

This is a repository copy of *Noise-resilient variational hybrid quantum-classical optimization*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/166636/>

Version: Accepted Version

Article:

Gentini, Laura, Cuccoli, Alessandro, Pirandola, Stefano orcid.org/0000-0001-6165-5615 et al. (2 more authors) (2020) Noise-resilient variational hybrid quantum-classical optimization. *Physical Review A*. 052414. ISSN: 1094-1622

<https://doi.org/10.1103/PhysRevA.102.052414>

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Noise-Assisted 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, UK*

⁴*ISC-CNR, UOS Dipartimento di Fisica, Università di Firenze, I-50019, Sesto Fiorentino (FI), Italy*

(Dated: December 17, 2019)

Variational hybrid quantum-classical optimization represents one the most promising avenue to show the advantage of nowadays 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 much 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 find the unexpected result that, in some regimes, noise can be beneficial, allowing a faster solution to the optimization problem.

Introduction:— Quantum computers are nowadays available as physical devices that are expected to perform calculations essentially impossible for our best classical supercomputers [1]. However, the *quantum advantage* has been proven only for a specifically designed problem whose practical application is currently unknown. In fact, the devices currently being built are noisy intermediate-scale quantum devices (NISQ) [2], for which many of the most promising uses can be formulated as hybrid optimizations using parametric quantum circuits [3–9]. These optimizations can solve useful problems, and potentially show quantum advantage, by using the quantum device to manipulate objects that live in a space whose dimension grows exponentially with the number of qubits. The manipulation is done via gates that depend on parameters which are iteratively updated via a feedback strategy: measurement outcomes of the device are classically processed to propose better parameters in the spirit of a variational approach.

Different authors, see for instance Refs. [5, 10, 11], 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. Meanwhile, the role of stochasticity of outcomes from quantum measurements has been described using the stochastic gradient descent framework [12, 13]. However, how to tame the combined effect of noise and stochasticity in hybrid variational optimization is still far from being understood.

Here 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 precision. We focus on the effects both of noisy gates and of stochastic measurement outcomes, not matter whether optimal observables are chosen to properly extract information from the noisy process, or not. We find that the attainable precision for fixed number of iterations is bounded by a quantity that depends on the Quantum Fisher Information [14–16]. Our analysis of such bound shows that, in some circumstances, noise can speed up the solution in the sense that it can pro-

vide better approximations for fixed number of iterations. The meaning of our theoretical prediction is corroborated by numerical experiments.

Variational Hybrid Optimization:— We consider the minimization of the cost function

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

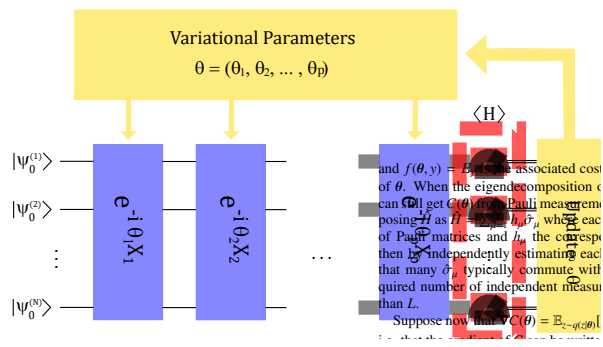
where $|\psi(\theta)\rangle$ is a variational quantum state of N qubits, namely a state that depends on P classical parameters $\theta = (\theta_1, \dots, \theta_P) \in \mathbb{R}^P$, and \hat{H} is a *cost operator* that depends on the problem. In the variational quantum eigensolver [3], for instance, \hat{H} is the Hamiltonian of a quantum many-body system and the task is to find a good variational approximation of the ground state; in the quantum approximate optimization algorithm (QAOA) [4] the task is to solve some combinatorial optimization problem and \hat{H} is an Ising-like Hamiltonian whose ground state contains the solution to the problem [17]; in quantum control [18], it is $\hat{H} = \hat{U} |\psi_0\rangle\langle\psi_0| \hat{U}^\dagger$ where \hat{U} is a target unitary, $|\psi_0\rangle$ is a reference state, and $C(\theta)$ is the fidelity of state preparation; finally, it is also possible to express in this language some machine learning applications, such as quantum classifiers [5, 19].

One of the most popular choices for the variational ansatz $|\psi(\theta)\rangle$ in (1) is the output of a *parametric quantum circuit*

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

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 nowadays NISQ devices [2] as long as \hat{X}_j contains 1- and 2-local interactions only. The fixed reference 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 by using a quantum device to prepare the variational state (2) and estimate the cost (1), and possibly its derivatives $\partial_{\theta_j} C$, via quantum measurements



show a theoretical experiment in the IBM's QASM Simulator [27] where our predictions are confirmed.

Noise-Assisted Variational Optimization:– Due to the unavoidable errors in their operation, NISQ devices cannot exactly prepare the ideal variational state (2), which must hence be substituted with $\hat{\rho}(\theta) = \mathcal{E}(\theta)[\hat{\rho}_0]$, where $\hat{\rho}_0$ is the noisy version of $|\psi_0\rangle$ and $\mathcal{E}(\theta)$ the noisy dynamical map. Although most of our results hold for more complex noise models, for the sake of simplicity in the following we use the decomposition

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

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 j -th 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 $\hat{\rho}(\theta)$ is a mixed state, the minimization of the cost function $C_{\text{noisy}}(\theta) := \text{Tr}[\hat{\rho}(\theta)\hat{H}]$ only provides an approximation to the minimum $C(\theta^{\text{opt}})$ that can be obtained in the noiseless case. The convergence rate of stochastic optimization towards the noisy minimum $C_{\text{noisy}}(\theta^{\text{opt}})$, with optimal parameters θ^{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 $C(\theta^{\text{opt}})$ and $C_{\text{noisy}}(\theta^{\text{opt}})$ we may write

$$C_{\text{noisy}}(\theta^{[1:I]}) - C(\theta^{\text{opt}}) \leq \text{Err}(\theta^{\text{opt}}, \theta^{\text{opt}}) + R \frac{G^{\text{noisy}}}{\sqrt{I}}, \quad (5)$$

where

$$\text{Err}(\theta, \theta) := C_{\text{noisy}}(\theta) - C(\theta). \quad (6)$$

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 cases, 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 $C_{\min} = C(\theta^{\text{opt}})$, and consequently ensures $\text{Err}(\theta^{\text{opt}}, \theta^{\text{opt}}) \geq 0$. This is typically not the case, as variational ansatz are normally chosen as simple circuits that are easy to implement in a NISQ device, for which one might get a negative $\text{Err}(\theta^{\text{opt}}, \theta^{\text{opt}})$. The worst-case error coming from the first term in the r.h.s. of (5) can be bounded by adapting the “peeling” technique from [28, 29]. Indeed, we show in the supplementary material that $\text{Err}(\theta, \theta) \leq P \|\hat{H}\|_{\infty} \max_k \|\mathcal{E}_k^{\theta_k} - \mathcal{U}_k^{\theta_k}\|_{\diamond}$, so the error increases at most linearly with the depth P and depends on the maximum distance, as measured by the diamond norm [30, 31], between the ideal gates and their noisy implementations. An alternative inequality $\text{Err}(\theta, \theta) \leq 2 \|\hat{H}\|_{\infty} \sqrt{1 - \langle \psi(\theta) | \hat{\rho}(\theta) | \psi(\theta) \rangle}$ 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 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 $\nabla_j C = \text{Tr}[\hat{\rho} \hat{g}_j]$ 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 [6, 12, 19, 20], either based on the Hadamard test or the so-called parameter-shift rule. However, those estimators may result biased if noisy gates only 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 SPSA [32] 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 [15, 33]

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

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

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

for any λ_j . The freedom in choosing λ_j follows from (7), since $\text{Tr}[\hat{L}_j \hat{\rho}] = \text{Tr}[\nabla_j \hat{\rho}] = \nabla_j \text{Tr}[\hat{\rho}] = 0$, implying the expectation value $\nabla_j C = \text{Tr}[\hat{g}_j \hat{\rho}]$ 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 [34]. The optimal λ_j s are discussed in the supplementary material. The measurement of the gradient operators provides stochastic outcomes $g_j^{\text{SLD}}(\theta, \gamma)$ with probabilities $\langle g_{\gamma,j} | \hat{\rho} | g_{\gamma,j} \rangle$, where we used the eigendecomposition $\hat{g}_j = \sum_{\gamma} g_{\gamma,j}^{\text{SLD}}(\theta, \gamma) |g_{\gamma,j}\rangle \langle g_{\gamma,j}|$. For pure states, the SLD operator has a simple form $\hat{L}_j = |\psi(\theta)\rangle \langle \nabla_j \psi(\theta)|$ and the above estimation strategy becomes equivalent to others already proposed in the literature [6, 12, 19, 20], which can be explicitly measured using a generalization of the Hadamard test [12].

An alternative estimator can be obtained using the logarithmic derivative (LD) trick [35], also called “reinforce” in the machine learning literature [36], which consists in writing the gradient of the cost function $\nabla_j C = \sum_{\gamma} E_{\gamma} \nabla_j p(\gamma|\theta)$ as an expectation value of $g_j^{\text{LD}}(\theta, \gamma) = E_{\gamma} \nabla_j \log p(\gamma|\theta)$ over the original distribution $p(\gamma|\theta) = \langle \gamma | \hat{\rho}(\theta) | \gamma \rangle$ where $\hat{H} = \sum_{\gamma} E_{\gamma} |\gamma\rangle \langle \gamma|$.

In the supplementary material, we show that all different estimators for the gradient satisfy the upper bound

$$G^{\text{noisy}} \leq \sqrt{P} G_{\infty}^{\text{noisy}} \leq \sqrt{P} \|\hat{H}\|_{\infty} \max_{j,\theta} \sqrt{\text{QFI}_j(\theta)}, \quad (9)$$

where QFI is the Quantum Fisher Information

$$\text{QFI}_j(\theta) = \text{Tr}[\hat{\rho}(\theta) \hat{L}_j(\theta)^2], \quad (10)$$

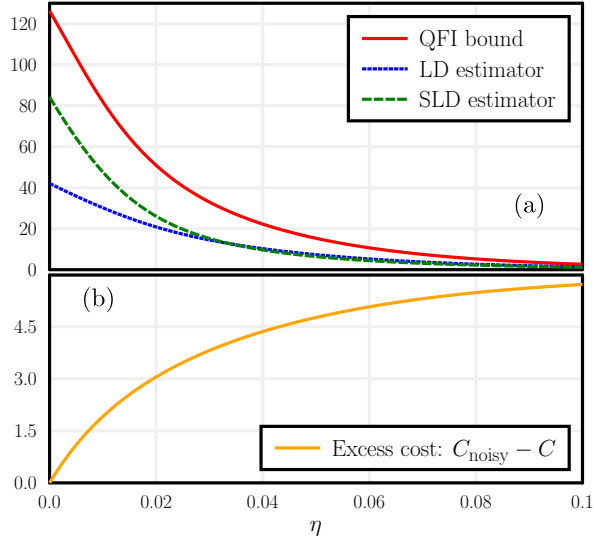


FIG. 2. Two sources of error, (a) the square root of second statistical moment of the gradient estimator $\max_{\theta} \sqrt{\mathbb{E}[\|\mathbf{g}(\theta)\|_2^2]} \leq G_{\text{noisy}}$ and (b) the excess cost $\text{Err}(\theta^{\text{opt}}, \theta^{\text{opt}})$, as a function of the depolarising noise strength 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 trick (LD) 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).

a central quantity in quantum metrology [15] that is also relevant for studying quantum phase transitions [37–39]. The bound (9), based on the QFI, shows a very important aspect: while the first term in the r.h.s. 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 $\mathcal{O}(N^2)$ (Heisenberg limit) to $\mathcal{O}(N)$ (standard quantum limit) [16, 40].

Our analysis thus shows that the convergence accuracy, as defined by the l.h.s. of (5), is bounded by the sum of two terms that typically display opposite behaviours as a function of the noise strength, with 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. Therefore, depending on the values of the constant R and on the number of iterations I , we may observe that noise does actually help. This will be shown with explicit simulations on the IBM QASM Simulator which effectively models the noisy evolution observed in the IBM-Q processors. Our analysis also shows that when I is very large, i.e. $I \gg \sqrt{PR^2\text{QFI}}$, then noise is always detrimental, as observed in some numerical experiments [41, 42]. In fact, noise-assisted optimization can only be observed for relatively few iterations, i.e. for a small number of queries of the quantum device, which is indeed the regime of interest for most variational problems on NISQ hardware.

Explicit example:— 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 non-commuting Hamiltonians, typically called \hat{H}_{γ} and \hat{H}_{β} . Here $\hat{H}_{\gamma} \equiv \hat{H}$ is equal to the cost operator appearing in Eq. (1) and is a function of the Pauli $\hat{\sigma}_j^z$ operators, where the indices $j = 1, \dots, N$ refer to the different qubits. In the *computational basis* defined by the eigenstates $\{|0\rangle, |1\rangle\}$ of $\hat{\sigma}_j^z$, \hat{H} is diagonal. The other Hamiltonian is fixed as $\hat{H}_{\beta} = -\sum_j \hat{\sigma}_j^x$, where $\hat{\sigma}_j^x$ 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}_{γ} and \hat{H}_{β}

$$|\psi(\gamma, \beta)\rangle = e^{-i\beta_P \hat{H}_{\beta}} e^{-i\gamma_P \hat{H}_{\gamma}} \dots e^{-i\beta_1 \hat{H}_{\beta}} e^{-i\gamma_1 \hat{H}_{\gamma}} |+\rangle^{\otimes N}. \quad (11)$$

The parameters are then split as $\theta = (\gamma, \beta)$ and the total depth of the circuit is $2P$. The initial state $|\psi_0\rangle = |+\rangle^{\otimes N}$, where $|+\rangle = (|0\rangle + |1\rangle)/\sqrt{2}$, is the ground state of \hat{H}_{β} . QAOA is a universal model for quantum computation [43, 44], meaning that, with specific choices of \hat{H}_{γ} , any state can be arbitrarily well approximated by $|\psi(\gamma, \beta)\rangle$ with suitable parameters γ_j, β_j and $P \rightarrow \infty$. For the specific choice $\gamma_j \propto j/P$ and $\beta_j \propto (1 - j/P)$, Eq. (11) can be interpreted as a discretization of an adiabatic evolution [4, 45] and QAOA is guaranteed to perform well for large enough P . Nonetheless, QFI can be very large when the adiabatic evolution crosses a dynamical phase transition [37–39]. Therefore, we expect that 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 [46] studied below, where \hat{H}_{β} models the global transverse field.

Here we study QAOA applied to an antiferromagnetic ring with $\hat{H}_{\gamma} = \sum_{j=1}^N \hat{\sigma}_j^z \hat{\sigma}_{j+1}^z$ and periodic boundary conditions $\hat{\sigma}_{N+1}^z \equiv \hat{\sigma}_1^z$. QAOA with this model has been studied in [10, 11], using the exact mapping to a free-fermion model. In particular, it has been proven [10] that the ground state can be exactly expressed with the QAOA ansatz (11) as long as $P \geq N/2$. 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 $\mathcal{E}_j^{\theta_j}[\hat{\rho}] = \mathcal{D}[e^{-i\theta_j \hat{X}_j} \hat{\rho} e^{i\theta_j \hat{X}_j}]$, $\mathcal{D} = \bigotimes_{j=1}^N \mathcal{D}_j$ and $\mathcal{D}_j(\rho) = (1 - \eta)\hat{\rho} + \eta\hat{\sigma}_j^x \hat{\rho} \hat{\sigma}_j^x$. 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 the Supplementary Material, we also study a different noise model, where the NISQ computer implements noisy yet unitary gates $e^{-i(\theta_j + \eta\epsilon_j)\hat{X}_j}$ 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 behaviour shown in Fig. 2.

We test our theoretical predictions using the QASM kit [27] that simulates QAOA on a physical hardware. In these simulations, the error model consists of single- and two-qubit gate errors, i.e. depolarizing error followed by a thermal relaxation error, and lastly single-qubit-readout errors. Furthermore, the gradient estimator is obtained using the SPSA algorithm [47]. In spite of the more complex model, the numerical results shown in Fig. 3 agree with our theoretical predictions. In Fig. 3 we show the probability of sampling from the different bit strings in the ideal and noisy case, for $P < N/2$ and $P > N/2$. We observe that the exact ground state, which cor-

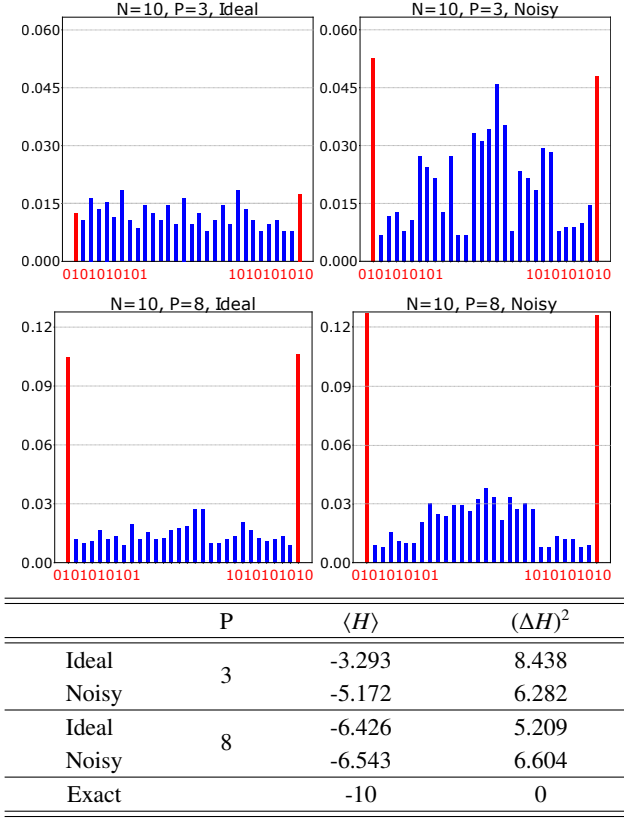


FIG. 3. Histograms of the probabilities of sampling the different bit strings after an optimization with 1024 samples. We use $N = 10$ and two values of P , $P = 3$ in the upper row and $P = 8$ in the second row. The first and second columns correspond, respectively, to the noiseless and noisy case. The bit strings are ordered via the Hamming distance from 0101..., and the two degenerate ground state configurations are shown in red. For visual clarity, the configurations with small probabilities are not shown. The total energy and energy variance of the different configurations is shown in the table.

responds to an equal superposition of the two antiferromagnetic configurations, is not even obtained for $P > N/2$, where the exact ground state is in principle achievable. This is due to the finite number of samples, and mere access to stochastic observations. Remarkably, the noise has a positive effect for both $P < N/2$ and $P > N/2$, enhancing the probability of sampling from the correct solutions at the end of the optimization. The enhancement is less pronounced for $P > N/2$, as the first term in the r.h.s. of (5) is larger for larger P and can only be positive in this specific example since $C(\theta_{\text{opt}}) = C_{\text{min}}$ for $P \geq N/2$.

Discussion:— Let us first comment upon the way QFI enters our results. 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). We also notice that in estimation theory one aims at a larger FI for a better determination of the wanted parameter via the sampling of a function that depends on it, and this is because a larger FI follows from larger local values of the

derivatives, and hence a higher sensitivity of the overall estimation procedure. Quite interestingly, though, in the scheme to which we are referring the role played by the parameter and the sampled function are reversed: we input different values of θ aiming at exploring the $C(\theta)$ -landscape, possibly locating its minimum. In fact, 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.

Getting into detail, we 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 [48]; this has been recently applied to noiseless parametric quantum circuits [49] based on the fact that, when $C = -\log p(x, \theta)$, the natural gradient is Fisher efficient, i.e. such that the variance of the estimator $\theta^{[1:l]}$ asymptotically 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.

Summarizing, we have shown that variational hybrid quantum-classical optimization algorithms provide results whose difference w.r.t. the exact ones 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. Due to the competition between these two terms, once the precision of the final result is chosen, the time the algorithm needs in order to get to its goal can be shorter in a noisy setting. In conclusion, we have theoretically found and numerically confirmed that there exist operational regimes where noise can be beneficial to speedup convergence, a result that we believe can inspire the development of new hybrid algorithms that fully take advantage of quantum effects.

ACKNOWLEDGMENTS

Acknowledgements:— L.B. acknowledges support by the program “Rita Levi Montalcini” for young researchers. S.P. acknowledges support by the QUARTET project funded by the European Unions Horizon 2020 (Grant agreement No 862644). This work is done in the framework of the Convenzione operativa between the Institute for Complex Systems of

-
- [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* **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, *Nature communications* **5**, 4213 (2014).
- [4] E. Farhi, J. Goldstone, and S. Gutmann, arXiv preprint arXiv:1411.4028 (2014).
- [5] M. Schuld, A. Bocharov, K. Svore, and N. Wiebe, arXiv preprint arXiv:1804.00633 (2018).
- [6] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, *Physical Review A* **98**, 032309 (2018).
- [7] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini, *Quantum Science and Technology* (2019).
- [8] R. LaRose, A. Tikku, É. O'Neil-Judy, L. Cincio, and P. J. Coles, *npj Quantum Information* **5**, 8 (2019).
- [9] S. Khatri, R. LaRose, A. Poremba, L. Cincio, A. T. Sornborger, and P. J. Coles, *Quantum* **3**, 140 (2019).
- [10] Z. Wang, S. Hadfield, Z. Jiang, and E. G. Rieffel, *Physical Review A* **97**, 022304 (2018).
- [11] G. B. Mbeng, R. Fazio, and G. Santoro, arXiv preprint arXiv:1906.08948 (2019).
- [12] A. Harrow and J. Napp, arXiv preprint arXiv:1901.05374 (2019).
- [13] R. Sweke, F. Wilde, J. Meyer, M. Schuld, P. K. Fährmann, B. Meynard-Piganeau, and J. Eisert, arXiv preprint arXiv:1910.01155 (2019).
- [14] S. L. Braunstein and C. M. Caves, *Physical Review Letters* **72**, 3439 (1994).
- [15] M. G. Paris, *International Journal of Quantum Information* **7**, 125 (2009).
- [16] V. Giovannetti, S. Lloyd, and L. Maccone, *Nature photonics* **5**, 222 (2011).
- [17] A. Lucas, *Frontiers in Physics* **2**, 5 (2014).
- [18] N. Khaneja, T. Reiss, C. Kehlet, T. Schulte-Herbrüggen, and S. J. Glaser, *Journal of magnetic resonance* **172**, 296 (2005).
- [19] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, *Nature communications* **9**, 4812 (2018).
- [20] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, *Physical Review A* **99**, 032331 (2019).
- [21] S. Bubeck *et al.*, *Foundations and Trends® in Machine Learning* **8**, 231 (2015).
- [22] D. P. Kingma and J. Ba, arXiv preprint arXiv:1412.6980 (2014).
- [23] D. Ruppert, *Efficient estimations from a slowly convergent Robbins-Monro process*, Tech. Rep. (Cornell University Operations Research and Industrial Engineering, 1988).
- [24] B. T. Polyak and A. B. Juditsky, *SIAM Journal on Control and Optimization* **30**, 838 (1992).
- [25] L. Banchi, N. Pancotti, and S. Bose, *NPJ Quantum Information* **2**, 16019 (2016).
- [26] L. Innocenti, L. Banchi, A. Ferraro, S. Bose, and M. Paternostro, in *Quantum Information and Measurement* (Optical Society of America, 2019) pp. F5A–28.
- [27] H. Abraham *et al.*, “Qiskit: An open-source framework for quantum computing,” (2019).
- [28] S. Pirandola, R. Laurenza, C. Ottaviani, and L. Banchi, *Nature communications* **8**, 15043 (2017).
- [29] S. Pirandola, S. L. Braunstein, R. Laurenza, C. Ottaviani, T. P. Cope, G. Spedalieri, and L. Banchi, *Quantum Science and Technology* **3**, 035009 (2018).
- [30] A. Y. Kitaev, *Uspekhi Matematicheskikh Nauk* **52**, 53 (1997).
- [31] J. Watrous, *The theory of quantum information* (Cambridge University Press, 2018).
- [32] J. C. Spall *et al.*, *IEEE transactions on automatic control* **37**, 332 (1992).
- [33] I. Bengtsson and K. Życzkowski, *Geometry of quantum states: an introduction to quantum entanglement* (Cambridge university press, 2017).
- [34] M. P. Deisenroth, G. Neumann, J. Peters, *et al.*, *Foundations and Trends® in Robotics* **2**, 1 (2013).
- [35] P. W. Glynn, *Communications of the ACM* **33**, 75 (1990).
- [36] R. J. Williams, *Machine learning* **8**, 229 (1992).
- [37] P. Zanardi, P. Giorda, and M. Cozzini, *Physical review letters* **99**, 100603 (2007).
- [38] L. C. Venuti and P. Zanardi, *Physical review letters* **99**, 095701 (2007).
- [39] L. Banchi, P. Giorda, and P. Zanardi, *Physical Review E* **89**, 022102 (2014).
- [40] B. Escher, R. de Matos Filho, and L. Davidovich, *Nature Physics* **7**, 406 (2011).
- [41] C. Xue, Z.-Y. Chen, Y.-C. Wu, and G.-P. Guo, arXiv preprint arXiv:1909.02196 (2019).
- [42] M. Alam, A. Ash-Saki, and S. Ghosh, arXiv preprint arXiv:1907.09631 (2019).
- [43] S. Lloyd, arXiv preprint arXiv:1812.11075 (2018).
- [44] M. E. Morales, J. Biamonte, and Z. Zimborás, arXiv preprint arXiv:1909.03123 (2019).
- [45] 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* **534**, 222 (2016).
- [46] E. Lieb, T. Schultz, and D. Mattis, *Annals of Physics* **16**, 407 (1961).
- [47] J. C. Spall, *Johns Hopkins apl technical digest* **19**, 482 (1998).
- [48] S.-I. Amari, *Neural computation* **10**, 251 (1998).
- [49] J. Stokes, J. Izaac, N. Killoran, and G. Carleo, arXiv preprint arXiv:1909.02108 (2019).
- [50] C. A. Fuchs and J. Van De Graaf, *IEEE Transactions on Information Theory* **45**, 1216 (1999).

Appendix A: Bound on $\text{Err}(\theta, \boldsymbol{\vartheta})$

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

$$\hat{\rho}(\theta) = \mathcal{E}(\theta_1, \dots, \theta_P)[\hat{\rho}_0]. \quad (\text{A1})$$

Here we show on the other hand that when the local error model (4) is assumed, then the error $\text{Err}(\theta, \boldsymbol{\vartheta})$ 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 $\text{Err}(\theta, \boldsymbol{\vartheta})$

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

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 $\|\hat{X}\|_\diamond$ is the diamond norm for quantum channels [30, 31]. In the last line it is

$$\mathcal{U}(\theta) := \mathcal{U}_P^{\theta_P} \circ \dots \circ \mathcal{U}_1^{\theta_1}, \quad (\text{A3})$$

and

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

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

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

where \mathcal{I} is the identity channel. We can now apply the “peeling” technique from [28, 29] 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

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

Iteratively applying the above inequality one gets

$$\delta_P \leq \sum_{k=1}^P \|\mathcal{E}_k - \mathcal{U}_k\|_\diamond \leq P \max_k \|\mathcal{E}_k - \mathcal{U}_k\|_\diamond. \quad (\text{A5})$$

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

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

An alternative bound can be obtained from (A2) via the Fuchs-van de Graaf inequality [50]

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

Appendix B: Bound on G_{noisy}

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

$$\nabla_j C = \mathbb{E}_{y \sim p(y|\theta)}[E_y \nabla_j \log p(y|\theta)]. \quad (\text{B1})$$

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

$$\mathbb{E}\left[\sum_j g_j^2\right] \leq G_{\text{noisy}}^2, \quad \max_j \mathbb{E}[g_j^2] \leq G_\infty^2. \quad (\text{B2})$$

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. [15] we find

$$\mathbb{E}[g_j^2] = \sum_y E_y^2 p(y|\theta) [\nabla_j \log p(y|\theta)]^2 \quad (\text{B3})$$

$$= \sum_y E_y^2 \frac{[\nabla_j p(y|\theta)]^2}{p(y|\theta)} \quad (\text{B4})$$

$$\stackrel{(a)}{=} \sum_y E_y^2 \frac{[\text{Tr} \hat{\Pi}_y (\hat{\rho} \hat{L}_j + \hat{L}_j \hat{\rho})/2]^2}{\text{Tr}[\hat{\Pi}_y \hat{\rho}]} \quad (\text{B5})$$

$$= \sum_y E_y^2 \frac{[\text{Re} \text{Tr}(\hat{\Pi}_y \hat{\rho} \hat{L}_j)]^2}{\text{Tr}[\hat{\Pi}_y \hat{\rho}]} \quad (\text{B6})$$

$$\leq \sum_y E_y^2 \frac{|\text{Tr}(\hat{\Pi}_y \hat{\rho} \hat{L}_j)|^2}{\text{Tr}[\hat{\Pi}_y \hat{\rho}]} \quad (\text{B7})$$

$$= \sum_y E_y^2 \left| \text{Tr} \left[\frac{\sqrt{\hat{\Pi}_y} \sqrt{\hat{\rho}}}{\sqrt{\text{Tr}[\hat{\Pi}_y \hat{\rho}]}} \sqrt{\hat{\rho}} \hat{L}_j \sqrt{\hat{\Pi}_y} \right] \right|^2 \quad (\text{B8})$$

$$\stackrel{(b)}{\leq} \sum_y E_y^2 \text{Tr}(\hat{\Pi}_y \hat{L}_j \hat{\rho} \hat{L}_j) \quad (\text{B9})$$

$$= \text{Tr}(\hat{H}^2 \hat{L}_j \hat{\rho} \hat{L}_j), \quad (\text{B10})$$

where in (a) we used 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

$$\mathbb{E}[g_j^2] \leq \|\hat{H}\|_\infty^2 \|\hat{L}_j \hat{\rho} \hat{L}_j\|_1 \leq \|\hat{H}\|_\infty^2 \text{QFI}_j, \quad (\text{B11})$$

where QFI_j is the Quantum Fisher Information (10). The upper bounds (B2) then follows with

$$G = \|\hat{H}\|_\infty \sqrt{P\left(\max_j \text{QFI}_j\right)}, \quad (\text{B12})$$

$$G_\infty = \|\hat{H}\|_\infty \sqrt{\max_j \text{QFI}_j}. \quad (\text{B13})$$

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

$$\nabla_j C = \text{Tr}[\hat{H}(\hat{\rho}\hat{L}_j + \hat{L}_j\hat{\rho})/2] = \frac{1}{2} \langle \hat{H}\hat{L}_j + \hat{L}_j\hat{H} \rangle_{\hat{\rho}(\theta)} \quad (\text{B14})$$

$$\equiv \langle \text{Re}(\hat{H}\hat{L}_j) \rangle_{\hat{\rho}(\theta)}, \quad (\text{B15})$$

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

$$\mathbb{E}[g_j^2] \equiv \langle \text{Re}(\hat{H}\hat{L}_j)^2 \rangle_{\hat{\rho}(\theta)} \quad (\text{B16})$$

$$\leq \langle \text{Re}(\hat{H}\hat{L}_j)^2 + \text{Im}(\hat{H}\hat{L}_j)^2 \rangle_{\hat{\rho}(\theta)} \quad (\text{B17})$$

$$= \frac{1}{2} \text{Tr}[\hat{\rho}(\hat{L}_j\hat{H}^2\hat{L}_j + \hat{H}\hat{L}_j^2\hat{H})] \quad (\text{B18})$$

$$= \frac{1}{2} \text{Tr}[\hat{L}_j\hat{\rho}\hat{L}_j(\hat{H}^2 + \hat{L}_j^{-1}\hat{H}\hat{L}_j^2\hat{H}\hat{L}_j^{-1})], \quad (\text{B19})$$

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

$$\mathbb{E}[g_j^2] \leq \frac{1}{2} \|\hat{L}_j\hat{\rho}\hat{L}_j\|_1 (\|\hat{H}\|_\infty^2 + \|\hat{L}_j^{-1}\hat{H}\hat{L}_j\|_\infty^2) \quad (\text{B20})$$

$$\leq \|\hat{H}\|_\infty^2 \text{QFI}_j, \quad (\text{B21})$$

which is equivalent to Eq. (B11).

Appendix C: 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

$$\begin{aligned} \mathbb{E}[g_j^2] &\equiv \left\langle \left(\frac{\{\hat{H}, \hat{L}_j\}}{2} + \lambda_j \hat{L}_j \right)^2 \right\rangle_{\hat{\rho}(\theta)} \quad (\text{C1}) \\ &= \left\langle \left(\frac{\{\hat{H}, \hat{L}_j\}}{2} \right)^2 + \lambda_j \frac{\{\hat{L}_j, \{\hat{H}, \hat{L}_j\}\}}{2} + \lambda_j^2 \hat{L}_j^2 \right\rangle_{\hat{\rho}(\theta)} \\ &= \left\langle \left(\frac{\{\hat{H}, \hat{L}_j\}}{2} \right)^2 + \lambda_j \frac{\{\hat{L}_j, \{\hat{H}, \hat{L}_j\}\}}{2} \right\rangle_{\hat{\rho}(\theta)} + \lambda_j^2 \text{QFI}_j, \end{aligned}$$

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

$$\lambda_j^{\text{opt}} = - \frac{\langle \{\hat{L}_j, \{\hat{H}, \hat{L}_j\}\} \rangle_{\hat{\rho}(\theta)}}{4 \text{QFI}_j}. \quad (\text{C2})$$

We note that the bound (B11) 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 non-optimal $\lambda_j = 0$.

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

$$\theta_j \rightarrow \mathcal{N}(\theta_j, \sigma_j^2), \quad (\text{D1})$$

namely that the parameters are normally distributed around a mean value θ_j with variance σ_j^2 . In the limit $\sigma_j \rightarrow 0$ 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

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

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

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

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

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

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

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

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

where

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

For $\sigma_j \rightarrow 0$ we see that $\eta_j \rightarrow 0$ 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 behaviour as observed in Fig. 2.