113**, 130503 (2014).
- [44] P. Mujal, R. Martínez-Peña, J. Nokkala, J. García-Beni, G. L. Giorgi, M. C. Soriano, and R. Zambrini, Opportunities in quantum reservoir computing and extreme learning machines, *Adv. Quantum Technol.* **4**, 2100027 (2021).
- [45] Z. Denis, I. Favero, and C. Ciuti, Photonic kernel machine learning for ultrafast spectral analysis, *Phys. Rev. Appl.* **17**, 034077 (2022).
- [46] G. Marcucci, D. Pierangeli, and C. Conti, Theory of neuro-morphic computing by waves: Machine learning by rogue waves, dispersive shocks, and solitons, *Phys. Rev. Lett.* **125**, 093901 (2020).
- [47] D. Pierangeli, G. Marcucci, and C. Conti, Photonic extreme learning machine by free-space optical propagation, *Photonics Research* **9**, 1446 (2021).
- [48] S. Thanasilp, S. Wang, M. Cerezo, and Z. Holmes, Exponential concentration and untrainability in quantum kernel methods, [arXiv:2208.11060](https://arxiv.org/abs/2208.11060) (2022).
- [49] J. Watrous, *The Theory of Quantum Information* (Cambridge University Press, Cambridge, UK, 2018).
- [50] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition* (Cambridge University Press, Cambridge, UK, 2010).
- [51] D. Gottesman, The Heisenberg representation of quantum computers, [arXiv:quant-ph/9807006](https://arxiv.org/abs/quant-ph/9807006) (1998).
- [52] S. Aaronson and D. Gottesman, Improved simulation of stabilizer circuits, *Phys. Rev. A* **70**, 052328 (2004).
- [53] We denote by  $\hat{H}$  the Hadamard gate and by  $\hat{S}$  the phase gate, which both belong to the Clifford group. For  $X$  rotations, we have  $\hat{X} = \hat{H}^\dagger \hat{Z} \hat{H}$  and, hence,  $e^{-i\theta_i \hat{X}/2} = \hat{H}^\dagger e^{-i\theta_i \hat{Z}/2} \hat{H}$  and we can replace  $\hat{W}_i$  and  $\hat{W}_{i+1}$  by  $\hat{H} \hat{W}_i$  and  $\hat{W}_{i+1} \hat{H}$ , respectively, to get another ansatz with the same form as the original one and with only  $Y$  and  $Z$  rotations. We proceed likewise for  $Y$  rotations using the fact that  $\hat{Y} = (\hat{S} \hat{H}) \hat{Z} (\hat{S} \hat{H})^\dagger$ . Note that in the case of the last layer one of the extra gates must be absorbed in the cost-function observable to get the same ansatz structure.
- [54] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Quantum circuit learning, *Phys. Rev. A* **98**, 032309 (2018).
- [55] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, Evaluating analytic gradients on quantum hardware, *Phys. Rev. A* **99**, 032331 (2019).
- [56] D. A. Roberts and B. Yoshida, Chaos and complexity by design, *J. High Energy Phys.* **2017**, 121 (2017).
- [57] S. Sim, P. D. Johnson, and A. Aspuru-Guzik, Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms, *Adv. Quantum Technol.* **2**, 1900070 (2019).
- [58] K. Nakaji and N. Yamamoto, Expressibility of the alternating layered ansatz for quantum computation, *Quantum* **5**, 434 (2021).
- [59] D. Gross, K. Audenaert, and J. Eisert, Evenly distributed unitaries: On the structure of unitary designs, *J. Math. Phys.* **48**, 052104 (2007).
- [60] J. T. Iosue, K. Sharma, M. J. Gullans, and V. V. Albert, Continuous-variable quantum state designs: Theory and applications, [arXiv:2211.05127](https://arxiv.org/abs/2211.05127) (2022).
- [61] A. W. Harrow and R. A. Low, Random quantum circuits are approximate 2-designs, *Commun. Math. Phys.* **291**, 257 (2009).
- [62] F. G. S. L. Brandão, A. W. Harrow, and M. Horodecki, Local random quantum circuits are approximate polynomial-designs, *Commun. Math. Phys.* **346**, 397 (2016).
- [63] J. Haferkamp, Random quantum circuits are approximate unitary  $t$ -designs in depth  $O(nt^{5+o(1)})$ , *Quantum* **6**, 795 (2022).

- [64] 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).
- [65] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, edited by F. Bach, Adaptive Computation and Machine Learning Series (MIT Press, Cambridge, MA, USA, 2016).
- [66] J. Liu, Z. Lin, and L. Jiang, Laziness, barren plateau, and noise in machine learning, [arxiv:2206.09313](https://arxiv.org/abs/2206.09313) (2022).
- [67] E. R. Anschuetz and B. T. Kiani, Quantum variational algorithms are swamped with traps, *Nat. Commun.* **13**, 7760 (2022).
- [68] L. Bittel and M. Kliesch, Training variational quantum algorithms is NP-hard, *Phys. Rev. Lett.* **127**, 120502 (2021).
- [69] M. Cerezo and P. J. Coles, Higher order derivatives of quantum neural networks with barren plateaus, *Quantum Sci. Technol.* **6**, 035006 (2021).
- [70] This encompasses distributions that are symmetric about angle  $k\pi/2$  for  $k \in \mathbb{Z}$ . In this case the bias can be factored out in the form of an extra fixed Clifford gate.
- [71] C. Zhao and X.-S. Gao, Analyzing the barren plateau phenomenon in training quantum neural networks with the ZX-calculus, *Quantum* **5**, 466 (2021).
- [72] C. Piveteau, D. Sutter, and S. Woerner, Quasiprobability decompositions with reduced sampling overhead, *Npj Quantum Inf.* **8**, 1 (2022).
- [73] C. McDiarmid *et al.*, On the method of bounded differences, *Surveys Combinatorics* **141**, 148 (1989).
- [74] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning, Second Edition* (MIT Press, Cambridge, Massachusetts, USA, 2018).
- [75] B. Baumgartner, An inequality for the trace of matrix products, using absolute values, [arxiv:1106.6189](https://arxiv.org/abs/1106.6189) (2011).