Noise-resilient variational hybrid quantum-classical optimization

Laura Gentini [1,2], Alessandro Cuccoli,[1,2] Stefano Pirandola,[3] Paola Verrucchi,[4,1,2] and Leonardo Banchi [1,2]

[1]*Dipartimento di Fisica e Astronomia, Università di Firenze, I-50019, Sesto Fiorentino (FI), Italy*
[2]*INFN, Sezione di Firenze, I-50019, Sesto Fiorentino (FI), Italy*
[3]*Computer Science and York Centre for Quantum Technologies, University of York, York YO10 5GH, United Kingdom*
[4]*ISC-CNR, UOS Dipartimento di Fisica, Università di Firenze, I-50019, Sesto Fiorentino (FI), Italy*

Variational hybrid quantum-classical optimization is one of the most promising avenues to show the advantages of noisy intermediate-scale quantum computers in solving hard problems, such as finding the minimum-energy state of a Hamiltonian or solving some machine-learning tasks. In these devices, noise is unavoidable and impossible to error correct, yet its role in the optimization process is not well understood, especially from the theoretical viewpoint. Here we consider a minimization problem with respect to a variational state, iteratively obtained via a parametric quantum circuit, taking into account both the role of noise and the stochastic nature of quantum measurement outcomes. We show that the accuracy of the result obtained for a fixed number of iterations is bounded by a quantity related to the quantum Fisher information of the variational state. Using this bound, we study the convergence property of the quantum approximate optimization algorithm under realistic noise models, showing the robustness of the algorithm against different noise strengths.

I. INTRODUCTION

Quantum computers are physical devices that are expected to perform calculations essentially impossible for our best classical supercomputers [1], although the *quantum advantage* has been proven only for a specifically designed problem whose practical application is unclear. As for the hardware, the devices that are currently being built are noisy intermediate-scale quantum (NISQ) ones [2], for which many of the most promising uses consist of solving optimization problems via hybrid quantum-classical algorithms that include parametric quantum circuits [3–10]. In these algorithms, the manipulation of the quantum register is done via gates that depend on some parameters: These are iteratively updated via a feedback strategy, where the measurement outcomes of the device are classically processed to propose better parameters, in the spirit of a variational approach. In what follows, we will refer to the above procedure as a "hybrid variational optimization."

There are various aspects that make a real quantum device different from an ideal one, among which are the noise due to any external environment and the stochasticity of outcomes due to the probabilistic nature of the quantum measurement process. Different authors (see, for instance, Refs. [10–13]) have studied the effect of noise (e.g., noisy gates, dephasing etc.) in protocols designed for the noiseless case and found that noise is usually detrimental. At the same time, the role of stochasticity of outcomes has been described using the stochastic gradient descent framework [14,15]. However, how to tame the combined effect of noise and stochasticity in hybrid variational optimization is still far from being understood.

In this work, we analytically study the convergence properties of hybrid variational optimizations in terms of the number of times, hereafter dubbed iterations, that the NISQ device must be queried to find the optimal parameters with a desired accuracy. We focus on the effects of noisy gates and stochasticity of the measurement outcomes, without considering the further problem of choosing the measured observable that is best suited to extract information from the noisy process. We show that the convergence speed is typically unaffected by the presence of a small amount of noise, while the accuracy of the solution typically deteriorates as the noise strength increases. Moreover, we demonstrate that the attainable accuracy for a fixed number of iterations is bounded by a quantity that plays a very relevant role in quantum estimation theory, namely the quantum Fisher information [16–18]. Our theoretical prediction is corroborated by the results of numerical experiments.

The paper is structured as follows: In Sec. II, we introduce the variational hybrid optimization procedure and describe an algorithm that implements it. In Sec. III, we study the role of noisy operations and demonstrate some general results about their effects. Results of our numerical experiments are presented and discussed in Sec. IV, while conclusions are drawn in Sec. V.

II. VARIATIONAL HYBRID OPTIMIZATION

Consider the expectation value

$$C(\boldsymbol{\theta}) := \langle \psi(\boldsymbol{\theta})| \hat{H} |\psi(\boldsymbol{\theta})\rangle, \tag{1}$$

where $|\psi(\boldsymbol{\theta})\rangle$ is a quantum state of N qubits that depends on P classical parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P) \in \mathbb{R}^P$, and $\hat{H}$ is a Hermitian operator. Many optimization problems have their

solution in the minimization of $C(\boldsymbol{\theta})$ with regard to the parameters $\boldsymbol{\theta}$. When this minimization is obtained via a variational procedure, the state $|\psi(\boldsymbol{\theta})\rangle$ is dubbed a "variational quantum state." As for the operator $\hat{H}$, its explicit form depends on the specific problem under analysis. In the variational quantum eigensolver [3], for instance, $\hat{H}$ is the Hamiltonian of a quantum many-body system and the task is that of finding a good variational approximation of the ground state. In the quantum approximate optimization algorithm (QAOA) [4], the task is that of solving some combinatorial problem, with $\hat{H}$ an Ising-like Hamiltonian whose ground state embodies the solution of the problem [19]. Further, it is possible to express in this language some machine learning applications, such as quantum classifiers [10,20]. In all of these cases, the function to be minimized, $C(\boldsymbol{\theta})$, is dubbed the "cost" function and only needs to be bound from below, with a range of attainable values that depend on the actual physical status of the observable represented by the operator $\hat{H}$. There also exist optimization instances where $C(\boldsymbol{\theta})$ can have a slightly different meaning. In quantum control [21] and simulation [22], for example, the goal is often that of obtaining a quantum state, or quantum operation, that is as similar as possible to a given target one; in this case, one can choose $\hat{H} = \hat{U}|\varphi\rangle\langle\varphi|\hat{U}^{\dagger}$, with $\hat{U}$ a target unitary and $|\varphi\rangle$ a reference state: The function (1) is then the square of the fidelity of state preparation, ranging from 0 to 1 by definition, and $1 - C(\boldsymbol{\theta})$ the function to be minimized. In this work, we will not explicitly refer to this case, as done elsewhere [23], but rather focus on problems where $C(\boldsymbol{\theta})$ is the expectation value of some physical observable that needs being minimized.

At the heart of many variational hybrid optimization processes is the variational ansatz for the state $|\psi(\boldsymbol{\theta})\rangle$ in (1). One of the most popular choices is to take such state as the output of a *parametric quantum circuit*

$$|\psi(\boldsymbol{\theta})\rangle = e^{-i\theta_P \hat{X}_P} \dots e^{-i\theta_1 \hat{X}_1} |\psi_0\rangle, \qquad (2)$$

i.e., of a series of evolutions generated by different, and yet *fixed*, Hamiltonian operators $\hat{X}_j$, for times θ_j representing the variational parameters. The reason for this choice is that parametric quantum circuits are implementable in NISQ devices [2] as long as $\hat{X}_j$ contains one- and two-local interactions only, i.e., when the gates $e^{-i\theta_j \hat{X}_j}$ act nontrivially on at most two qubits. The state $|\psi_0\rangle$ is chosen among states that are easy to prepare and it is typically separable, $|\psi_0\rangle \equiv \bigotimes_{j=1}^{N} |\psi_0^{(j)}\rangle$.

Variational hybrid quantum-classical algorithms, schematically shown in Fig. 1, operate iteratively a quantum device and a classical processor. At the ith iteration, the quantum device generates the variational state $|\psi(\boldsymbol{\theta}^{(i)})\rangle = e^{-i\theta_P^{(i)}\hat{X}_P} \dots e^{-i\theta_1^{(i)}\hat{X}_1} |\psi(\boldsymbol{\theta}^{(i-1)})\rangle$ and estimates the cost (1), and possibly its derivatives $\partial_{\theta_j} C$, via quantum measurements [5,14,24] of the observable $\hat{H}$. This is the computationally hardest part, as it requires the manipulation of states that belong to Hilbert spaces whose dimension exponentially increases with the number of qubits N. Afterward, a classical algorithm processes the estimated values of $C(\boldsymbol{\theta}^{(i)})$, or derivatives $\partial_{\theta_j} C$, and proposes new parameters $\boldsymbol{\theta}^{(i+1)}$ that are expected to flow toward the minimum of the cost function. Classical optimization, quantum evolution, and quantum measurements are thus performed iteratively until convergence.
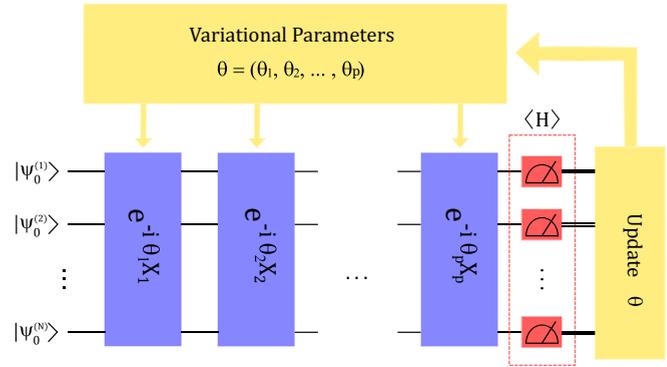
FIG. 1. Variationl hybrid quantum-classical optimization. A quantum computer is used to prepare the variational state (2) by sequentially applying some gates that depend on parameters θ_j and then measuring the observable $\hat{H}$ to estimate the cost (1). A classical algorithm iteratively processes these outcomes and updates the parameters θ_j to iteratively minimize the cost (1).

The advantage of this hybrid approach is that the quantum computer is always reset after each iteration so that the coherence times required are just those necessary to operate a circuit with depth $O(P)$ and then perform a measurement.

The main difference with other common variational approaches used in quantum mechanics is that $C(\boldsymbol{\theta})$, or derivatives $\partial_{\theta_j} C$, are estimated from measurement outcomes and, as such, are affected by uncertainty due to the probabilistic nature of quantum measurements, even in the noiseless case.

Having access to stochastic values of the cost function dramatically changes the convergence time [25]. Algorithms for stochastic optimization are classified as zeroth order, or derivative free, when only $C(\boldsymbol{\theta})$ is measured; first order when it is possible to directly measure the derivatives with regard to θ_j of the cost function; or, in general, kth order when also kth-order derivatives are available. It has been recently shown [14] that first-order methods can lead to substantially faster convergence than zeroth-order methods. On the other hand, the convergence time is not more strictly bounded when using higher order derivatives, although some advantage may be observed in practical implementations. Motivated by that analysis, here we focus on the convergence of first-order methods using the framework of stochastic optimization.

When dealing with stochastic optimization problems, where only the stochastic outcomes $f(\boldsymbol{\theta}, y)$ are directly measurable by sampling different values of y that are distributed according to a distribution $p(y|\boldsymbol{\theta})$, the cost function is usually written [14,25,26] as $C(\boldsymbol{\theta}) = \mathbb{E}_{y \sim p(y|\boldsymbol{\theta})}[f(\boldsymbol{\theta}, y)]$, where $\mathbb{E}_{y \sim p(y|\boldsymbol{\theta})}[f(\boldsymbol{\theta}, y)]$ is the expectation value $\sum_y p(y|\boldsymbol{\theta}) f(\boldsymbol{\theta}, y)$. The cost function (1) can be written in the above form by using the (possibly unknown) eigendecomposition of $\hat{H} \equiv \sum_y E_y |y\rangle\langle y|$: The measurement outcomes y are distributed with probability $p(y|\boldsymbol{\theta}) = \langle y| \hat{\rho}(\boldsymbol{\theta}) |y\rangle$, where $\hat{\rho}(\boldsymbol{\theta}) = |\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})|$ and $f(\boldsymbol{\theta}, y) = E_y$ is the associated cost, which is independent of $\boldsymbol{\theta}$.

When the eigendecomposition of $\hat{H}$ is not known, one can still get $C(\boldsymbol{\theta})$ from Pauli measurements, namely by decomposing $\hat{H}$ as $\hat{H} = \sum_{\mu=1}^{L} h_\mu \hat{\sigma}_\mu$, where each $\hat{\sigma}_\mu$ is a

tensor product of Pauli matrices and h_μ is the corresponding coefficient, and then by independently estimating each $\langle\psi(\boldsymbol{\theta})|\hat{\sigma}_\mu|\psi(\boldsymbol{\theta})\rangle$. Note that many $\hat{\sigma}_\mu$ typically commute with each other, so the required number of independent measurements can be smaller than L.

Suppose now that $\nabla C(\boldsymbol{\theta}) = \mathbb{E}_{z\sim q(z|\boldsymbol{\theta})}[\boldsymbol{g}(\boldsymbol{\theta}, z)]$, with $\nabla_j := \frac{\partial}{\partial\theta_j}$; i.e., that the gradient of C can be written as an expectation of a vector-valued function $\boldsymbol{g}(\boldsymbol{\theta}, z)$ over some stochastic outcomes z, distributed with a probability distribution q, possibly different from p. The simplest first-order method for stochastic optimization is stochastic gradient descent (SGD) that, intuitively, acts as a gradient descent algorithm where ∇C is substituted with an unbiased estimate $\boldsymbol{g}$. If the parameters are updated at each iteration i as $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \alpha_i \boldsymbol{g}(\boldsymbol{\theta}^{(i)})$, then, after I iterations, the algorithm converges [14,25,27,28] to a local optimum $\boldsymbol{\theta}^{\mathrm{opt}}$ with

$$\mathbb{E}[C(\boldsymbol{\theta}^{[1:I]})] - C(\boldsymbol{\theta}^{\mathrm{opt}}) \leqslant R\frac{G}{\sqrt{I}}. \tag{3}$$

The left-hand side of (3), where $\boldsymbol{\theta}^{[1:I]} = \frac{1}{I}\sum_{i=1}^{I}\boldsymbol{\theta}^{(i)}$, formally defines what we call "accuracy" in this work; in the right-hand side of the inequality R is a constant that depends on the function and on the parameter space, while G is an upper bound on the norm of the gradient estimate, $\mathbb{E}[\|\boldsymbol{g}(\boldsymbol{\theta})\|_2^2] \leqslant G^2$. Such rate is achieved with $\alpha_i \equiv \alpha = RI^{-1/2}/G$. The inequality (3) means that a larger gradient variance implies slower convergence. Note that, due to the stochastic nature of $\boldsymbol{g}$, even the parameters $\boldsymbol{\theta}^{(i)}$ are stochastic. On the other hand, Eq. (3) shows that $\boldsymbol{\theta}^{[1:I]}$ is a *good estimator* of the optimal value $\boldsymbol{\theta}^{\mathrm{opt}}$ in the limit of many iterations I and an arbitrarily small error $\epsilon \propto G/\sqrt{I}$ may be achieved. In other algorithms [14,25,26], the convergence depends on the bound $\mathbb{E}[\|\boldsymbol{g}(\boldsymbol{\theta})\|_\infty^2] \leqslant G_\infty^2$, obtained with a different norm. Since norm inequalities imply $G \leqslant \sqrt{P}G_\infty$, we can always focus on G_∞. Although different algorithms may have different convergence times, for instance with adaptive α_i and other definitions of $\boldsymbol{\theta}^{[1:I]}$, most upper bounds have a form similar to (3). Faster convergence, $\epsilon \approx G^2/I$, can be obtained when $C(\boldsymbol{\theta})$ satisfies extra properties [14,25], such as strong convexity, with a slightly different definition of $\boldsymbol{\theta}^{[1:I]}$. The bound (3) assumes that the parameters are updated after each query, namely after a single measurement outcome $\boldsymbol{g}$. An alternative is *minibatch learning* [25], where $M > 1$ queries are used to better estimate the gradient. Although this yields a less noisy gradient estimator, which, for instance, provides better numerical results in training quantum dynamical systems [29,30], the theoretical worst-case convergence rate is similar to (3). Indeed, a bound like (3) can be written with $I = MN_{\mathrm{iter}}$, with N_{iter} being the number of iterations and I being the total number of measurements.

III. HYBRID OPTIMIZATION WITH NOISY OPERATIONS

Because of the unavoidable errors in their operation, NISQ devices cannot exactly prepare the ideal variational state (2), which must hence be substituted with $\hat{\rho}(\boldsymbol{\theta}) = \mathcal{E}(\boldsymbol{\theta})[\hat{\rho}_0]$, where $\hat{\rho}_0$ is the noisy version of $|\psi_0\rangle$ and $\mathcal{E}(\boldsymbol{\theta})$ is the noisy dynamical map. Although most of our theoretical bounds hold for more complex noise models, for the sake of simplicity in the

following we use the decomposition

$$\hat{\rho}(\boldsymbol{\theta}) = \mathcal{E}_P^{\theta_P} \circ \cdots \circ \mathcal{E}_1^{\theta_1}[\hat{\rho}_0], \tag{4}$$

where $\circ$ indicates composition and $\mathcal{E}_j^{\theta_j}$ is the noisy version of the ideal parametric unitary channel $\mathcal{U}_j^{\theta_j}[\hat{\rho}] = e^{-i\theta_j\hat{X}_j}\hat{\rho}e^{i\theta_j\hat{X}_j}$ implemented by the jth parametric gate of the NISQ device. In what follows, $C_{\min} := \min_\psi\langle\psi|H|\psi\rangle$ is the exact minimum of the cost function. Since $\hat{\rho}(\boldsymbol{\theta})$ is a mixed state, the minimization of the cost function $C_{\mathrm{noisy}}(\boldsymbol{\theta}) := \mathrm{Tr}[\hat{\rho}(\boldsymbol{\theta})\hat{H}]$ only provides an approximation to the minimum $C(\boldsymbol{\theta}^{\mathrm{opt}})$ that can be obtained in the noiseless case. Although not explicitly considered in this paper, our analysis may be straightforwardly extended to some cases where even the ideal target is a mixed state, e.g., the nonequilibrium steady state of a noisy evolution [31–33].

The convergence rate of stochastic optimization toward the noisy minimum $C_{\mathrm{noisy}}(\boldsymbol{\vartheta}^{\mathrm{opt.}})$, with optimal parameters $\boldsymbol{\vartheta}^{\mathrm{opt}}$, can be bounded as in Eq. (3). Considering both the error due to the finite number of iterations and the error due to the difference between $C(\boldsymbol{\theta}^{\mathrm{opt}})$ and $C_{\mathrm{noisy}}(\boldsymbol{\vartheta}^{\mathrm{opt}})$, we may write

$$C_{\mathrm{noisy}}(\boldsymbol{\theta}^{[1:I]}) - C(\boldsymbol{\theta}^{\mathrm{opt}}) \leqslant \mathrm{Err}(\boldsymbol{\theta}^{\mathrm{opt}}, \boldsymbol{\vartheta}^{\mathrm{opt}}) + R\frac{G^{\mathrm{noisy}}}{\sqrt{I}}, \tag{5}$$

where we define $\mathrm{Err}(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ as the difference between the noisy and noiseless costs, namely

$$\mathrm{Err}(\boldsymbol{\theta}, \boldsymbol{\vartheta}) := C_{\mathrm{noisy}}(\boldsymbol{\vartheta}) - C(\boldsymbol{\theta}). \tag{6}$$

For some noise models, it can be shown that $\boldsymbol{\vartheta}^{\mathrm{opt}} = \boldsymbol{\theta}^{\mathrm{opt}}$, namely that the optimal parameters in the noiseless and noisy case are the same [13]. The inequality (5) shows a simple and yet important aspect: After a fixed number of iterations I, our best approximation to the noiseless variational minimum has an error that is given by two different terms. The first one follows from the difference between the noiseless and noisy case, while the second one depends on the gradient estimator and always decreases for increasing I. To simplify our discussion and provide a worst-case scenario, we assume that we know how to choose an ideal variational ansatz (2) that provides $C_{\min} = C(\boldsymbol{\theta}^{\mathrm{opt}})$ and consequently ensures $\mathrm{Err}(\boldsymbol{\vartheta}^{\mathrm{opt}}, \boldsymbol{\theta}^{\mathrm{opt}}) \geqslant 0$. This is typically not the case, as variational ansatze are normally chosen as simple circuits that are easy to implement on a NISQ device, for which one might get a negative $\mathrm{Err}(\boldsymbol{\vartheta}^{\mathrm{opt}}, \boldsymbol{\theta}^{\mathrm{opt}})$. The worst-case error coming from the first term in the right-hand side of (5) can be bounded by adapting the "peeling" technique from Refs. [34,35]. Indeed, we show in Appendix A that $\mathrm{Err}(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \leqslant P\|\hat{H}\|_\infty \max_k \|\mathcal{E}_k^{\vartheta_k} - \mathcal{U}_k^{\theta_k}\|_\diamond$ so the error increases at most linearly with the depth P and depends on the maximum distance, as measured by the diamond norm [36,37], between the ideal gates and their noisy implementations; see also Eq. (A4). An alternative inequality $\mathrm{Err}(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \leqslant 2\|\hat{H}\|_\infty\sqrt{1 - \langle\psi(\boldsymbol{\theta})|\hat{\rho}(\boldsymbol{\vartheta})|\psi(\boldsymbol{\theta})\rangle}$ shows that the first error term is bounded by the fidelity between the optimal pure state and its noisy version.

We now focus on G^{noisy} in (5), which depends on the procedure used to estimate the gradient from quantum measurements. The measurement of an observable with associated operator $\hat{g}_j$ provides an unbiased estimator of the gradient if $\nabla_j C = \mathrm{Tr}[\hat{\rho}\hat{g}_j]$ for each j. In this sense, we refer to the observables $\hat{g}_j$ as *estimators* of the gradient. In the noiseless case, different estimators have been proposed [5,14,20,24],

based on either the Hadamard test or the so-called parameter-shift rule. However, those estimators may be biased if only noisy gates are available: Therefore, a rigorous generalization to the noisy regime is still lacking. The convergence of SGD with biased gradient estimators is not much understood, aside from specific algorithms such as the simultaneous perturbation stochastic approximation (SPSA) [38], where the bias can be controlled. In order to define an unbiased estimator in the general case, we use the geometry of quantum states, from which we know that any derivative can be written as [17,39]

$$\nabla_j \hat{\rho} = \frac{\hat{L}_j \hat{\rho} + \hat{\rho} \hat{L}_j}{2}, \qquad (7)$$

where the operator $\hat{L}_j$ is called the symmetric logarithmic derivative (SLD). The gradient of the cost $C(\boldsymbol{\theta}) = \mathrm{Tr}[\hat{\rho}(\boldsymbol{\theta})\hat{H}]$ can hence be obtained by measuring observables with associated operators

$$\hat{g}_j(\boldsymbol{\theta}) = \frac{\hat{L}_j(\boldsymbol{\theta})\hat{H} + \hat{H}\hat{L}_j(\boldsymbol{\theta})}{2} + \lambda_j \hat{L}_j(\boldsymbol{\theta}) \qquad (8)$$

for any λ_j. The freedom in choosing λ_j follows from Eq. (7), as $\mathrm{Tr}[\hat{L}_j \hat{\rho}] = \mathrm{Tr}[\nabla_j \hat{\rho}] = \nabla_j \mathrm{Tr}[\hat{\rho}] = 0$ implies that the expectation value $\nabla_j C = \mathrm{Tr}[\hat{g}_j \hat{\rho}]$ is independent of λ_j. Therefore, the free parameters λ_j are analogous to the so-called baselines, commonly employed in reinforcement learning for variance reduction [40]. The optimal λ_js are discussed in Appendix B. The measurement of the gradient operators provides stochastic outcomes $g_j^{\mathrm{SLD}}(\boldsymbol{\theta}, \gamma)$ with probabilities $\langle g_{\gamma,j} | \hat{\rho} | g_{\gamma,j} \rangle$, where we used the eigendecomposition $\hat{g}_j = \sum_\gamma g_j^{\mathrm{SLD}}(\boldsymbol{\theta}, \gamma) | g_{\gamma,j} \rangle \langle g_{\gamma,j} |$. For pure states, the SLD operator has a simple form $\hat{L}_j = |\psi(\boldsymbol{\theta})\rangle \langle \nabla_j \psi(\boldsymbol{\theta})|$ and the above estimation strategy becomes equivalent to others, proposed in the literature [5,14,20,24], which can be explicitly measured using a generalization of the Hadamard test [14].

An alternative estimator can be obtained using the log-derivative (LD) method [41], also called "reinforce" in the machine learning literature [42], which consists in writing the gradient of the cost function $\nabla_j C = \sum_y E_y \nabla_j p(y|\boldsymbol{\theta})$ as en expectation value of $g_j^{\mathrm{LD}}(\boldsymbol{\theta}, y) = E_y \nabla_j \ln p(y|\boldsymbol{\theta})$ over the original distribution $p(y|\boldsymbol{\theta}) = \langle y| \hat{\rho}(\boldsymbol{\theta}) |y\rangle$, where $\hat{H} = \sum_y E_y |y\rangle\langle y|$.

In Appendix C, we show that all different estimators for the gradient satisfy the upper bound

$$G^{\mathrm{noisy}} \leqslant \sqrt{P} G_\infty^{\mathrm{noisy}} \leqslant \sqrt{P} \|\hat{H}\|_\infty \max_{j,\boldsymbol{\theta}} \sqrt{F_j(\boldsymbol{\theta})}, \qquad (9)$$

where F_j is the Quantum Fisher Information (QFI),

$$F_j(\boldsymbol{\theta}) = \mathrm{Tr}[\hat{\rho}(\boldsymbol{\theta})\hat{L}_j(\boldsymbol{\theta})^2], \qquad (10)$$

a central quantity in quantum metrology [17] that is also relevant for studying quantum phase transitions [33,43,44].

Before proceeding further, let us briefly comment upon the emergence of the QFI in this context. There are two slightly different way of understanding the QFI. First, there is the QFI of a state $\hat{\rho}$ *with respect to a certain observable* $\hat{X}$: This is $F_Q[\hat{\rho}, \hat{X}] := 2 \sum_{mn} \frac{(\rho_m - \rho_n)^2}{(\rho_m + \rho_n)} |\langle m| \hat{X} |n\rangle|^2$, where $\hat{\rho} = \sum_n \rho_n |n\rangle\langle n|$ is the eigendecomposition of $\hat{\rho}$. When $\hat{\rho}$ is a pure state, the QFI is nothing but twice the expectation value of $\hat{X}$ on such state. On the other hand, if a reference state $\hat{\rho}_0$ and a
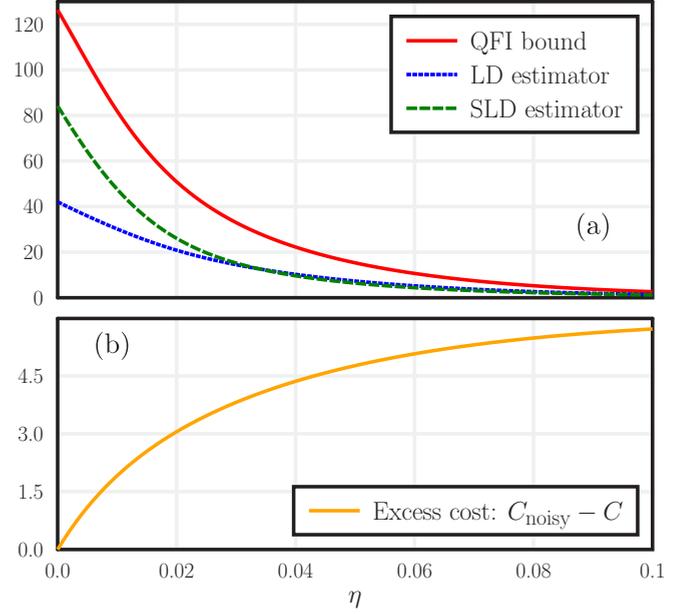
FIG. 2. Opposite behavior of the two sources of error: (a) the square root of second statistical moment of the gradient estimator $\max_{\boldsymbol{\theta}} \sqrt{\mathbb{E}[\|\boldsymbol{g}(\boldsymbol{\theta})\|_2^2]} \leqslant G_{\mathrm{noisy}}$ and (b) the excess cost $\mathrm{Err}(\boldsymbol{\theta}^{\mathrm{opt}}, \boldsymbol{\vartheta}^{\mathrm{opt}})$, as a function of the depolarising noise strength η (see Sec. IV for the definition of η). The variational circuit corresponds to a QAOA for a ring of $N = 6$ qubits with 20 variational parameters. Different gradient estimators are considered: The one based on the log-derivative (LD) method and the one based on the symmetric logarithmic derivative (SLD). Those are plotted against the upper bound (9) based on the quantum Fisher information (QFI).

real parameter θ (that we here take as one-dimensional for the sake of clarity) exist such that $\hat{\rho} = e^{-i\theta\hat{X}} \hat{\rho}_0 e^{i\theta\hat{X}} = \hat{\rho}(\theta)$, then it is $F_Q[\hat{\rho}, \hat{X}] = F_Q[\hat{\rho}(\theta)] = \mathrm{Tr}[\hat{\rho}(\theta)\hat{L}^2]$, where $\hat{L}$ is the SLD we have encountered above, whose relation with the observable $\hat{X}$ is via $i[\hat{\rho}, \hat{X}] = \frac{1}{2}(\hat{L}\hat{\rho} + \hat{\rho}\hat{L})$. The expression $F_Q[\hat{\rho}(\theta)]$ is common in quantum metrology, where it quantifies the sensitivity of the parametric state $\hat{\rho}(\theta)$ to the variations of θ itself. In this respect, notice that if θ represents the classical quantity to be measured via a quantum-metrology protocol, a higher QFI guarantees a more precise result. The above comments are easily generalized to the case of a multidimensional parameter $\boldsymbol{\theta}$, as the one used elsewhere in this work.

Let us now get back to the bound (9) to underline a very important aspect: While the first term in the right-hand side of (5) increases as a function of the noise strength, the second one can decrease. Indeed, it is known that noise is normally detrimental for metrology, as it can reduce the QFI from $O(N^2)$ (Heisenberg limit) to $O(N)$ (standard quantum limit) [18,45]. Our analysis thus shows that the convergence accuracy, as defined by the left-hand side of (5), is bounded by the sum of two terms that typically display opposite behaviors as a function of the noise strength, with the first one increasing and the second one decreasing, as shown in Fig. 2 for the specific example that will be described in the following section. From the physical point of view, this is due to how the cost-function landscape is affected by noise. As shown in Fig. 3 for strong noise, the landscape is "flattened" and
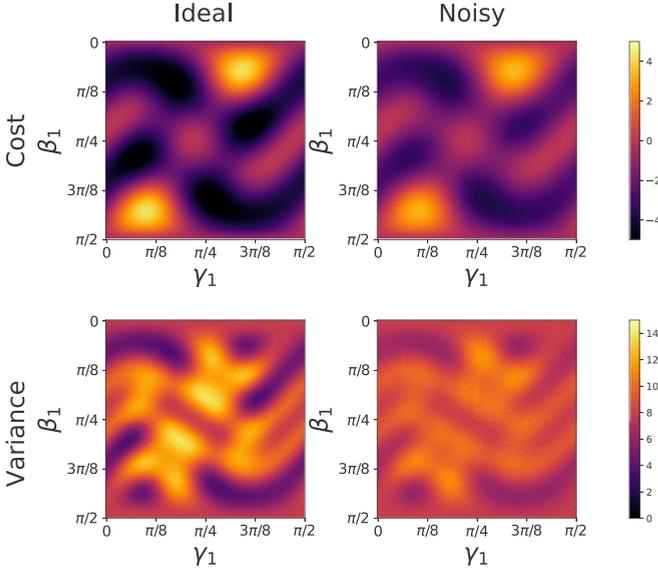
FIG. 3. Typical form of the cost-function landscape in the noiseless (left) to the noisy (right) regimes as a function of two parameters, γ_1 and β_1; see Sec. IV for the model description. Costs $\langle \hat{H} \rangle$ and variances $\langle \hat{H}^2 \rangle - \langle \hat{H} \rangle^2$ are respectively shown in the top and bottom panels.

hence the minimum cost is higher, but so is the variance, and hence the maximum variance is lower. Therefore, in the noisy regime the number of operations needed to correctly find the minimum may be reduced. Our analysis of Eq. (5) quantifies this intuition: When the number of iterations I is small, we may observe that noise is actually beneficial. However, due to measurement uncertainty, the number of iterations required for convergence is typically large and $\max_{j,\boldsymbol{\theta}} F_j$. In this regime noise is always detrimental, as observed in some numerical experiments [46,47].

The bound (5), together with the above analysis, shows that the convergence speed is mostly unaffected by noise, while the quality of the solution typically deteriorates due to the error $\mathrm{Err}(\boldsymbol{\theta}, \boldsymbol{\vartheta})$, which still increases at most linearly with the depth P of the circuit. Therefore, we may expect that hybrid variational optimization is robust against noise. In the next section, we numerically study such robustness for a relevant optimization problem.

IV. NUMERICAL EXAMPLES

QAOA [4] is a specific ansatz for variational hybrid optimization which consists in the repetition of two types of parametric quantum evolutions generated by two different noncommuting Hamiltonians, typically called $\hat{H}_\gamma$ and $\hat{H}_\beta$. Here $\hat{H}_\gamma \equiv \hat{H}$ is equal to the cost operator appearing in Eq. (1) and is a function of the Pauli $\hat{\sigma}_l^z$ operators only, where the indices $l = 1, \ldots, N$ refer to the different qubits. In the *computational basis* defined by the eigenstates $\{|0\rangle, |1\rangle\}$ of $\hat{\sigma}_l^z$, H is diagonal. The other Hamiltonian is fixed as $\hat{H}_\beta = -\sum_l \hat{\sigma}_l^x$, where $\hat{\sigma}_l^x$ are other Pauli operators, which are not diagonal in the computational basis. The QAOA evolution can be written as in Eq. (2) with sequential applications of $\hat{H}_\gamma$ and $\hat{H}_\beta$

$$|\psi(\boldsymbol{\gamma}, \boldsymbol{\beta})\rangle = e^{-i\beta_{\mathcal{P}}\hat{H}_\beta} e^{-i\gamma_{\mathcal{P}}\hat{H}_\gamma} \ldots e^{-i\beta_1\hat{H}_\beta} e^{-i\gamma_1\hat{H}_\gamma} |+\rangle^{\otimes N}. \quad (11)$$

The parameters are then split as $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta})$ and the total depth of the circuit is $P = 2\mathcal{P}$. The initial state $|\psi_0\rangle = |+\rangle^{\otimes N}$, where $|+\rangle = (|0\rangle + |1\rangle)/\sqrt{2}$, is the ground state of $\hat{H}_\beta$. QAOA is a universal model for quantum computation [48,49], meaning that, with specific choices of $\hat{H}_\gamma$, any state can be arbitrarily well approximated by $|\psi(\boldsymbol{\gamma}, \boldsymbol{\beta})\rangle$ with suitable parameters γ_j, β_j, and $\mathcal{P} \rightarrow \infty$. For the specific choice $\gamma_j \propto j/\mathcal{P}$ and $\beta_j \propto (1 - j/\mathcal{P})$, Eq. (11) can be interpreted as a discretization of an adiabatic evolution [4,50] and QAOA is guaranteed to perform well for large enough $\mathcal{P}$. The QFI can be large when the adiabatic evolution crosses a dynamical phase transition [33,43,44] and the error from G_{noisy} in (5) can be significant when the Hamiltonian $\beta\hat{H}_\beta + \gamma\hat{H}_\gamma$ displays a quantum phase transition for some choices of (β, γ). One such example is the Ising ring [51] studied below, where $\hat{H}_\beta$ models the global transverse field. In such a model, the QFI scales as $O(N^2)$ in the translational invariant case, while for steady states of more complex noisy evolution it can be as large as $O(N^6)$ [33].

Here we study QAOA applied to a translational invariant antiferromagnetic ring with $\hat{H}_\gamma = \sum_{l=1}^N \hat{\sigma}_l^z \hat{\sigma}_{l+1}^z$ and periodic boundary conditions $\hat{\sigma}_{N+1}^z \equiv \hat{\sigma}_1^z$. QAOA with this model has been studied in Refs. [11,12], using the exact mapping to a free-fermion model. In particular, it has been proven [11] that the ground state can be exactly expressed with the QAOA ansatz (11) as long as $\mathcal{P} \geqslant N/2$. The effect of noise in an overparameterized QAOA is shown in Fig. 2, where we consider the effect of a local depolarising error, as in (4) with $\mathcal{E}_j^{\theta_j}[\hat{\rho}] = \mathcal{D}[e^{-i\theta_j\hat{X}_j}\hat{\rho}e^{i\theta_j\hat{X}_j}]$, $\mathcal{D} = \bigotimes_{l=1}^N \mathcal{D}_l$, and $\mathcal{D}_l(\rho) = (1 - \eta)\hat{\rho} + \eta\hat{\sigma}_l^z\hat{\rho}\hat{\sigma}_l^z$. All bounds are computed by numerically finding the operators $\hat{L}_j$ from Eq. (7). In Fig. 2, we see that our theory predicts a decreasing G_{noisy} in (5) as a function of η. In Appendix D, we also study a different noise model, where the NISQ computer implements noisy yet unitary gates $e^{-i(\theta_j+\eta\epsilon_j)X_j}$, where $\epsilon_j \sim \mathcal{N}(0, 1)$ is a Gaussian random variable. We found that also with this noise, the error terms display the same behavior shown in Fig. 2.

We test our theoretical predictions using the QISKIT framework [52] to simulate QAOA on a physical hardware by numerical experiments. In these simulations, the error model consists of single- and two-qubit gate errors, i.e., depolarizing error followed by a thermal relaxation error, as discussed in Appendix E. In Fig. 4, we show the optimization of the cost function (1) using a simple gradient descend update, where each gradient value is estimated by using a quantum circuit followed by local measurements (see Appendix F for details); these operations are repeated 200 times to estimate a single value. Because of the stochastic measurement outcomes, each optimization takes a different stochastic route. Figure 4 is divided into three panels, differing from each other only for noise strength. In fact, in order to study the performance of QAOA in different noisy regimes, we progressively increase noise strength in our simulations: In the results reported in Fig. 4, noise strength has been progressively increased by factors $f = 2$ [Fig. 4(a)], factor $f = 4$ [Fig. 4(b)], and factor $f = 10$ [Fig. 4(c)]; namely gate errors and relaxation and decoherence rates are all increased by the factor f compared to the typical values of noise parameters obtained from IBM quantum processors (see Appendix E). Results obtained without increase, namely with $f = 1$, are not reported,
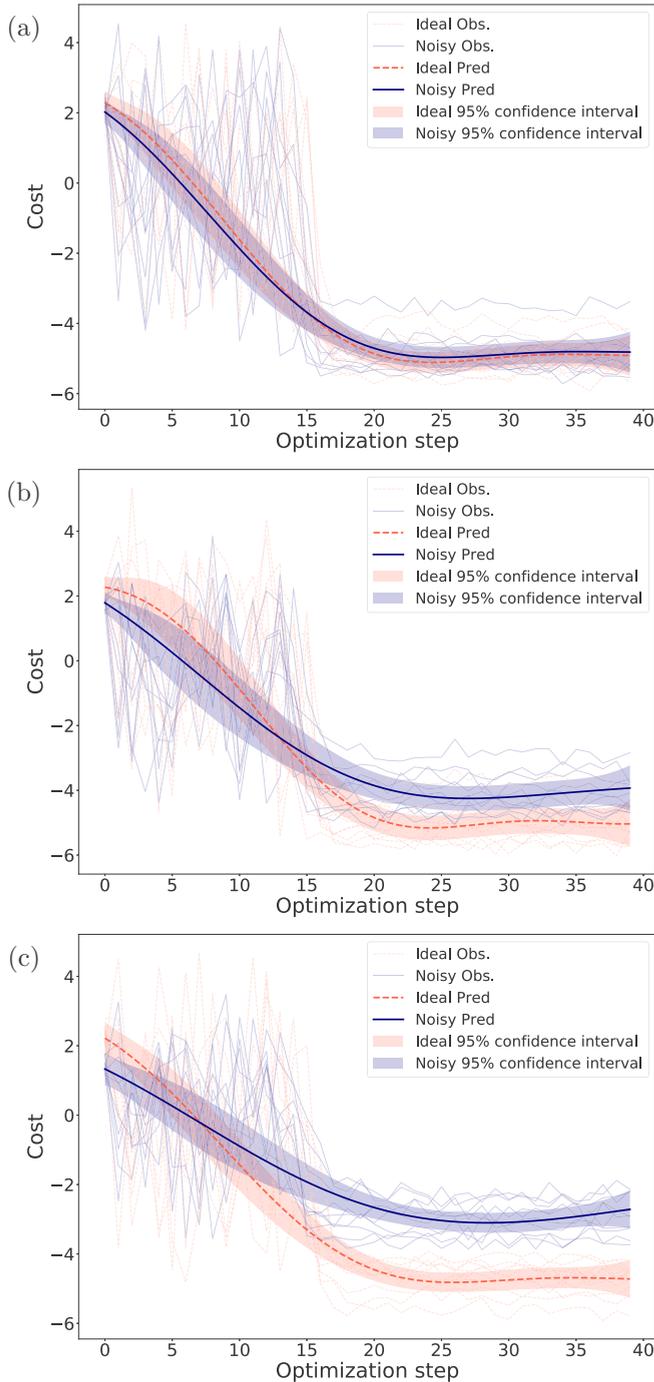
FIG. 4. Numerical results on the convergence of QAOA. We use $N = 8$ and $\mathcal{P} = 3$ in all panels (a), (b), and (c). Red and blue curves, respectively, represent noiseless and noisy evolution. The noise parameters are taken from typical values of IBM quantum processors (see Appendix E), yet increased by factors of 2 (a), 4 (b), and 10 (c). The convergence of the cost function (1) is shown during the optimization procedure; faded thin lines are 10 trials of the same realization of QAOA, while solid lines show the empirical average evolution from the 10 trials. For more information, see the main text.

since they are basically indistinguishable from those obtained in the noiseless case. We expect that noise effects can become visible even with $f = 1$ when the number of qubits N and/or

circuit depth P are sensibly larger than those employed in our simulations.

The following description holds for all three individual panels regardless of the increase factor f of noise strength: The fade thin lines correspond to 10 different trials of optimization of the cost function in both noisy (blue) and noiseless (red) cases with the same initial conditions. We observe that the cost, estimated like the gradient, repeating 200 times the quantum circuit and measurements, initially highly fluctuates among different runs, but then all runs converge toward the local minimum (since $\mathcal{P} < N/2$ the global minimum cannot be achieved). Higher accuracy in the convergence may be obtained using more sophisticated first-order methods [26]. Optimizations starting from different initial parameters display similar behavior (not shown). Using Gaussian process regression, we have also computed average "path-integral-like" optimization curves, which are shown in the same graph in Fig. 4 with solid lines for both the noiseless and noisy cases. These lines show the empirical average evolution from the 10 trials, and colored zones represent 95% confidence intervals associated with such average.

We observe that for small noise strength, e.g., Fig. 4(a), the average optimization curves are basically indistinguishable from the noisy ones, aside from finite-size effects due to the finite number of measurements. For moderate noise values, Fig. 4(b), as predicted by our theory, the final point in the optimization is slightly higher in the noisy regime, but the convergence speed is comparable. Moreover, at the initial points, the average cost is smaller in the noisy case. Although this prediction may be due to the finite number of curves, it is consistent with our analysis from (5). For large noise strengths, Fig. 4(c), we observe that the final cost significantly deviates from the noiseless case, but the convergence speed is again comparable.

Overall our analysis shows that the convergence speed of hybrid variational optimization is not significantly affected by noise, while the quality of the final solution is.

V. DISCUSSION

We have studied the convergence speed of variational hybrid quantum-classical optimization algorithms, showing that the error after a finite number of steps can be upper bounded by the sum of two terms: The first one is the difference between the noisy and the noiseless result, and typically increases for stronger noise; the second term, though, is proportional to the square root of the quantum Fisher information that usually decreases with noise. Because of the competition between these two terms, depending on the noise strength, different results can be observed. For small to intermediate noise, we find that the convergence time is mostly unaffected by the presence of noise. This said, for a number of iterations such that the convergence time has not passed yet, the accuracy may actually be higher in some noisy regimes. On the other hand, if the convergence time is reached, the error term typically dominates, showing the foreseeable negative role of noise.

Let us also comment upon the way QFI enters our results. From the formal viewpoint, we understand its occurrence as due to the use of stochastic optimization methods, which

involve the gradient of the cost function with respect to the variational parameters, and hence the operators $\hat{L}_j$ in (7) and QFI via its definition (10). However, the substantial reason for the QFI to appear in (9) in a somehow counterintuitive way (the lower the better) can be explained as follows. In estimation theory, a larger FI guarantees a better determination of the wanted parameter via the sampling of a function that depends on it; this is because a larger FI follows from larger local values of the derivatives and hence a higher sensitivity of the overall estimation procedure. On the other hand, in the scheme to which we are referring the role played by the parameter and the sampled function is reversed: We input different values of $\boldsymbol{\theta}$ aiming at exploring the $C(\boldsymbol{\theta})$ landscape, possibly locating its minimum; this exploration is more agile if the above landscape is more level, which corresponds to a lower FI. This general argument holds both in a classical and in a quantum setting, and we think it lies underneath the result Eq. (9) in the following sense: Noise can help an algorithmic procedure to more easily explore the landscape of the cost function one wants to minimize, thus increasing, at least as far as its detrimental effect on the cost-function evaluation is not too strong, the overall efficiency of the optimization scheme.

We finally underline that QFI enters our analysis by only providing a theoretical upper bound that never needs being evaluated. In fact, should the QFI be efficiently measurable, one could use more sophisticated stochastic algorithms, such as Amari's natural gradient [53]; this has been recently applied to noiseless parametric quantum circuits [54] based on the fact that, when $C = -\ln p(x, \boldsymbol{\theta})$, the natural gradient is Fisher efficient, i.e., such that the variance of the estimator $\boldsymbol{\theta}^{[1:I]}$ asimptotically meets the Cramér-Rao lower bound. However, such a result does not hold for more general cost functions like (1). Furthermore, no efficient method [e.g., poly(N)] for estimating the QFI from measurements is currently available in the noisy regime and, even if it existed, estimating the QFI at each step would require further quantum measurements that would increase the query complexity. In fact, understanding whether one can obtain Fisher efficient estimators of the optimal parameters is currently an open question.

APPENDIX A: BOUND ON Err($\boldsymbol{\theta}, \boldsymbol{\vartheta}$)

Many of our results hold irrespective of assumption (4), and are valid for any error model

$$\hat{\rho}(\boldsymbol{\theta}) = \mathcal{E}(\theta_1, \ldots, \theta_P)[\hat{\rho}_0]. \tag{A1}$$

Indeed, all results that we derive in Appendixes B and C do not depend on the assumption (4) and are valid for any noise model as in (A1). Here we show, on the other hand, that when the local error model (4) is assumed, then the error Err($\boldsymbol{\theta}, \boldsymbol{\vartheta}$) grows at most linearly with the number of parameters. We study an upper bound to the first error in (5), which is clearly valid irrespective of the sign of Err($\boldsymbol{\theta}, \boldsymbol{\vartheta}$)

$$
\begin{aligned}
\text{Err}(\boldsymbol{\theta}, \boldsymbol{\vartheta}) &:= \text{Tr}[\hat{H}(\hat{\rho}(\boldsymbol{\vartheta}) - |\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})|)] \\
&\overset{(a)}{\leqslant} \|\hat{H}\|_\infty \|\hat{\rho}(\boldsymbol{\vartheta}) - |\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})|\|_1 \\
&\overset{(b)}{\leqslant} \|\hat{H}\|_\infty \|\mathcal{E}(\boldsymbol{\vartheta}) - \mathcal{U}(\boldsymbol{\theta})\|_\diamond,
\end{aligned}
\tag{A2}
$$

where $\|\hat{X}\|_\infty$ is the maximum singular value of $\hat{X}$, namely the maximum absolute value $|x_j|$ where x_j are the eigenvalues of $\hat{X}$, $\|\hat{X}\|_1 = \text{Tr}[\sqrt{\hat{X}\hat{X}^\dagger}]$ is the trace norm, and $\|X\|_\diamond$ is the diamond norm for quantum channels [36,37]. In the last line, it is

$$\mathcal{U}(\boldsymbol{\theta}) := \mathcal{U}_P^{\theta_P} \circ \cdots \circ \mathcal{U}_1^{\theta_1}, \tag{A3}$$

and

$$\hat{\rho}(\boldsymbol{\vartheta}) = \mathcal{E}(\boldsymbol{\vartheta})[|\psi_0\rangle\langle\psi_0|], \quad |\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})| = \mathcal{U}(\boldsymbol{\theta})[|\psi_0\rangle\langle\psi_0|],$$

where for simplicity we have absorbed the noisy preparation of $|\psi_0\rangle$ into $\mathcal{E}_1$. To derive (A2), in (a) we used the Hölder inequality and in (b) we used the distance induced by the diamond norm

$$\|\mathcal{E} - \mathcal{U}\|_\diamond = \max_\rho \|\mathcal{I} \otimes \mathcal{E}(\rho) - \mathcal{I} \otimes \mathcal{U}(\rho)\|_1, \tag{A4}$$

where $\mathcal{I}$ is the identity channel. We can now apply the "peeling" technique from Refs. [34,35] to bound the error in the diamond distance. To this aim, we now use the decomposition from Eq. (4) from the main text, and let $\delta_P = \|\mathcal{E}_{1:P} - \mathcal{U}_{1:P}\|_\diamond$, where the 1:$k$ refers to the composition of the first k channels. Then, using the monotonicity of the diamond norm over CPTP maps and the triangle inequality, we may write

$$
\begin{aligned}
\delta_P &= \|\mathcal{E}_P \circ \mathcal{E}_{1:P-1} - \mathcal{E}_P \circ \mathcal{U}_{1:P-1} + \mathcal{E}_P \circ \mathcal{U}_{1:P-1} \\
&\quad - \mathcal{U}_P \circ \mathcal{U}_{1:P-1}\|_\diamond \\
&\leqslant \|\mathcal{E}_P \circ \mathcal{E}_{1:P-1} - \mathcal{E}_P \circ \mathcal{U}_{1:P-1}\|_\diamond \\
&\quad + \|\mathcal{E}_P \circ \mathcal{U}_{1:P-1} - \mathcal{U}_P \circ \mathcal{U}_{1:P-1}\|_\diamond \\
&\leqslant \delta_{P-1} + \|\mathcal{E}_P - \mathcal{U}_P\|_\diamond.
\end{aligned}
$$

Iteratively applying the above inequality, one gets

$$\delta_P \leqslant \sum_{k=1}^P \|\mathcal{E}_k - \mathcal{U}_k\|_\diamond \leqslant P \max_k \|\mathcal{E}_k - \mathcal{U}_k\|_\diamond. \tag{A5}$$

Combining (A5) and (A2), we find that the error increases at most linearly with P, according to

$$\text{Err}(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \leqslant P\|\hat{H}\|_\infty \max_k \|\mathcal{E}_k - \mathcal{U}_k\|_\diamond. \tag{A6}$$

An alternative bound can be obtained from (A2) via the Fuchs–van de Graaf inequality [55]

$$\text{Err}(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \leqslant 2\|\hat{H}\|_\infty \sqrt{1 - \langle\psi(\boldsymbol{\theta})| \hat{\rho}(\boldsymbol{\vartheta}) |\psi(\boldsymbol{\theta})\rangle}. \tag{A7}$$

APPENDIX B: OPTIMAL BASELINES

We discuss the role of the free parameters λ_j, dubbed "baselines," in the optimization. In principle, such parameters should be chosen to minimize $\mathbb{E}[g_j^2]$. We may write

$$
\begin{aligned}
\mathbb{E}\big[g_j^2\big] &\equiv \left\langle \left(\frac{\{\hat{H}, \hat{L}_j\}}{2} + \lambda_j \hat{L}_j \right)^2 \right\rangle_{\hat{\rho}(\boldsymbol{\theta})} \\
&= \left\langle \left(\frac{\{\hat{H}, \hat{L}_j\}}{2} \right)^2 + \lambda_j \frac{\{\hat{L}_j, \{\hat{H}, \hat{L}_j\}\}}{2} + \lambda_j^2 \hat{L}_j^2 \right\rangle_{\hat{\rho}(\boldsymbol{\theta})} \\
&= \left\langle \left(\frac{\{\hat{H}, \hat{L}_j\}}{2} \right)^2 + \lambda_j \frac{\{\hat{L}_j, \{\hat{H}, \hat{L}_j\}\}}{2} \right\rangle_{\hat{\rho}(\boldsymbol{\theta})} + \lambda_j^2 F_j,
\end{aligned}
$$

(B1)

where $\{\hat{A}, \hat{B}\} = \hat{A}\hat{B} + \hat{B}\hat{A}$. Since QFI is always positive, the optimal value of the "baseline" λ_j is the vertex of the above parabola, namely

$$
\lambda_j^{\text{opt}} = -\frac{\langle \{\hat{L}_j, \{\hat{H}, \hat{L}_j\}\} \rangle_{\hat{\rho}(\boldsymbol{\theta})}}{4 F_j}.
$$

(B2)

We note that the bound (C11) continues to hold even when the optimal baseline is used, as by definition $\mathbb{E}[g_j^2]$ with the optimal baseline is smaller than $\mathbb{E}[g_j^2]$ for the nonoptimal $\lambda_j = 0$.

APPENDIX C: BOUND ON G_{noisy}

We first focus on the estimator based on the log-derivative trick. We write the cost function as $C = \sum_y E_y p(y|\boldsymbol{\theta})$, where $p(y|\boldsymbol{\theta}) = \langle y | \hat{\rho}(\boldsymbol{\theta}) | y \rangle$, $\hat{H} = \sum_y E_y \hat{\Pi}_y$ is the possibly unknown eigendecomposition of H and $\hat{\Pi}_y = |y\rangle\langle y|$. Then

$$
\boldsymbol{\nabla}_j C = \mathbb{E}_{y \sim p(y|\boldsymbol{\theta})}[E_y \boldsymbol{\nabla}_j \ln p(y|\boldsymbol{\theta})].
$$

(C1)

From the above, we find that $g_j = E_y \boldsymbol{\nabla}_j \ln p(y|\boldsymbol{\theta})$ is an unbiased estimator of $\boldsymbol{\nabla}_j C$. We recall the definition of the constants G_{noisy} and G_∞ such that

$$
\mathbb{E}\left[\sum_j g_j^2 \right] \leqslant G_{\text{noisy}}^2, \qquad \max_j \mathbb{E}[g_j^2] \leqslant G_\infty^2.
$$

(C2)

To get those constants, we need to find upper bounds for $\mathbb{E}[g_j^2]$. By explicit calculation, following a similar derivation of Ref. [17] we find

$$
\mathbb{E}[g_j^2] = \sum_y E_y^2 p(y|\boldsymbol{\theta})[\boldsymbol{\nabla}_j \ln p(y|\boldsymbol{\theta})]^2
$$

(C3)

$$
= \sum_y E_y^2 \frac{[\boldsymbol{\nabla}_j p(y|\boldsymbol{\theta})]^2}{p(y|\boldsymbol{\theta})}
$$

(C4)

$$
\overset{(a)}{=} \sum_y E_y^2 \frac{[\text{Tr}\hat{\Pi}_y (\hat{\rho}\hat{L}_j + \hat{L}_j \hat{\rho})/2]^2}{\text{Tr}[\hat{\Pi}_y \hat{\rho}]}
$$

(C5)

$$
= \sum_y E_y^2 \frac{[\Re \text{Tr}(\hat{\Pi}_y \hat{\rho}\hat{L}_j)]^2}{\text{Tr}[\hat{\Pi}_y \hat{\rho}]}
$$

(C6)

$$
\leqslant \sum_y E_y^2 \frac{|\text{Tr}(\hat{\Pi}_y \hat{\rho}\hat{L}_j)|^2}{\text{Tr}[\hat{\Pi}_y \hat{\rho}]}
$$

(C7)

$$
= \sum_y E_y^2 \left| \text{Tr}\left(\frac{\sqrt{\hat{\Pi}_y}\sqrt{\hat{\rho}}}{\sqrt{\text{Tr}[\hat{\Pi}_y \hat{\rho}]}} \sqrt{\hat{\rho}}\hat{L}_j \sqrt{\hat{\Pi}_y} \right) \right|^2
$$

(C8)

$$
\overset{(b)}{\leqslant} \sum_y E_y^2 \text{Tr}(\hat{\Pi}_y \hat{L}_j \hat{\rho}\hat{L}_j)
$$

(C9)

$$
= \text{Tr}(\hat{H}^2 \hat{L}_j \hat{\rho}\hat{L}_j),
$$

(C10)

where we used in (*a*) the definition of the SLD (7) and in (*b*) the Cauchy-Schwartz inequality. Using then the Hölder inequality and the fact that $\hat{L}_j \hat{\rho}\hat{L}_j$ is a positive operator, we find then

$$
\mathbb{E}[g_j^2] \leqslant \|\hat{H}\|_\infty^2 \|\hat{L}_j \hat{\rho}\hat{L}_j\|_1 \leqslant \|\hat{H}\|_\infty^2 F_j,
$$

(C11)

where QFI$_j$ is the quantum Fisher information (10). The upper bounds (C2) then follow with

$$
G = \|\hat{H}\|_\infty \sqrt{P\left(\max_j F_j \right)},
$$

(C12)

$$
G_\infty = \|\hat{H}\|_\infty \sqrt{\max_j F_j}.
$$

(C13)

A similar bound is obtained with another unbiased estimator of the gradient. Here we set $\lambda_j = 0$, while the general case is studied in the next section. Using the SLD, we note that

$$
\boldsymbol{\nabla}_j C = \text{Tr}[\hat{H}(\hat{\rho}\hat{L}_j + \hat{L}_j \hat{\rho})/2] = \tfrac{1}{2}\langle \hat{H}\hat{L}_j + \hat{L}_j \hat{H} \rangle_{\hat{\rho}(\boldsymbol{\theta})}
$$

(C14)

$$
\equiv \langle \text{Re}(\hat{H}\hat{L}_j) \rangle_{\hat{\rho}(\boldsymbol{\theta})},
$$

(C15)

where $\langle \hat{A} \rangle_{\hat{\rho}} = \text{Tr}[\hat{\rho}\hat{A}]$, $\text{Re}[\hat{A}] := (\hat{A} + \hat{A}^\dagger)/2$, so the gradient can be estimated by quantum measurements of the operator $\text{Re}(\hat{H}\hat{L}_j)$. An upper bound is then obtained as

$$
\mathbb{E}[g_j^2] \equiv \langle \text{Re}(\hat{H}\hat{L}_j)^2 \rangle_{\hat{\rho}(\boldsymbol{\theta})}
$$

(C16)

$$
\leqslant \langle \text{Re}(\hat{H}\hat{L}_j)^2 + \text{Im}(\hat{H}\hat{L}_j)^2 \rangle_{\hat{\rho}(\boldsymbol{\theta})}
$$

(C17)

$$
= \tfrac{1}{2}\text{Tr}[\hat{\rho}(\hat{L}_j \hat{H}^2 \hat{L}_j + \hat{H}\hat{L}_j^2 \hat{H})]
$$

(C18)

$$
= \tfrac{1}{2}\text{Tr}[\hat{L}_j \hat{\rho}\hat{L}_j (\hat{H}^2 + \hat{L}_j^{-1} \hat{H}\hat{L}_j^2 \hat{H}\hat{L}_j^{-1})],
$$

(C19)

where we have assumed that $\hat{L}_j^{-1}$ exists. Using again the Hölder inequality, we get

$$
\mathbb{E}[g_j^2] \leqslant \tfrac{1}{2}\|\hat{L}_j \hat{\rho}\hat{L}_j\|_1 (\|\hat{H}\|_\infty^2 + \|\hat{L}_j^{-1} \hat{H}\hat{L}_j\|_\infty^2)
$$

(C20)

$$
\leqslant \|\hat{H}\|_\infty^2 F_j,
$$

(C21)

which is equivalent to Eq. (C11).

APPENDIX D: FLUCTUATING PARAMETERS

We consider an experimentally motivated noise model where the parameters θ_j cannot be tuned exactly. The lack of exact accuracy is modeled by a Gaussian noise with variance σ_j^2. This corresponds to the following substitution,

$$
\theta_j \to \mathcal{N}(\theta_j, \sigma_j^2),
$$

(D1)

namely that the parameters are normally distributed around a mean value θ_j with variance σ_j^2. In the limit $\sigma_j \to 0$, we recover the deterministic unitary operation (2). For $\sigma_j \neq 0$,

we show that the above noise can be expressed into the form of Eq. (4). We first note that

$$\mathcal{E}_j^{\theta_j}[\hat{\rho}] = \int d\vartheta \, \frac{e^{-\frac{(\vartheta-\theta_j)^2}{2\sigma_j^2}}}{\sqrt{2\pi\sigma_j^2}} \, e^{-i\vartheta\hat{X}_j} \hat{\rho} e^{i\vartheta\hat{X}_j} \qquad \text{(D2)}$$

$$= \mathcal{D}_j \circ \mathcal{U}_j^{\theta_j}[\hat{\rho}] \equiv \mathcal{U}_j^{\theta_j} \circ \mathcal{D}_j[\hat{\rho}], \qquad \text{(D3)}$$

where $\mathcal{U}_j^{\theta_j}[\hat{\rho}] = e^{-i\theta_j\hat{X}_j}\hat{\rho}e^{i\theta_j\hat{X}_j}$ is the noiseless gate and

$$\mathcal{D}_j[\hat{\rho}] = \int d\vartheta \, \frac{e^{-\frac{\vartheta^2}{2\sigma_j^2}}}{\sqrt{2\pi\sigma_j^2}} \, e^{-i\vartheta\hat{X}_j} \hat{\rho} e^{i\vartheta\hat{X}_j}, \qquad \text{(D4)}$$

is independent on θ_j. To simplify our discussion, we assume that $\hat{X}_j^2 = \mathbb{1}$. Although a more general form can also be obtained in other cases, any tensor product of Pauli matrices satisfies the constraint $\hat{X}_j^2 = \mathbb{1}$, so we believe that this restriction covers the most common gates that can be implemented in current NISQ devices. From series expansion, it is simple to show that

$$e^{-i\vartheta\hat{X}_j}[\hat{\rho}]e^{i\vartheta\hat{X}_j} = \hat{\rho} + \sin^2(\vartheta)(\hat{X}_j\hat{\rho}\hat{X}_j - \hat{\rho}) - \frac{i}{2}\sin(2\vartheta)[\hat{X}_j, \hat{\rho}].$$
$$\text{(D5)}$$

Performing the integration in (D4), we get a dephasing-like channel, but with more general operators $\hat{X}_j$

$$\mathcal{D}_j[\hat{\rho}] = (1 - \eta_j)\hat{\rho} + \eta_j\hat{X}_j\hat{\rho}\hat{X}_j, \qquad \text{(D6)}$$

where

$$\eta_j = \frac{1 - e^{-2\sigma_j^2}}{2}. \qquad \text{(D7)}$$

For $\sigma_j \to 0$, we see that $\eta_j \to 0$ and $\mathcal{D}_j$ reduces to the identity channel.

We have studied the effect of Gaussian fluctuations in the parameters of a QAOA as a function of the noise rate $\eta_j \equiv \eta$. We found that the two terms in the bound (5) display the same behavior as observed in Fig. 2.

APPENDIX E: NOISE MODEL AND PARAMETERS

In our numerical simulations, we have used a custom noise model built up applying a depolarizing channel and thermal relaxation errors after each gate. For each of them, three parameters have to be set: a dimensionless *gate error*, connected with the parameter of the depolarizing channel that follows every single- and two-qubit gate, and the relaxation and decoherence times. In order to have a realistic noise model, in our numerical experiments relaxation and decoherence times are chosen to be different for each qubit, and gate error varies not only among qubits but also with the specific gate it is associated to. Indeed, all these parameters are taken from the available current calibration data of the various IBM quantum processors: After obtaining them, we have constructed the noise model using the class *NoiseModel* available at the moment of writing in the *Aer* library of QISKIT [52], the quantum computing software development framework from IBM.

As we are interested in analyzing how the algorithm performance depends on noise "strength," in our numerical

TABLE I. Gate error rates.

Gate	Gate error
U1	0 (virtual gate)
U2	1×10^{-3}
U3	3×10^{-3}
CNOT	4×10^{-2}

experiments we have simulated an increasing noise strength by multiplying every gate error value originally obtained querying IBM calibration data by a factor $f = 2, 4, 10$, at the same time decreasing by the same factor f every relaxation and decoherence time value (we remark that we scale up the full collection of original parameters obtained from the calibration data, so that a realistic noise model is preserved).

An additional set of parameters appearing in the noise model are gate times, i.e., the time required to apply the desired gate, during which relaxation and decoherence phenomena take place. Even gate time depends on the specific type of gate and on the qubit they are applied to: In all our simulations, we left them unchanged, using the original values provided by IBM calibration data sets.

In Tables I and II, we report as a reference the average values for the IBMQ-16-Melbourne processor at the time of writing and running our simulations. The reader may refer to the QISKIT documentation [52] for the explanation of the noise model and how it affects the different elementary gates.

APPENDIX F: ANALYTIC GRADIENT EVALUATION

We use a classical gradient descent algorithm to minimize the cost function (1). The information about the gradient of the cost function can be directly extracted by measuring the corresponding quantum observables: This procedure is often referred to as *analytically evaluated gradient*, in the sense that we can analytically define a quantum circuit to estimate the gradient value for fixed parameters. Here we focus on the Hadamard test [14].

To find the analytic gradient, we first denote with $\hat{U}_{m:n}$, with $m, n \in \{1, \ldots, P\}$, $m \leqslant n$, the unitary operator that applies the gates from the mth to the nth steps in the variational ansatz (2):

$$\hat{U}_{m:n} = e^{-i\theta_m\hat{X}_m} \ldots e^{-i\theta_n\hat{X}_n}. \qquad \text{(F1)}$$

By deriving the cost function with respect to the kth parameter, $\theta_k, k \in \{1, \ldots, P\}$, we obtain

$$\frac{\partial C(\boldsymbol{\theta})}{\partial\theta_k} = \langle\psi_0|\,\partial_{\theta_k}\hat{U}_{1:P}^{\dagger}\hat{H}\hat{U}_{1:P}\,|\psi_0\rangle + \langle\psi_0|\,\hat{U}_{1:P}^{\dagger}\hat{H}\partial_{\theta_k}\hat{U}_{1:P}\,|\psi_0\rangle$$

$$= -2\text{Im}\,\langle\psi_0|\,\hat{U}_{1:k}^{\dagger}\hat{X}_k\hat{U}_{k+1:P}^{\dagger}\hat{H}\hat{U}_{1:P}\,|\psi_0\rangle\,, \qquad \text{(F2)}$$

TABLE II. Parameters of the noise model.

Parameters	Value
Relaxation time	$55\,\mu\text{s}$
Decoherence time	$68\,\mu\text{s}$
Single-qubit gate time	$0.08\,\mu\text{s}$
Two-qubit gate time	$0.7\,\mu\text{s}$

where the last equality is obtained by using the definition (F1). Using the Pauli decompositions for both $\hat{X}_k$ and $\hat{H}$,

$$\begin{cases} \hat{X}_k = \sum_\mu \beta_\mu^{(k)} \hat{Q}_\mu^{(k)} \\ \hat{H} = \sum_\nu \alpha_\nu \hat{P}_\nu \, , \end{cases} \quad \text{(F3)}$$

where $\hat{Q}_\mu^{(k)}$ and $\hat{P}_\nu$ are Pauli operators, we can write the above derivative as

$$\frac{\partial C(\boldsymbol{\theta})}{\partial \theta_k} = -2 \sum_\mu \sum_\nu \beta_\mu^{(k)} \alpha_\nu \text{Im} \, \langle \psi_0 | \hat{U}_{1:k}^\dagger \hat{Q}_\mu^{(k)} \hat{U}_{k+1:P}^\dagger$$

$$\times \hat{P}_\nu \hat{U}_{1:P} | \psi_0 \rangle \, . \quad \text{(F4)}$$

Every term in the sum (F4) can be evaluated with a generalized Hadamard test, that requires an additional ancilla qubit. The Hadamard test is performed with the following steps:

(1) Initialize the ancilla qubit in the state $|+\rangle_A$ and the *principal register* in the state $|\psi_0\rangle$.

(2) Apply $\hat{U}_{1:k}$ to the principal register.

(3) Apply $\hat{Q}_\mu^{(k)}$ to the principal register, controlled by the ancilla.

(4) Apply $\hat{U}_{k+1:P}$ to the principal register.

(5) Apply $\hat{P}_\nu$ to the principal register, controlled by the ancilla.

(6) Apply a $\pi/2$ rotation around the x axis to the ancilla and measure the latter on the computational basis.

The probability of getting the outcome 0 after the above steps is proportional to $\text{Im} \, \langle \psi_0 | \hat{U}_{1:k}^\dagger \hat{Q}_\mu^{(k)} \hat{U}_{k+1:P}^\dagger \hat{P}_\nu \hat{U}_{1:P} | \psi_0 \rangle$. Repeating the Hadamard test for all l and all j, and performing the sum expressed in (F4), we can obtain an estimation of the analytic expression (F4) of the kth derivative, so repeating all these steps for $k \in \{1 \dots P\}$ we can evaluate the gradient.

---

[1] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell *et al.*, Nature (London) **574**, 505 (2019).

[2] J. Preskill, Quantum **2**, 79 (2018).

[3] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'brien, Nat. Commun. **5**, 4213 (2014).

[4] E. Farhi, J. Goldstone, and S. Gutmann, arXiv:1411.4028.

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

[6] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini, Quantum Sci. Technol. **4**, 043001 (2020).

[7] R. LaRose, A. Tikku, É. O'Neel-Judy, L. Cincio, and P. J. Coles, npj Quantum Inf. **5**, 8 (2019).

[8] S. Khatri, R. LaRose, A. Poremba, L. Cincio, A. T. Sornborger, and P. J. Coles, Quantum **3**, 140 (2019).

[9] X. Yuan, S. Endo, Q. Zhao, Y. Li, and S. C. Benjamin, Quantum **3**, 191 (2019).

[10] M. Schuld, A. Bocharov, K. Svore, and N. Wiebe, Phys. Rev. A **101**, 032308 (2020).

[11] Z. Wang, S. Hadfield, Z. Jiang, and E. G. Rieffel, Phys. Rev. A **97**, 022304 (2018).

[12] G. B. Mbeng, R. Fazio, and G. Santoro, arXiv:1906.08948.

[13] K. Sharma, S. Khatri, M. Cerezo, and P. J. Coles, New J. Phys. **22**, 043006 (2020).

[14] A. Harrow and J. Napp, arXiv:1901.05374.

[15] R. Sweke, F. Wilde, J. Meyer, M. Schuld, P. K. Fährmann, B. Meynard-Piganeau, and J. Eisert, Quantum **4**, 314 (2020).

[16] S. L. Braunstein and C. M. Caves, Phys. Rev. Lett. **72**, 3439 (1994).

[17] M. G. Paris, Int. J. Quantum. Inform. **7**, 125 (2009).

[18] V. Giovannetti, S. Lloyd, and L. Maccone, Nat. Photon. **5**, 222 (2011).

[19] A. Lucas, Front. Phys. **2**, 5 (2014).

[20] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Nat. Commun. **9**, 4812 (2018).

[21] N. Khaneja, T. Reiss, C. Kehlet, T. Schulte-Herbrüggen, and S. J. Glaser, J. Magn. Reson. **172**, 296 (2005).

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

[23] K. Beer, D. Bondarenko, T. Farrelly, T. J. Osborne, R. Salzmann, and R. Wolf, Nat. Commun. **11**, 808 (2020).

[24] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, Phys. Rev. A **99**, 032331 (2019).

[25] S. Bubeck, Found. Trends Machine Learn. **8**, 231 (2015).

[26] D. P. Kingma and J. Ba, arXiv:1412.6980.

[27] D. Ruppert, Efficient estimations from a slowly convergent Robbins-Monro process, Tech. Rep., Cornell University Operations Research and Industrial Engineering, Ithaca, NY, 1988 (unpublished).

[28] B. T. Polyak and A. B. Juditsky, SIAM J. Control Optim. **30**, 838 (1992).

[29] L. Banchi, N. Pancotti, and S. Bose, npj Quantum Inf. **2**, 16019 (2016).

[30] L. Innocenti, L. Banchi, A. Ferraro, S. Bose, and M. Paternostro, New J. Phys. **22**, 065001 (2019).

[31] N. Yoshioka, Y. O. Nakagawa, K. Mitarai, and K. Fujii, arXiv:1908.09836.

[32] A. Carollo, B. Spagnolo, A. A. Dubkov, and D. Valenti, J. Stat. Mech.: Theory Exp. (2019) 094010.

[33] L. Banchi, P. Giorda, and P. Zanardi, Phys. Rev. E **89**, 022102 (2014).

[34] S. Pirandola, R. Laurenza, C. Ottaviani, and L. Banchi, Nat. Commun. **8**, 15043 (2017).

[35] S. Pirandola, S. L. Braunstein, R. Laurenza, C. Ottaviani, T. P. Cope, G. Spedalieri, and L. Banchi, Quantum Sci. Technol. **3**, 035009 (2018).

[36] A. Y. Kitaev, Usp. Mat. Nauk **52**, 53 (1997).

[37] J. Watrous, *The Theory of Quantum Information* (Cambridge University Press, Cambridge, UK, 2018).

[38] J. C. Spall, IEEE Trans. Automatic Control **37**, 332 (1992).

[39] I. Bengtsson and K. Życzkowski, *Geometry of Quantum States: An Introduction to Quantum Entanglement* (Cambridge University Press, Cambridge, UK, 2017).

[40] M. P. Deisenroth, G. Neumann, and J. Peters, Found. Trends Robotics **2**, 1 (2013).

[41] P. W. Glynn, Commun. ACM **33**, 75 (1990).

[42] R. J. Williams, Mach. Learn. **8**, 229 (1992).

[43] P. Zanardi, P. Giorda, and M. Cozzini, Phys. Rev. Lett. **99**, 100603 (2007).

[44] L. Campos Venuti and P. Zanardi, Phys. Rev. Lett. **99**, 095701 (2007).

[45] B. Escher, R. de Matos Filho, and L. Davidovich, Nat. Phys. **7**, 406 (2011).

[46] C. Xue, Z.-Y. Chen, Y.-C. Wu, and G.-P. Guo, arXiv:1909.02196.

[47] M. Alam, A. Ash-Saki, and S. Ghosh, arXiv:1907.09631.

[48] S. Lloyd, arXiv:1812.11075.

[49] M. E. S. Morales, J. D. Biamonte, and Z. Zimborás, Quantum Inf Process **19**, 291 (2020).

[50] R. Barends, A. Shabani, L. Lamata, J. Kelly, A. Mezzacapo, U. Las Heras, R. Babbush, A. G. Fowler, B. Campbell, Y. Chen *et al.*, Nature (London) **534**, 222 (2016).

[51] E. Lieb, T. Schultz, and D. Mattis, Ann. Phys. **16**, 407 (1961).

[52] H. Abraham *et al.*, QISKIT: An open-source framework for quantum computing, doi: 10.5281/zenodo.2562110.

[53] S.-I. Amari, Neural Comput. **10**, 251 (1998).

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

[55] C. A. Fuchs and J. Van De Graaf, IEEE Trans. Inf. Theory **45**, 1216 (1999).