

Modeling and mitigation of cross-talk effects in readout noise with applications to the Quantum Approximate Optimization Algorithm

Filip B. Maciejewski¹, Flavio Baccari², Zoltán Zimborás^{3,4,5}, and Michał Oszmaniec¹

¹Center for Theoretical Physics, Polish Academy of Sciences, Al. Lotników 32/46, 02-668 Warsaw, Poland

²Max-Planck-Institut für Quantenoptik, Hans-Kopfermann-Straße 1, 85748 Garching, Germany

³Wigner Research Centre for Physics, H-1525 Budapest, P.O.Box 49, Hungary

⁴BME-MTA Lendület Quantum Information Theory Research Group, Budapest, Hungary

⁵Mathematical Institute, Budapest University of Technology and Economics, P.O.Box 91, H-1111, Budapest, Hungary

Measurement noise is one of the main sources of errors in currently available quantum devices based on superconducting qubits. At the same time, the complexity of its characterization and mitigation often exhibits exponential scaling with the system size. In this work, we introduce a correlated measurement noise model that can be efficiently described and characterized, and which admits effective noise-mitigation on the level of marginal probability distributions. Noise mitigation can be performed up to some error for which we derive upper bounds. Characterization of the model is done efficiently using Diagonal Detector Overlapping Tomography – a generalization of the recently introduced Quantum Overlapping Tomography to the problem of reconstruction of readout noise with restricted locality. The procedure allows to characterize k -local measurement cross-talk on N -qubit device using $O(k2^k \log(N))$ circuits containing random combinations of X and identity gates. We perform experiments on 15 (23) qubits using IBM’s (Rigetti’s) devices to test both the noise model and the error-mitigation scheme, and obtain an average reduction of errors by a factor > 22 (> 5.5) compared to no mitigation. Interestingly, we find that correlations in the measurement noise do not correspond to the physical layout of the device. Furthermore, we study numerically the effects of readout noise on the performance of the Quantum Approximate Optimization Algorithm (QAOA). We observe in simulations that for numerous objective Hamiltonians, including random MAX-2-SAT instances and the Sherrington-Kirkpatrick model, the noise-mitigation improves the quality of the optimization. Finally, we provide arguments why in the course of QAOA optimization the estimates of the local energy (or cost) terms often behave like uncorrelated variables, which greatly reduces sampling complexity of the energy estimation compared to the pessimistic error analysis. We also show that similar effects are expected for Haar-random quantum states and states generated by shallow-depth random circuits.

1 Introduction

1.1 Motivation

Outstanding progress has been made in the last years on the path to development of scalable and

Filip B. Maciejewski: filip.b.maciejewski@gmail.com

Michał Oszmaniec: oszmaniec@cft.edu.pl

fully functional quantum devices. With state of the art quantum processors reaching a scale of 50-100 qubits [1], the scientific community is approaching a regime in which quantum systems cannot be modeled on modern supercomputers using known methods [2]. This situation is unarguably exciting since it opens possibilities for the first demonstrations of quantum advantage and,

potentially, useful applications [3–5]. At the same time, however, it entails a plethora of non-trivial problems related to fighting effects of experimental imperfections on the performance of quantum algorithms. As quantum devices in the near-term will be unable to implement proper error correction [5], various methods of *noise mitigation* have been recently developed [6–15]. Those methods aim at reducing the effects of errors present in quantum gates and/or in quantum measurements. In this work, we focus on the latter, i.e., noise affecting quantum detectors.

Indeed, it has been found in currently available quantum devices that the noise affecting measurements is quite significant. Specifically, errors of the order of a few percent in a single qubit measurement and non-negligible effects of cross-talk were reported [1, 13–16]. Motivated by this, a number of methods to characterize and/or reduce measurement errors were proposed [13–23]. Readout noise mitigation usually relies on the classical post-processing of experimental statistics, preceded by some procedure of noise characterization. Existing techniques typically suffer from the curse of dimensionality due to characterization cost, sampling complexity, and the complexity of post-processing – which scale exponentially with the number of qubits N . Fortunately, some interesting problems in quantum computing do not require measurements to be performed across the whole system. An important class of algorithms that have this feature are the Quantum Approximate Optimization Algorithms (QAOA) [24], which only require the estimation of a number of few-particle *marginals*. QAOA is a heuristic, hybrid quantum-classical optimization technique [24, 25], that was later generalized as a standalone ansatz [26] shown to be computationally universal [27, 28]. In its original form, QAOA aims at finding approximate solutions for hard combinatorial problems, notable examples of which are maximum satisfiability (SAT) [29], maximum cut (MAXCUT) [30], or the Sherrington-Kirkpatrick (SK) spin-glass model [31, 32]. Regardless of the underlying problem, the main QAOA subroutine is the estimation of the energy of the local classical Hamiltonians on a quantum state generated by the device. Those Hamiltonians are composed of a number of a few-body commuting operators, and hence estimation of energy can be done via

estimation of the local terms (which can be performed simultaneously). Since the estimation of marginal distributions is the task of low sampling and post-processing complexity, this suggests that error mitigation techniques can be efficiently applied in QAOA. In this work, we present a number of contributions justifying the usage of measurement error-mitigation in QAOA, even in the presence of significant cross-talk effects.

1.2 Results outline

Our first contribution is to provide an efficiently describable measurement noise model that incorporates asymmetric errors and cross-talk effects. Importantly, our noise model admits efficient error-mitigation on the marginal probability distributions, which can be used, e.g., for improvement of the performance of variational quantum algorithms. We show how to efficiently characterize the noise using a number of circuits scaling logarithmically with the number of qubits. To this aim, we generalize the techniques of the recently introduced Quantum Overlapping Tomography (QOT) [33] to the problem of readout noise reconstruction. Specifically, we introduce notion of Diagonal Detector Overlapping Tomography (DDOT) which allows to reconstruct noise description with k -local cross-talk on N -qubit device using $O(k2^k \log(N))$ quantum circuits consisting of single layer of X and identity gates. Furthermore, we explain how to use that characterization to mitigate noise on the marginal distributions and provide a bound for the accuracy of the mitigation. Importantly, assuming that cross-talk in readout noise is of bounded locality, the sampling complexity of error-mitigation is not significantly higher than that of the starting problem of marginals estimation.

We test our error-mitigation method in experiments on 15 qubits using IBM’s device and on 23 qubits using Rigetti’s device, both architectures based on superconducting transmon qubits [34]. We obtain a significant advantage by using a correlated error model for error mitigation in the task of ground state energy estimation. Interestingly, the locality structure of the reconstructed errors in these devices does not match the spatial locality of qubits in these systems.

We also study statistical errors that appear in the simultaneous estimation of multiple local Hamiltonian terms that appear frequently in

QAOAs. In particular, we provide arguments why one can expect that the estimated energies of local terms behave effectively as uncorrelated variables, for the quantum states appearing at the beginning and near the end of the QAOA algorithm. This allows to prove significant reductions in sampling complexity of the total energy estimation (compared to the worst-case upper bounds).

Finally, we present a numerical study and detailed discussion of the possible effects that measurement noise can have on the performance of the Quantum Approximate Optimization Algorithm. This includes the study of how noise distorts the quantum-classical optimization in QAOAs. We simulate the QAOA protocol on an 8-qubit system affected by correlated measurement noise inspired by the results of IBM’s device characterization. For a number of random Hamiltonians, we conclude that the error-mitigation highly improves the accuracy of energy estimates, and can help the optimizer to converge faster than when no error-mitigation is used.

1.3 Related works

The effects of simple, uncorrelated noisy quantum channels on QAOAs were analyzed in Refs. [35–37]. In particular, the symmetric bitflip noise analyzed in [35] can be used also to model symmetric, uncorrelated measurement noise - a type of readout noise which has been demonstrated to be not very accurate in currently available devices based on transmon qubits [13, 16]. A simple readout noise-mitigation technique on the level of marginals for QAOAs was recently experimentally implemented on up to 23 qubits in a work by Google AI Quantum team and collaborators [38]. Importantly, the authors assumed uncorrelated noise on each qubit – we believe that our approach which accounts for correlations could prove beneficial in those kinds of experiments. In a similar context, readout noise-mitigation using classical post-processing on global probability distributions was implemented in QAOA and Variational Quantum Eigensolver (VQE) experiments on up to 6 qubits [39, 40]. While for such small system sizes it is possible to efficiently perform global error mitigation, we emphasize that our approach based on the noise-mitigation on the marginals could be performed also in larger experiments of this type.

Alternative methods of characterization of correlated measurement noise were recently proposed in Refs. [15, 16, 20]. Out of the above-mentioned works, perhaps the most related to ours in terms of studied problems is Ref. [15], therefore we will now comment on it thoroughly. The authors introduced a correlated readout noise model based on Continuous Time Markov Processes (CTMP). They provide both error-characterization and error-mitigation methods. The CTMP noise model assumes that the noisy stochastic process in a measurement device can be described by a set of two-qubit generators with corresponding error rate parameters. The total number of parameters required to describe a generic form of such noise for N -qubit device is $2N^2$. The authors propose a method of noise characterization that requires preparation of a set of suitably chosen classical states and performing a post-selection on the noise-free outcomes (i.e., correct outcomes given known input classical state) on the subset of $(N - 2)$ qubits. Since probability of noise-free outcomes can be exponentially small even for the uncorrelated readout noise, this method can prove relatively costly in practice. We believe that our DDOT characterization technique could prove useful in reducing the number of parameters needed to describe CTMP model (by showing which pairs of qubits are correlated, and for which the crosstalk can be neglected), hence allowing for much more efficient version of noise reconstruction presented in Ref. [15]. The authors also present a novel noise-mitigation method that allows to estimate the noise-reduced expected values of observables. The method is based on decomposition of inverse noise matrix into the linear combination of stochastic matrices, and constructing a random variable (with the aid of classical randomness and post-processing) that agrees in the expected value of noise-free observable. In general, both sampling complexity and classical post-processing cost scale exponentially with the number of qubits. Since the authors aim to correct observables with arbitrary locality, the problem they consider is different to our approach that aims to correct only *local* observables (and therefore does not exhibit exponential scalings). It is an interesting problem to see whether the CTMP noise model can also be interpreted on the level of marginal probability distributions in a way that

allows for mitigation analogous to ours in terms of complexity.

2 Correlated readout noise model

2.1 Preliminaries

In quantum mechanics, the most general measurement \mathbf{M} that can be performed is represented by a Positive Operator-Valued Measure (POVM) [41]. A POVM \mathbf{M} with r outcomes on a d -dimensional Hilbert space is a set of positive-semidefinite operators that sum up to identity, i.e.,

$$\mathbf{M} = \{M_i\}_i^r, \quad \forall i \ M_i \geq 0, \quad \sum_{i=1}^r M_i = \mathbb{1}, \quad (1)$$

where $\mathbb{1}$ is the identity operator. When one performs a measurement \mathbf{M} on the quantum state ρ , the probability of obtaining outcome i is given by Born's rule: $\Pr(i|\mathbf{M}, \rho) = \text{tr}(\rho M_i)$.

In quantum computing a perfect measurement is often modeled as a so-called *projective* measurement $\mathbf{P} = \{P_i\}_i^r$ for which the measurement operators, in addition to the requirements from Eq. (1), are also projectors, i.e., $\forall_i P_i^2 = P_i$. While non-projective measurements have many applications in certain quantum information processing tasks, in this work we are interested in using them to model imperfections in measurement devices. The relationship between an ideal measurement \mathbf{P} and its noisy implementation \mathbf{M} can be modeled by the application of a generic quantum channel. Recently it has been experimentally demonstrated for superconducting quantum devices that in practice those channels, to a good approximation, belong to a restricted class of stochastic maps [13, 14], which we will refer to as 'classical measurement noise'. In such a model, the relation between \mathbf{M} and \mathbf{P} is given by some stochastic transformation Λ . Namely, we have $\mathbf{M} = \Lambda \mathbf{P}$, i.e., $M_i = \sum_j \Lambda_{ij} P_j$. Due to the linearity of Born's rule, it follows that probabilities \mathbf{p}^{noisy} from which noisy detector samples are related to the noiseless probabilities \mathbf{p}^{ideal} via the same stochastic map, hence [13]

$$\mathbf{p}^{noisy} = \Lambda \mathbf{p}^{ideal}. \quad (2)$$

Specifically, in the convention where probability vectors are columns, the noise matrix Λ is left-stochastic, meaning that each of its columns contains non-negative numbers that sum up to 1.

Such noise is thus equivalent to a stochastic process, in which an outcome from a perfect device probabilistically changes to another (possibly erroneous) one. Equation (2) suggests a simple way to mitigate errors on the noisy device – via left-multiplying the estimated statistics by the inverse of noise matrix Λ^{-1} [13, 14]. From stochasticity of Λ it follows that its inverse does preserve the sum of the elements of probability vectors, however it may introduce some unphysical (i.e., lower than 0 or higher than 1) terms in corrected vector. A common practice in such a scenario is to solve optimization problem

$$\mathbf{p} = \underset{\mathbf{q}}{\text{argmin}} \|\mathbf{q} - \Lambda^{-1} \mathbf{p}^{noisy}\|_2^2, \quad (3)$$

where minimization goes over all proper probability distributions. This introduces additional error in the final estimations which can be easily upper-bounded [13].

Before going further, we note that while multiplication by Λ^{-1} is perhaps the most natural (and simple) method to reduce the noise and has been shown to be useful in practical situations [13, 14], there exist more sophisticated techniques of noise-mitigation that do not exhibit this problem. For example, Iterative Bayesian Unfolding [18] always returns physical probability vectors. We note that our noise model and noise-mitigation on marginal probability distributions (to be introduced in following sections) is consistent with using methods different than Λ^{-1} correction, and we intend to test them in the future.

To finish this introduction, we recall that, as mentioned above, Eq. (2) does not present the most general model of quantum measurement noise. Specifically, *coherent* errors might occur, and reduce the effectiveness of error-mitigation (a detailed analysis of this effect was presented in [13]). However, it was validated experimentally on multiple occasions, that coherent errors in superconducting quantum devices are small, and the error-mitigation by classical post-processing was shown to work very well in few-qubit scenarios [13, 14].

2.2 Correlations in readout noise

The size of the matrix Λ scales exponentially with the number of qubits. Thus, if one wants to estimate such a generic Λ using standard methods, both the number of circuits and sampling complexity scale exponentially. Indeed, the standard

method of reconstructing Λ is to create all the 2^n computational basis and estimate the resulting probability distributions (which constitute the columns of Λ). We refer to such characterization as Diagonal Detector Tomography (DDT), since it probes the diagonal elements of the measurement operators describing the detector. This is restricted version of more general Quantum Detector Tomography (QDT) [42–45].

These complexity issues can be circumvented if one assumes some *locality* structure in the measurement errors. For example, in the simplest model with completely uncorrelated readout noise, the Λ matrix is a simple tensor product of single-qubit error matrices Λ_{Q_i}

$$\Lambda = \bigotimes_i \Lambda_{Q_i} \quad (\text{uncorrelated noise}). \quad (4)$$

In this model, we need to estimate only single-qubit matrices, which renders the complexity of the problem to be linear in the number of qubits. However, for contemporary quantum devices based on superconducting qubits, it was demonstrated that such a noise model is not very accurate due to the cross-talk effects [1, 13, 14, 16]. At the same time, the completely correlated noise is not realistic as well, which motivates the search for a model that can account for correlations in readout errors while still giving an efficient description of Λ .

In this work, we propose such a model and give a method to characterize it. Let us lay out the basic concepts of our model. Consider the correlated errors between some group of qubits \mathcal{C}_χ . The most general way of describing those errors is to treat the qubits in \mathcal{C}_χ as a single object, i.e., to always consider their measurement outcomes together. In terms of the noise matrix description, this means that the noise matrix on \mathcal{C}_χ is some generic $\Lambda_{\mathcal{C}_\chi}$ acting on \mathcal{C}_χ . This gives rise to the first basic object in our model – the *clusters* of qubits. The cluster \mathcal{C}_χ is a group of qubits with correlations between them so strong, that one can not consider outcomes of their measurements separately. At the same time, it is unlikely that in actual devices the correlations between all the qubits will be so strong that one should assign them all to a single cluster. This motivates the introduction of another, milder possibility. Consider a measurement performed on qubits in cluster \mathcal{C}_χ and some other qubits \mathcal{N}_χ (not being in that cluster). It is conceivable to imag-

ine some complicated physical process, which results in the situation in which the noise matrix $\Lambda_{\mathcal{C}_\chi}$ on cluster \mathcal{C}_χ slightly depends on the state of the qubits in \mathcal{N}_χ . To account for that, we introduce the second basic object of our model – the *neighborhood* of the cluster. The neighborhood \mathcal{N}_χ of a cluster \mathcal{C}_χ is a group of qubits the state of which *just before the measurement* affects slightly the noise matrix acting on the cluster \mathcal{C}_χ .

For example, if \mathcal{C}_χ contains only a single qubit, say Q_0 , it is possible that due to some effective ferromagnetic-type interaction, the probability of erroneously detecting state “0” of Q_0 as “1” rises when the neighboring qubit Q_1 is in state “1” (compared to when it is in state ‘0’).

A notion related to our “neighborhood” has appeared in recent literature. Specifically in the context of measurement error characterization, ‘spectator qubits’ are the qubits that affect measurement noise on other qubits [16, 20]. However, so far this effect was treated rather as an undesired complication, while here it is an inherent element of the proposed noise model.

Now we are ready to provide an efficient noise model. We construct a global noise matrix Λ with matrix elements of the following form

$$\Lambda_{X_1 \dots X_N | Y_1 \dots Y_N} = \prod_\chi \Lambda_{\mathbf{X}_{\mathcal{C}_\chi} | \mathbf{Y}_{\mathcal{N}_\chi}}^{\mathbf{Y}_{\mathcal{N}_\chi}}. \quad (5)$$

In the next subsection we give some illustrative examples, but first let us thoroughly describe the notation used in the above equation. A collection $\{\mathcal{C}_\chi\}_\chi$ gives us the partitioning of the set of all qubits. Explicitly, $\mathcal{C}_\chi \cap \mathcal{C}_{\chi'} = \emptyset$ if $\chi \neq \chi'$ and $\cup_\chi \mathcal{C}_\chi = [N]$, where N is the number of qubits. To each cluster \mathcal{C}_χ the model associate its *neighborhood* \mathcal{N}_χ . Equation (5) can be now understood in the following way. The noise matrix $\Lambda^{\mathbf{Y}_{\mathcal{N}_\chi}}$ describing the measurement noise occurring in cluster \mathcal{C}_χ depends on the state *just before measurement* of the qubits in the neighborhood \mathcal{N}_χ of that cluster (hence the superscript $\mathbf{Y}_{\mathcal{N}_\chi}$ denoting that state). Importantly, each $\Lambda^{\mathbf{Y}_{\mathcal{N}_\chi}}$ is left-stochastic for any state of the neighbors. By $\mathbf{X}_{\mathcal{C}_\chi}$ (or $\mathbf{Y}_{\mathcal{C}_\chi}$) we denote bit-strings of qubits belonging to cluster \mathcal{C}_χ which were measured (or put inside the device just before measurement). Finally, $\mathbf{Y}_{\mathcal{N}_\chi}$ indicates the bit-string denoting the state just before the measurement of the qubits from the neighborhood \mathcal{N}_χ of the cluster \mathcal{C}_χ (see Fig. 1a for illustration). Note that in general the

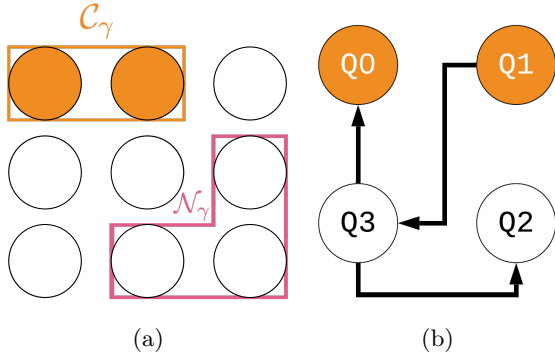


Figure 1: Exemplary correlations in the measurement noise that can be captured by our model. Each circle represents a qubit. Subfigure a) represents a 9-qubit device with some cluster \mathcal{C}_x (orange envelope) consisting of two qubits. The noise on the qubits from that cluster is dependent on the state of qubits from its neighborhood \mathcal{N}_x (magenta envelope). Note that the “neighborhood” does not have to correspond to spatial arrangement of qubits in the device. Subfigure b) shows a more detailed example of a four-qubit device. Qubits Q_0 and Q_1 are in one cluster, which is indicated by coloring them with the same color. Qubits Q_2 and Q_3 are white, meaning they do not belong to any cluster. Qubits at the beginning of the arrow are the neighbors of the qubits at the end of the arrow. Explicitly, the clusters in the example are $\mathcal{C}_1 = \{0, 1\}$, $\mathcal{C}_2 = \{2\}$, and $\mathcal{C}_3 = \{3\}$, while their neighborhoods are $\mathcal{N}_1 = \{3\}$, $\mathcal{N}_2 = \{3\}$, and $\mathcal{N}_3 = \{1\}$. Correlations in the readout errors for qubits Q_0 and Q_1 can be arbitrary, while for the rest of the qubits the dependencies are restricted by the structure of clusters and the neighbors. In particular, the noise matrix on Q_3 depends on the state of Q_1 , while the state of Q_3 affects the noise on qubits Q_0 and Q_2 . At the same time, qubit 2 does not affect the noise matrix on any other qubit. See the description in the main text.

correlations in measurement errors (expressed by the structure of the clusters and neighborhoods) do not need to be directly correlated with the physical layout of the device. In general Λ in Eq.(5) is specified by only $\sum_{\chi} (2^{|\mathcal{C}_{\chi}|}) 2^{|\mathcal{N}_{\chi}|}$ parameters, where $|\mathcal{C}_{\chi}|$ and $|\mathcal{N}_{\chi}|$ are sizes of χ 'th cluster and its neighbourhood respectively. Therefore this description is *efficient* provided sizes of the clusters and their neighborhoods are bounded by a constant.

2.3 Illustrative examples

In what follows present examples of readout correlation structures that can be described with our model. It is instructive to start with a simple

example of a hypothetical four-qubit device depicted in Fig. 1b. Note that in this example we have only one non-trivial (i.e., with size ≥ 2) cluster. Let us write explicitly the matrix elements of the global noise matrix acting on that exemplary 4-qubit device

$$\Lambda_{X_0 X_1 X_2 X_3 | Y_0 Y_1 Y_2 Y_3} = \Lambda_{X_0 X_1 | Y_0 Y_1}^{Y_3} \Lambda_{X_2 | Y_2}^{Y_3} \Lambda_{X_3 | Y_3}^{Y_1}. \quad (6)$$

Note that on the RHS of Eq. (6), the superscript Y_3 appears two times, indicating that noise matrices on the cluster \mathcal{C}_1 and on the cluster \mathcal{C}_2 both depend on the state just before measurement of the qubit 3. At the same time, there is no superscript Y_2 , which follows from the fact that qubit 2 does not affect the noise on any other qubits. Note that while a generic noise matrix on 4 qubits would require reconstruction of 16×16 matrix, here we need a number of smaller dimensional matrices to fully describe the noise.

We now move to a more general readout error model, which is particularly inspired by current superconducting qubits implementations of quantum computing devices. Consider a collection of qubits arranged on a device with a limited connectivity. This can be schematically represented as a graph $\mathcal{G}(V, E)$, where each vertex in V represents a qubit and each edge in E connects two qubits that can interact in the device. In such a scenario, a natural first step beyond an uncorrelated readout noise model can be a nearest-neighbour correlated model, where readout errors on each qubit are assumed to be influenced at most by the state of the neighbouring ones. By using the notation introduced in the previous section, we can represent such a model by associating a single-qubit cluster to each vertex, $\mathcal{C}_i = \{i\}$, for $i \in V$, and defining the neighbourhoods according to the graph structure, namely $\mathcal{N}_i = \{j | (i, j) \in E\}$. The global noise matrix then reads

$$\Lambda_{X_1 \dots X_{|V|} | Y_1 \dots Y_{|V|}} = \prod_{i \in V} \Lambda_{\mathbf{X}_i | \mathbf{Y}_i}^{\mathbf{Y}_{\mathcal{N}_i}}. \quad (7)$$

If we specialise this to the case of a 2D rectangular lattice of size L , the neighbourhood of generic (i.e. not belonging to the boundary) vertex becomes $\mathcal{N}_i = \{i + 1, i - 1, i + L, i - L\}$. It follows that each $\Lambda_{\mathbf{X}_i | \mathbf{Y}_i}^{\mathbf{Y}_{\mathcal{N}_i}}$ can be represented by a collection of $2^4 = 16$ matrices of size 2×2 , which is an

exponential improvement with respect to a general $2^{L^2} \times 2^{L^2}$.

Although the above correlated noise model seems a very natural one, we will see in the following Section that it does not encompass all the correlated readout errors in current superconducting devices, for which it will be more convenient to resort to models (5) with more general cluster and neighbourhood structures that do not necessarily correspond to the physical layout of the devices.

3 Characterization of readout noise

Here we outline a strategy to determine a noise matrix in the form (5) which closely represents the readout noise of a given device. We proceed in two steps: at first we infer the structure of clusters ($\{\mathcal{C}_\chi\}$) and neighbourhoods ($\{\mathcal{N}_\chi\}$) by making use of Diagonal Detector Tomography (DDT); then we proceed to experimentally determine noise matrices $\{\Lambda_{\mathbf{X}_{\mathcal{C}_\chi}^{\mathbf{Y}_{\mathcal{N}_\chi}}|\mathbf{Y}_{\mathcal{C}_\chi}}\}$

For the first step, we propose to reconstruct all two-qubit noise matrices (averaged over all other qubits) by means of DDT and calculate the following quantities

$$c_{j \rightarrow i} = \frac{1}{2} \|\Lambda_{Q_i}^{Y_j=0'} - \Lambda_{Q_i}^{Y_j=1'}\|_{1 \rightarrow 1} \quad (8)$$

where $\|A\|_{1 \rightarrow 1} := \sup_{\|v\|_1=1} \|Av\|_1 = \max_j \sum_i |A_{ij}|$. The above quantity has a strong operational motivation in terms of Total-Variation Distance (TVD). This distance quantifies statistical distinguishability of probability distributions \mathbf{p} and \mathbf{q} and can be defined by

$$\text{TVD}(\mathbf{p}, \mathbf{q}) = \frac{1}{2} \|\mathbf{p} - \mathbf{q}\|_1 = \frac{1}{2} \sum_i |p_i - q_i|. \quad (9)$$

We can give the following, intuitive interpretation of the quantity from Eq. (8): $c_{j \rightarrow i}$ represents the maximal TVD for which the output probability distributions on qubit Q_i differs due to the impact of the state of the qubit Q_j on the readout noise on Q_i . Note that in general $c_{j \rightarrow i} \neq c_{i \rightarrow j}$, which encapsulates the asymmetry in the correlations which is built into our noise model.

We propose to use the values of $c_{j \rightarrow i}$ to decide whether the qubits should belong to the same clusters, to the neighborhoods, or should be considered uncorrelated (a simple, intuitive algorithm for this procedure is presented in Appendix B.2., Algorithm 3 – in the future, we

intend to extend those methods). After doing so, the noise matrices $\{\Lambda_{\mathbf{X}_{\mathcal{C}_\chi}^{\mathbf{Y}_{\mathcal{N}_\chi}}|\mathbf{Y}_{\mathcal{C}_\chi}}\}$ can be reconstructed by means of joint DDT over the sets of qubits $\{\mathcal{C}_\chi \cup \mathcal{N}_\chi\}_\chi$.

In the above construction we assumed that the joint size of a cluster and its neighborhood is at most k . This makes it so that one has to gather DDT data on subsystems of fixed size, implying a number of different circuits that scales at most as $O(N^k)$. However, for any characterization procedure, it is expedient to utilize as few resources as possible. In order to reduce the number of circuits even further, in the next Section we generalize the recently introduced Quantum Overlapping Tomography (QOT) [33] (see also recent followups [46, 47]) to the context of Diagonal Detector Tomography. We will refer to our method as Diagonal Detector Overlapping Tomography (DDOT).

3.1

Diagonal Detector Overlapping Tomography

Quantum Overlapping Tomography is a technique that was introduced for a problem of efficient estimation of all k -particle marginal quantum states. The main result of Ref. [33] was to use the concept of hashing functions [48–51] to reduce the number of circuits needed to reconstruct all k -qubit marginal states. Specifically, it was shown there that $O(\log(N) e^k)$ circuits suffice for this purpose. Here we propose to use an analogous technique to estimate all noise matrices corresponding to k -particle subsets of qubits. Specifically, we propose to construct a collection of circuits consisting of certain combinations of $\mathbb{1}$ and X gates in order to initialize qubits in states $|0\rangle$ or $|1\rangle$. With fixed k , the collection of quantum circuits for DDOT must have the following property – for each subset of qubits of size k , all of the computational-basis states on that subset must appear at least once in the whole collection of circuits. Intuitively, if a collection has this property, then the implementation of all circuits in the collection allows us to perform tomographic reconstruction (via standard DDT) of noise matrices on all k -qubit subsets. One can think about DDOT as a method of parallelizing multiple local DDTs in order to minimize number of circuits needed to obtain description of all local k -qubit noise processes. In Appendix B.4 we show that

it suffices to implement $O(k2^k \log(N))$ quantum circuits consisting of random combinations of X and identity gates in order to construct a DDOT circuits collection that allows to capture all k -qubit correlations in readout errors (see Algorithm 1 and Algorithm 2). It is an exponential improvement over standard technique of performing local DDTs separately (which, as mentioned above, requires $O(N^k)$ circuits). For example, if one chooses $k = 5$ for $N = 15$ -qubit device, the naive estimation of all 5-qubit marginals would require the implementation of $2^5 \binom{15}{5} \approx 10^5$ quantum circuits, while DDOT allows doing so using ≈ 350 circuits. We note that this efficiency, however, comes with a price. Namely, since different circuits are sampled with different frequencies, some false-positive correlations might appear. This may cause some correlations in the reconstructed noise model to be overestimated (see Appendix B.6 for a detailed explanation of this effect). This effect can be mitigated either by certain post-processing of experimental results (see Appendix B.6), or by constructing DDOT collections that sample each term the same number of times. Using probabilistic arguments in Appendix B.3 we show that still the number of circuits exponential in k and logarithmic in N suffices if we want to have all k particle subsets sampled with *approximately* equal frequency.

3.2 Experimental results

We implemented the above procedure with $k = 5$ for IBM’s 15q *Melbourne* device and 23-qubit subset of Rigetti’s *Aspen-8* device (the details of the experiments are moved to Appendices). The obtained correlation models are depicted in Fig. 2. In the case of Rigetti’s device, our procedure reports a very complicated structure of multiple correlations in readout noise, while in the case of IBM’s device the correlations are fairly simple. We discuss this issue in detail in further sections while presenting results of noise-mitigation benchmarks. Here we conclude by making an observation that, despite common intuition, the structure of the correlations in the readout noise can not be directly inferred from the physical layout of the device.

4 Noise mitigation on marginals

Now we will analyze the estimation and correction of marginal probability distributions affected by the kind of correlated noise model introduced in the previous section. In the next subsection, we will use those findings to propose a benchmark of the adopted noise model.

4.1 Noise on marginal probability distributions

Let us denote by \mathbf{p}^{noisy} a global probability distribution generated by measurement of arbitrary quantum state on the noisy detector for which Eq. (5) holds. As mentioned previously, for many interesting problems, such as QAOA or VQE algorithms, one is interested not in the estimation of \mathbf{p} itself (which is an exponentially hard task), but instead in the estimation of multiple marginal probability distributions obtained from \mathbf{p} . Let us say that we are interested in the marginal on a subset \mathcal{S} formed by clusters \mathcal{C}_χ indexed by a set \mathcal{A} , $\mathcal{S} := \cup_{\chi \in \mathcal{A}} \mathcal{C}_\chi$, where each \mathcal{C}_χ is some cluster of qubits (see green envelope on Fig. 3 for illustration). Our goal is to perform error mitigation on \mathcal{S} . To achieve this, we need to understand how our model of noise affects marginal distribution on \mathcal{S} .

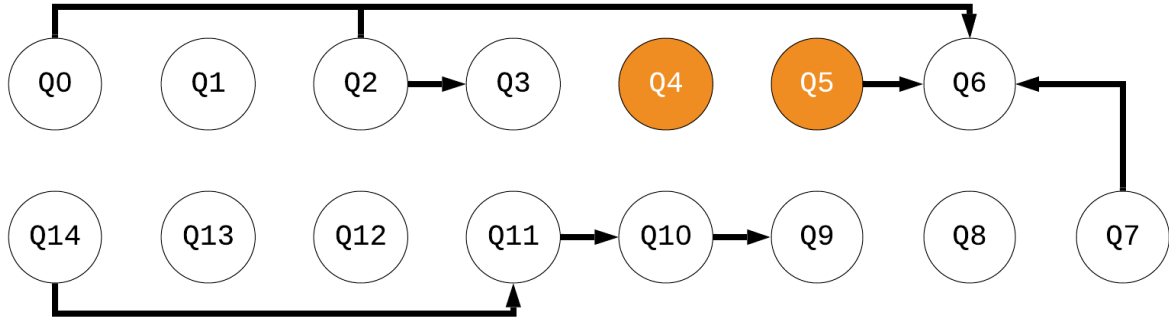
From the definition of the noise model in Eq. (5) we get that the marginal probability distribution $\mathbf{p}_{\mathcal{S}}^{noisy}$ on \mathcal{S} , is a function of the local noise matrices acting on the qubits from \mathcal{S} and the “joint neighborhood” of \mathcal{S} , $\mathcal{N}(\mathcal{S}) := \cup_{\chi \in \mathcal{A}} \mathcal{N}_\chi \setminus \mathcal{S}$. The set $\mathcal{N}(\mathcal{S})$ consists of qubits which are neighbors of points from \mathcal{S} but are not in \mathcal{S} .

Because of this one can not simply use the standard mitigation strategy: i.e., estimate $\mathbf{p}_{\mathcal{S}}^{noisy}$ and reconstruct probability distribution $\mathbf{p}_{\mathcal{S}}^{ideal}$ by applying the inverse of Λ (more discussion of this matter is given in Appendices).

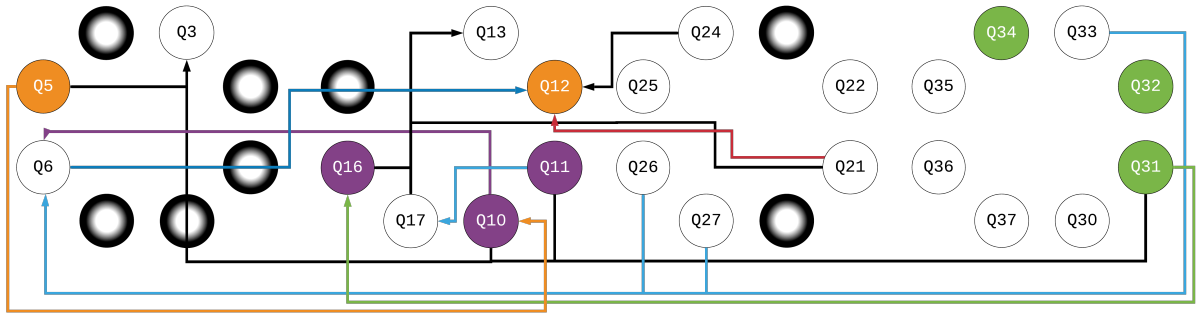
To circumvent the above problem we propose to use the following natural ansatz for the construction of an *approximate* effective noise model on the marginal \mathcal{S}

$$\Lambda_{av}^{\mathcal{S}} := \frac{1}{2^{|\mathcal{N}(\mathcal{S})|}} \sum_{\mathbf{Y}_{\mathcal{N}(\mathcal{S})}} \Lambda^{\mathbf{Y}_{\mathcal{N}(\mathcal{S})}}, \quad (10)$$

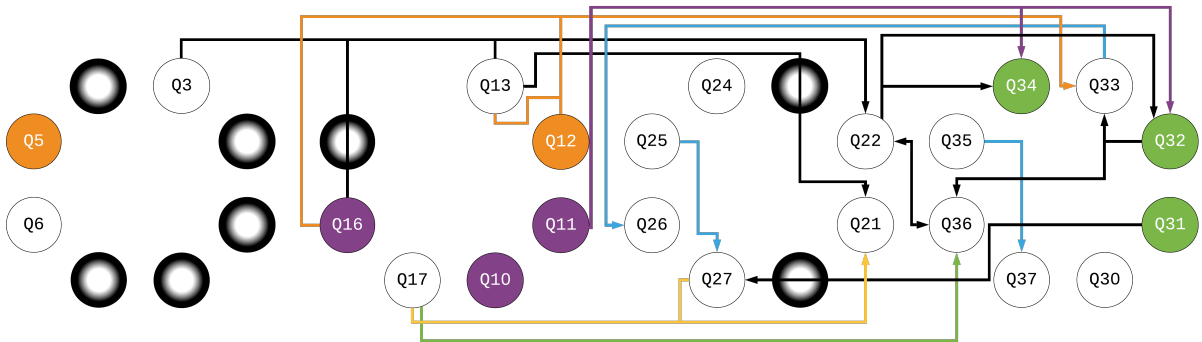
where summation is over states of qubits in the joint neighborhood $\mathcal{N}(\mathcal{S})$ defined above. In other words, it is a noise matrix averaged over all states



(a) IBM's *Melbourne* device, 15 qubits.



(b) Rigetti's *Aspen-8* device. Arrows indicate qubits which affect the measurement noise on the left half of the device.



(c) Rigetti's *Aspen-8* device. Arrows indicate qubits which affect the measurement noise on the right half of the device.

Figure 2: Depiction of the correlation model obtained with Diagonal Detector Overlapping Tomography on a) IBM's 15-qubit *Melbourne* device and b,c) a 23-qubit subset of Rigetti's *Aspen-8* device. Due to the complicated structure of Rigetti's correlations, for clarity we divided the plots into two parts which show correlations on the left and right halves of the device (the merged plot can be found in the Appendix E.2). The 8 black-and-white circles without label represent qubits which were not included in the characterization due to poor fidelity of the single-qubit gates (below 98%). The meaning of the rest of the symbols is described in the caption of Fig. 1b. The colors of the lines connecting the neighbors on b,c) are provided such that the crossings of the lines are unambiguous (and have no other meaning otherwise). For the layout of the graphs, we used the qubits actual connectivity in the devices (i.e., it is possible to physically implement two-qubit entangling gates on all nearest-neighbors in the graph). For IBM's device, we included the qubits in the cluster if the correlations given by Eq. (8) were higher than 4% in any direction, while we marked qubits as neighbors if the correlations were higher than 1%. In the case of Rigetti, the respective thresholds were chosen to be 6% and 2%. Moreover, for Rigetti we imposed locality constraints by forcing the joint size of the cluster and the neighborhood to be at most 5 by disregarding the smallest correlations. In practice the correlations within clusters were significantly higher than the chosen thresholds – heatmaps of all correlations can be found in Appendix E.2.

of the neighbors of the clusters in \mathcal{S} , *excluding* potential neighbors which themselves belong to the clusters in \mathcal{S} . Indeed, note that it might happen that a qubit from one cluster is a neighbor of a qubit from another cluster – in that case, one does not average over it but includes it in a noise model. Importantly, the average matrices $\Lambda_{av}^{\mathcal{S}}$ can be calculated explicitly using data obtained in the characterization of the readout noise.

4.2 Approximate noise mitigation

The noise matrix $\Lambda_{av}^{\mathcal{S}}$ can be used to construct the corresponding *effective correction matrix*

$$A_{av}^{\mathcal{S}} := \left(\Lambda_{av}^{\mathcal{S}}\right)^{-1}. \quad (11)$$

Correcting the marginal distribution via left-multiplication by the above ansatz matrix is not perfect and can introduce error in the mitigation. In the following Proposition 1 we provide an upper bound on that error measured in TV distance.

Proposition 1. *Let \mathbf{p}^{noisy} be a probability distribution on N qubits obtained from the N qubit probability distribution \mathbf{p}^{ideal} via stochastic transformation Λ of the form given in Eq. (5). Consider the subset of qubits $\mathcal{S} = \cup_{\chi \in \mathcal{A}\mathcal{C}_{\chi}}$. Let $\mathbf{p}_{\mathcal{S}}^{corr} = A_{av}^{\mathcal{S}} \mathbf{p}_{\mathcal{S}}^{noisy}$ be the result of the application to the marginal distribution $\mathbf{p}_{\mathcal{S}}^{noisy}$ of the correction procedure using the effective correction matrix $A_{av}^{\mathcal{S}}$ from Eq. (11). We then have the following inequality*

$$\begin{aligned} \text{TVD}\left(\mathbf{p}_{\mathcal{S}}^{corr}, \mathbf{p}_{\mathcal{S}}^{ideal}\right) &\leq \\ &\leq \frac{1}{2} \|\Lambda_{av}^{\mathcal{S}}\|_{1 \rightarrow 1} \max_{\mathbf{Y}_{\mathcal{N}(\mathcal{S})}} \|\Lambda_{av}^{\mathcal{S}} - \Lambda^{\mathbf{Y}_{\mathcal{N}(\mathcal{S})}}\|_{1 \rightarrow 1}, \end{aligned} \quad (12)$$

where the maximization goes over all possible states of the neighbors of \mathcal{S} .

The proof of the above Proposition is given in Appendix A.2 – it uses the convexity of the set of stochastic matrices, together with standard properties of matrix norms and with a triangle inequality (similar methods were used for providing error bounds on mitigated statistics in Ref. [13]). Note that the quantity on RHS of Eq. (12) shows resemblance to $c_{i \rightarrow j}$ in Eq. (8) (which we used to quantify correlations). Hence $\frac{1}{2} \max_{\mathbf{Y}_{\mathcal{N}(\mathcal{S})}} \|\Lambda_{av}^{\mathcal{S}} - \Lambda^{\mathbf{Y}_{\mathcal{N}(\mathcal{S})}}\|_{1 \rightarrow 1}$ can be interpreted as the maximal

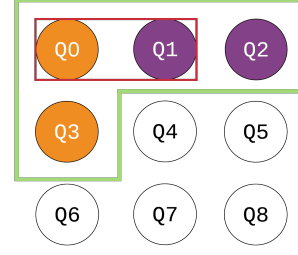


Figure 3: Illustration of the cluster structure of an exemplary 9-qubit device. There are two non-trivial clusters present ($\{0, 3\}$ and $\{1, 2\}$). For clarity, no neighborhood dependencies are shown, though in reality noise on the clusters can be dependent on the qubits outside the clusters. When one measures all the qubits but the goal is to estimate four-qubit marginal distribution on $\mathcal{S} = \{0, 3\} \cup \{1, 2\} = \{0, 1, 2, 3\}$ (green envelope), noise-mitigation should be performed based on the noise model for the whole set of qubits \mathcal{S} . On the other hand, when the goal is to estimate the two-qubit marginal on qubits $\mathcal{S}_k = \{0, 1\}$ (red envelope), it is still preferable to first perform error-mitigation on the four-qubit marginal on qubits $\mathcal{S} = \{0, 1, 2, 3\}$, and then take marginals over qubits 2 and 3 to obtain corrected marginal on \mathcal{S}_k . See the description in the text.

TVD between states on \mathcal{S} generated by $\Lambda_{av}^{\mathcal{S}}$ and states generated by $\Lambda^{\mathbf{Y}_{\mathcal{N}(\mathcal{S})}}$ (which appear in the description of the noise model). This can be also interpreted as a measure of dependence of noise between qubits in \mathcal{S} and the state of their neighbors just before measurement. Indeed, if the true noise does not depend on the state of the neighbors, the RHS of inequality Eq. (12) yields 0, and it grows when the noise matrices $\{\Lambda^{\mathbf{Y}_{\mathcal{N}(\mathcal{S})}}\}$ increasingly differ.

In practice, it might happen that one is interested in the marginal distribution on the qubits from some subset $\mathcal{S}_k \subset \mathcal{S}$ (red envelope in Fig. 3). In principle, one could then consider a coarse-graining of noise-model *within* \mathcal{S} (i.e., construction of noise model averaged over qubits from \mathcal{S} that do not belong to \mathcal{S}_k , treating those qubits like neighbors) and perform error-mitigation on the coarse-grained subset \mathcal{S}_k . However, due to the high level of correlations within clusters, we expect such a strategy to work worse than performing error-mitigation on \mathcal{S} , and then taking marginal to \mathcal{S}_k . Indeed, we observed numerous times that the latter strategy works better in practice. Yet, it is also more costly (since, by definition, \mathcal{S} is bigger than \mathcal{S}_k), hence in actual implementations with restricted resources

one may also consider implementing a coarse-grained strategy. In the following sections, we will focus on error-mitigation on the set \mathcal{S} . All those considerations can be easily generalized to the case of $\mathcal{S}_k \subset \mathcal{S}$.

4.3 Sampling complexity of error-mitigation

Let us now briefly comment on the sampling complexity of this error-mitigation scheme (the detailed discussion is postponed to Section 6). If one is interested in estimating an expected value of local Hamiltonian, a standard strategy is to estimate the local marginals and calculate the expected values of local Hamiltonian terms on those marginals. Without any error-mitigation, this has exponential sampling complexity in locality of marginals (which for local Hamiltonians is small), and logarithmic complexity in the number of local terms (hence, for typically considered Hamiltonians, also logarithmic in the number of qubits) – see Eq. (18) and its derivation in Appendix A.3. Now, if one adds to this picture error-mitigation *on marginals*, this, under reasonable assumptions, does not significantly change the scaling of the sampling complexity. We identify here two sources of sampling complexity increase (as compared to the non-mitigated marginal estimation). First, the noise mitigation via inverse of noise matrix does propagate statistical deviations – the bound on this quantitatively depends on the norm of the correction matrix (see Ref. [13] and detailed discussion around Eq. (18) in Section 6). Assuming that the local noise matrices are not singular (which is anyway required for error-mitigation to work), this increases sampling complexity by a constant factor. Second, the additional errors can come from the fact that, as described above, sometimes it is desirable to perform noise-mitigation on higher-dimensional marginals (if some qubits are highly correlated). However, assuming that readout noise has bounded locality, this can increase a sampling complexity only by a constant factor (this factor is proportional to the increase of the marginal size as compared to estimation without error-mitigation). In both cases, for a fixed size of marginals (as is the case for local Hamiltonians), it does not change the *scaling* of the sampling complexity with the number of qubits, which remains logarithmic.

5 Benchmark of the noise model

5.1 Energy estimation of local Hamiltonians

After having characterized the noise model, how to assess whether it is accurate? To answer this question we propose the following, application-driven heuristic benchmark. The main idea is to test whether the error-mitigation of local marginals based on the adopted noise model is accurate. To check this we propose to consider the problem of estimation of the expectation value $\langle \mathcal{H} \rangle$ of a local classical Hamiltonian

$$\mathcal{H} = \sum_{\alpha} H_{\alpha} \quad (13)$$

measured on its ground state $|\psi_0(\mathcal{H})\rangle$. Here by “local” we mean that the maximal number of qubits on which each H_{α} acts non-trivially does not scale with the system size. Classicality of the Hamiltonian means that every H_{α} is a linear combination of products of σ_z Pauli matrices. In turn the ground state $|\psi_0(\mathcal{H})\rangle$ can be chosen as classical i.e.,

$$|\psi_0(\mathcal{H})\rangle = |\mathbf{X}(\mathcal{H})\rangle, \quad (14)$$

for some bit-string $\mathbf{X}(\mathcal{H})$ representing one of the states from the computational basis. This problem is a natural candidate for error-mitigation benchmark due to at least three reasons. First, a variety of interesting optimization problems can be mapped to Ising-type Hamiltonians from Eq. (13). Indeed, this is the type of Hamiltonians appearing in the Quantum Approximate Optimization Algorithm. The goal of the QAOA is to get as close as possible to the ground state $|\psi_0(\mathcal{H})\rangle$. Second, the estimation of $\langle \mathcal{H} \rangle$ can be solved by the estimation of energy of local terms $\langle H_{\alpha} \rangle$ and therefore error-mitigation on marginals can be efficiently applied. Finally, the preparation of the classical ground state $|\psi_0\rangle$, once it is known, is very easy and requires only the application of local σ_x (NOT) gates. This works in our favor because we want to extract the effects of the readout noise, and single-qubit gates are usually of high quality in existing devices.

To perform the benchmark we propose to implement quantum circuits preparing ground states of many different local classical Hamiltonians, measure them on the noisy device, and perform two estimations of the energy - first from the raw data, and second with error-mitigation based

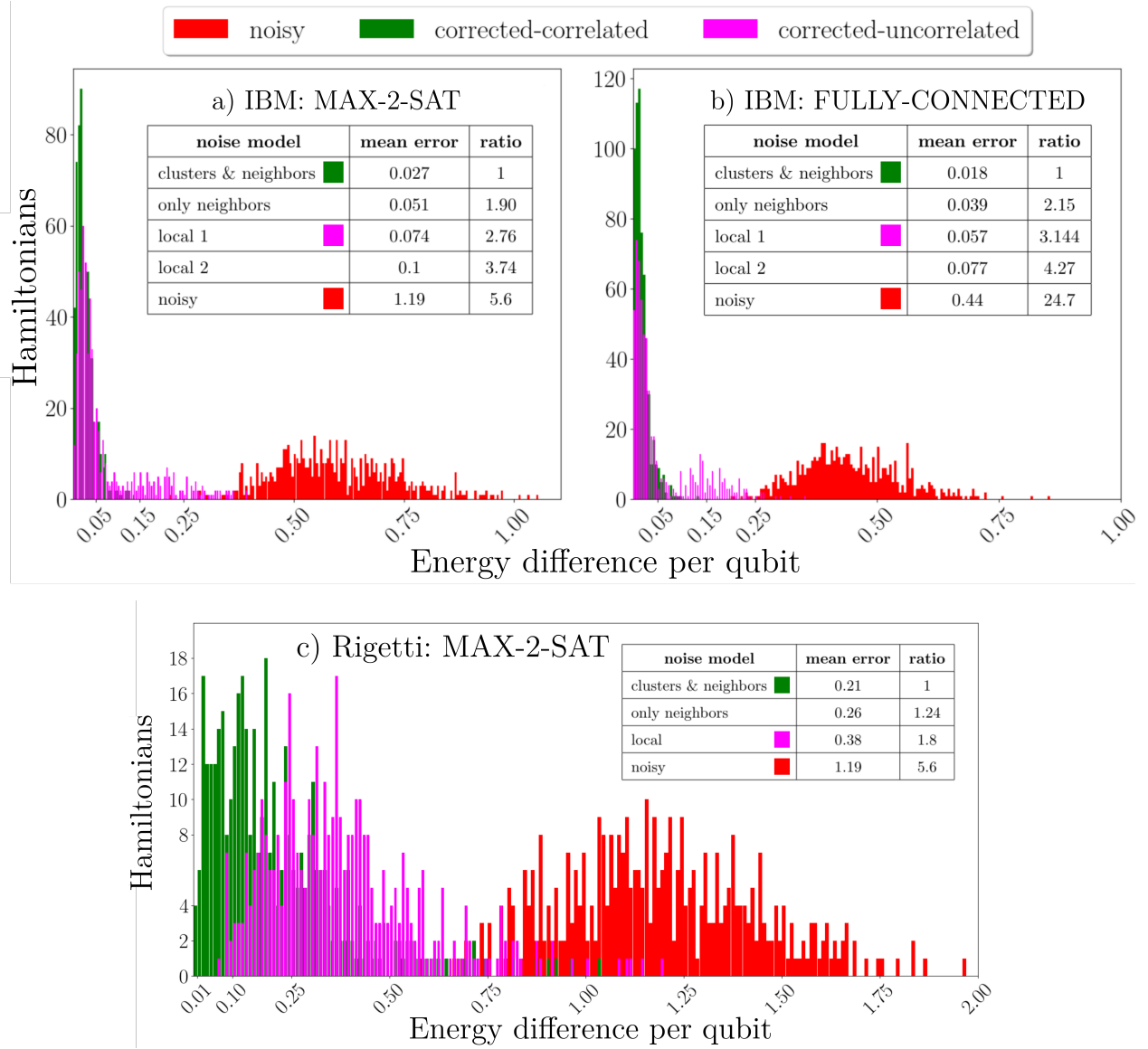


Figure 4: Results of an experimental benchmark of the readout noise mitigation on (a-b) IBM’s 15-qubit *Melbourne* device, and (c) a subset of 23 qubits on Rigetti’s *Aspen-8* device. Each histogram shows data for 600 (IBM’s) or 399 (Rigetti) different random Hamiltonians – (a,c) random MAX-2-SAT and (b) fully-connected graph with random interactions and local fields. The horizontal axis shows the absolute energy difference (between estimated and theoretical) divided by the number of qubits. The histogram comparison is done with no mitigation and with uncorrelated noise model characterization. The embedded tables show average errors depending on the adopted noise model. Here "ratio" refers to ratio of means. Additional second row in each figure shows data for noise model with only trivial (single-qubit) clusters and their neighborhoods. In case of IBM data, the additional fifth row illustrates memory effects. See description in the main text.

on our characterization. Naturally, it is also desirable to compare both with the error-mitigation based on a completely uncorrelated noise model (cf. Eq. (4)). We propose that if the mitigation based on a particular noise model works well on average (over some number of Hamiltonians), one can infer that the model is more accurate as well.

5.2 Experimental results

We applied the benchmark strategy described above on the 15-qubit IBM’s *Melbourne* device and 23-qubit subset of Rigetti’s *Aspen-8* device. We implemented Hamiltonians encoding random MAX-2-SAT instances with clause density 4 (600 on IBM and 399 on Rigetti) and fully-connected graphs with random interactions and local fields (600 on IBM). MAX-2-SAT instances were generated by considering all possible sets of $4 \times 15 = 60$ clauses with 8 variables, choosing one randomly and mapping it to Ising Hamiltonian (see, e.g., Ref. [52]). Random interactions and local fields for fully-connected graphs were chosen uniformly at random from $[-1, 1]$. Figure 4 presents the results of our experiments, together with a comparison with the uncorrelated noise model.

Let us first analyze the results of experiments performed on a 15-qubit IBM’s device. Here it is clear that the error-mitigation based on our noise model performs well, often reducing errors in estimation by as much as one order of magnitude. We further note that the uncorrelated noise model performs quite well (yet being visibly worse than ours). To compare the accuracy by using a single number (as opposed to looking at the whole histogram), we take the ratio of the mean deviations from the ideal energy for the error-mitigated data based on two models. The results are presented in tables embedded into Fig. 4. In those tables, we provide also additional experimental data. Namely, for each figure there the second row shows data for noise model labeled as "only neighbors". This corresponds to noise model in which each qubit is a trivial, single-qubit cluster, and correlations are included only via neighborhoods. The worse results of error-mitigation for such model as compared to full clusters-neighborhoods model motivates the introduction of non-trivial clusters. Furthermore, we found experimentally that the characterization of the uncorrelated noise model exhibits significant memory effects (see, for ex-

ample, Ref. [53]). Particularly, if one performs uncorrelated noise characterization in a standard way, i.e., by performing characterization in a separate job request to a provider, without any other preceding experiments (“local 2” in tables), the accuracy (measured by the error in energy after mitigation based on a given noise model) is much lower than for the characterization with some other experiments performed *prior* to the characterization of the uncorrelated model (“local 1” in tables). Indeed, the difference in mean accuracy can be as big as $\approx 26\%$.

Clearly, the overall performance of Rigetti’s 23 qubit device is lower than that for IBM’s device. First, the effects of noise (measured in energy error per qubit) are stronger. Second, the mean error with error-mitigation is only around ≈ 5.6 times smaller than the error without error-mitigation (as opposed to factor over 22 for IBM’s device). Third, the comparison to the uncorrelated noise model shows that the uncorrelated model performs not much worse than the correlated one.

Here we provide possible explanations of this poorer quality of experiments performed on Rigetti’s device. Due to the limited availability of the Rigetti’s device, we used a much lower number of samples to estimate Hamiltonian’s energies in those experiments. Specifically, each energy estimator for Rigetti’s experiments was calculated using only 1000 samples, while for IBM’s experiments the number of samples was 40960. This should lead to statistical errors higher by a factor of roughly $\sqrt{41} \approx 6.4$ (and note that the errors in error-mitigated energy estimation in Rigetti’s device are around 7.8 times higher than corresponding errors for IBM’s device for the same class of Hamiltonians). Similarly, we used fewer measurements to perform DDOT – on Rigetti’s device, we implemented 504 DDOT circuits sampled 1000 each, while on IBM’s device we performed 749 circuits sampled 8192 each. Less DDOT circuits imply less balanced collection, hence, as already mentioned, some correlations might have been overestimated. In summary, our characterization of this device was in general less accurate than on IBM’s device. This might be further amplified by the fact that single-qubit gates (which are used to implement DDOT circuits) were of lower quality for Rigetti’s device. Finally, as illustrated in Fig. 2, we observed that correlations in

measurement noise for Rigetti’s device are much more complex than in the case of IBM’s. As mentioned in the Figure’s description, to work around this we imposed locality constraints in the constructed noise model by disregarding the lowest correlations between qubits, which made the model less accurate.

We note that due to the limited availability of Rigetti’s device, we did not perform the study of memory effects similar to that performed in IBM. The local noise model presented for this device originates from a separate uncorrelated noise characterization performed prior to the rest of the experiments (hence it is analogous to the “local 2” model in IBM’s case).

To summarize, presented results suggest that in experiments on near-term quantum devices it will be indispensable to account for cross-talk effects in measurement noise. For both studied quantum devices we provided proof-of-principle experiments showing significant improvements in ground state energy estimation on the systems of sizes in the NISQ regime. Motivated by those results, we hope that the framework developed in this work will prove useful in the future, more complex experiments on even larger systems.

6 Error analysis for QAOA

In this section, we analyze the magnitude of errors resulting from our noise-mitigation scheme when applied to an energy estimation problem. Those errors result as a combination of two independent sources. First, from the fact that we use approximate correction matrices instead of the exact ones (see in Proposition 1). Secondly, by statistical errors due to the common practice of measuring *multiple* marginals *simultaneously* in a single run of the experiment. In the following, we will analyze the first and second effects separately and then provide a bound that takes both into account. We restrict our analysis to local Hamiltonians diagonal in the computational basis. A detailed derivation of the results below can be found in Appendices.

6.1 QAOA overview

Before starting, let us provide a short overview of the QAOA algorithm. In standard implementation [24], one initializes quantum system to be

in $|+\rangle^{\otimes N}$ state, where $|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$. Then p -layer QAOA is performed via implementation of unitaries of the form

$$U_p(\alpha, \beta) = \prod_j^p U_{\alpha_j} U_{\beta_j}, \quad (15)$$

where α, β are the angles to-be-optimized. Unitary matrices are given by $U_{\alpha_j} := \exp(-i \alpha_j \mathcal{H}_D)$, and $U_{\beta_j} := \exp(-i \beta_j \mathcal{H}_O)$, where \mathcal{H}_D is driver Hamiltonian (which we take to be $\mathcal{H}_D = \sum_k^N \sigma_x^k$), and \mathcal{H}_O is objective Hamiltonian that one wishes to optimize (i.e., to find approximation for its ground state energy). The quantum state after p -th layer is $|\psi_p\rangle = U_p |+\rangle^{\otimes N}$ and the function which is passed to classical optimizer is the estimator of the expected value $\langle \psi_p | \mathcal{H}_O | \psi_p \rangle$ (note that this makes those estimators to effectively be a function of parameters $\{\alpha_j\}, \{\beta_j\}$). The estimator is obtained by sampling from the distribution defined by the measurement of $|\psi_p\rangle$ in the computational basis, taking the relevant marginals, and calculating the expected value of \mathcal{H}_O using values of those estimated marginals. Let us now proceed to the analysis of possible sources of errors while performing noise-mitigation on the level of marginals to estimate the energy of local Hamiltonians, such as those present in QAOA.

6.2 Approximation errors

We start by recalling that performing noise mitigation with the average noise matrix instead of the exact one subjects the estimation of each marginal to an error upper bounded by Eq. (12). It follows that the correction of multiple marginal distributions can lead to the accumulation of errors which for each marginal α (we label subset of qubits by α so that local term H_α acts non-trivially on qubits from α) take the form

$$\delta^\alpha := \frac{1}{2} \|C_{av}^{\mathcal{S}_\alpha}\|_{1 \rightarrow 1} \max_{\mathbf{Y}_{\mathcal{N}(\mathcal{S}_\alpha)}} \|\Lambda_{av}^{\mathcal{S}_\alpha} - \Lambda^{\mathbf{Y}_{\mathcal{N}(\mathcal{S}_\alpha)}}\|_{1 \rightarrow 1} \quad (16)$$

where set $\mathcal{S}_\alpha = \cup_{\gamma \in \mathcal{A}} \mathcal{C}_\gamma$, $\mathcal{A} = \{\chi \mid \mathcal{C}_\chi \cap \alpha \neq \emptyset\}$, consists of clusters to which qubits from α belong, and $C_{av}^{\mathcal{S}_\alpha}$ is the average correction matrix for the marginal on that set. It is straightforward to show that the total possible deviation between the error-mitigated expected value $\langle \mathcal{H}^{\text{corr}} \rangle$

and the noiseless one $\langle \mathcal{H} \rangle$ is upper bounded by

$$|\langle \mathcal{H}^{\text{corr}} \rangle - \langle \mathcal{H} \rangle| \leq 2 \sum_{\alpha} \delta^{\alpha} \|H_{\alpha}\|$$

(additive approximation bound) . (17)

6.3 Additive statistical bound

Moving to the effect of measuring several marginals simultaneously, let us start by considering the simplest bound on the propagation of statistical deviations under our error-mitigation. In Appendix A we derive that the TVD (Eq. (9)) between the estimator $\mathbf{p}_{\alpha}^{\text{est}}$ and the actual local marginal \mathbf{p}_{α} is upper bounded by

$$\begin{aligned} \text{TVD}(\mathbf{p}_{\alpha}^{\text{est}}, \mathbf{p}_{\alpha}) &\leq \\ &\leq \epsilon^* := \sqrt{\frac{\log(2^n - 2) + \log\left(\frac{1}{P_{\text{err}}}\right) + \log(K)}{2s}}, \end{aligned}$$

(18)

where n is the number of qubits in the support of each local term (for simplicity we assume it to be the same for all H_{α}), K is the total number of local terms, s is the number of samples, and $1 - P_{\text{err}}$ is the confidence with which the above bound is stated. Importantly, the above bound is satisfied for each marginal *simultaneously*, hence the logarithmic overhead $\log(K)$. Using Eq. (18) together with standard norm inequalities one obtains the following bound for the total energy estimation

$$|\mathcal{H}_{\text{corr}}^{\text{est}} - \langle \mathcal{H} \rangle| \leq \sum_{\alpha} \|H^{\alpha}\| \|A_{\alpha}\|_{1 \rightarrow 1} \epsilon^*$$

(additive statistical bound) . (19)

Here $\mathcal{H}_{\text{corr}}^{\text{est}}$ denotes the estimator of the total energy with error-mitigation performed on each local term independently and A_{α} is the exact (not approximate) correction matrix on marginal α .

6.4 Joint approximation and statistical bound

The two bounds provided above took into account the two considered sources of errors independently. By using the triangle inequality (see Appendix A.4), we can now combine them to obtain

$$|\mathcal{H}_{\text{corr}}^{\text{est}} - \langle \mathcal{H} \rangle| \leq 2 \sum_{\alpha} \|H^{\alpha}\| \left(\underbrace{\epsilon^* \|A_{\alpha}^S\|_{1 \rightarrow 1}}_{\text{statistical errors}} + \underbrace{\delta_{\alpha}}_{\text{approximation errors}} \right).$$

(20)

It follows that the dominant scaling in the overall error are linear in the number of terms K caused by summing over all of them and the logarithmic overhead in ϵ^* added by the statistical errors.

6.5 Sampling complexity of energy estimation

While the additive bound from Eq. (19) could be tight *in principle*, we observed numerically on many occasions that in practice the statistical errors are much smaller (see Fig. 5 for exemplary results).

Here we will provide arguments that show that natural estimators of local energy terms H_{α} effectively behave as uncorrelated for a broad class of quantum states, hence leading to a significantly smaller total error than that obtained from an additive bound.

We start by describing in detail the natural strategy for energy estimation in the considered scenarios. In this work we are concerned with classical local Hamiltonians. This means that all local terms H_{α} can be measured simultaneously via a single computational basis measurement. The natural estimation procedure amounts to repeating s independent computational basis measurements on a quantum state ρ of interests. Outcomes of these measurements are then used to obtain local energy estimators $E_{\alpha}^{\text{est}} = \frac{1}{s} \sum_{i=1}^s E_{\alpha}^i$, where E_{α}^i are values of the local energy terms obtained in the i 'th experimental run. Now to perform estimation of expected value of energy, $\langle \mathcal{H} \rangle$, we simply sum the local estimators E_{α}^{est}

$$\mathcal{H}^{\text{est}} = \sum_{\alpha} H_{\alpha}^{\text{est}} = \frac{1}{s} \sum_{i=1}^s \sum_{\alpha} E_{\alpha}^i.$$

(21)

It is clear that \mathcal{H}^{est} is an unbiased estimator of $\langle \mathcal{H} \rangle$. Likewise E_{α}^{est} are unbiased estimators of $\langle H_{\alpha} \rangle$.

We would like to understand statistical properties of \mathcal{H}^{est} (specifically its variance) as a function of number of experimental runs (samples) s and the number of local terms in the Hamiltonian K . To this end we observe that random variables $E_{\alpha}^i, E_{\beta}^j$ are independent unless $i = j$ and therefore

$$\text{Var}(\mathcal{H}^{\text{est}}) = \frac{1}{s} \sum_{\alpha, \beta} \text{Cov}(E_{\alpha}^i, E_{\beta}^i),$$

(22)

Assuming that measurements of the energy E_{α}^i are distributed according to the probability com-

patible with the Born rule allows us to write

$$\text{Cov}(E_\alpha^i, E_\beta^i) = \text{Cov}(H_\alpha, H_\beta) = \langle H_\alpha H_\beta \rangle - \langle H_\alpha \rangle \langle H_\beta \rangle \quad (23)$$

Consequently we have

$$\text{Var}(\mathcal{H}^{\text{est}}) = \frac{1}{s} \text{Var}(\mathcal{H}) = \sum_{\alpha, \beta} \text{Cov}(H_\alpha, H_\beta). \quad (24)$$

The variance of \mathcal{H}^{est} can be related to the *sample complexity* of the energy estimation. Let $\Delta E > 0$ be some positive number. Then using Chebyshev inequality we get

$$\text{Prob}(|\mathcal{H}^{\text{est}} - \langle \mathcal{H} \rangle| \geq \Delta E) \leq \frac{\text{Var}(\mathcal{H}^{\text{est}})}{(\Delta E)^2}. \quad (25)$$

Choosing some parameter P_f as an upper bound on the RHS of this inequality, i.e., $\frac{\text{Var}(\mathcal{H}^{\text{est}})}{(\Delta E)^2} \leq P_f$. Then we obtain using (24) that for the number of samples s satisfying

$$\frac{\text{Var}(\mathcal{H})}{(\Delta E)^2 P_f} \leq s. \quad (26)$$

the estimator \mathcal{H}^{est} will be within accuracy from the expectation value $\langle \mathcal{H} \rangle$ with probability at least $1 - P_f$. Now, if different Hamiltonian terms \mathcal{H}_α are correlated then according to the above bound the sample complexity grows like K^2 , where K is the total number of terms in \mathcal{H} . Conversely, if $\text{Cov}(H_\alpha, H_\beta) \approx 0$ (for $\alpha \neq \beta$) then get sample complexity scaling linearly with K .

The above consideration can be equivalently translated to the estimates of the confidence intervals associated with estimator \mathcal{H}^{est} for a fixed value of samples s . Specifically, if local terms are uncorrelated, then the confidence interval (statistical error) will scale as square-root \sqrt{K} of the number of Hamiltonian terms, contrary to the pessimistic bound in Eq. (20) which is linear in K . We now show that one should expect that the sub-linear scaling of energy errors described above holds for a variety of quantum states, which in turn greatly reduces the sampling complexity compared to the pessimistic (linear) bound. We want to emphasize that our results are of immense practical importance for near-term devices. Our findings indicate a reduction (compared to the naive bounds) of sample complexity of energy estimation by orders of magnitude, even for relatively small systems ($K \approx 100$ and larger).

6.6 Generic 2-local Hamiltonians in QAOA

We start by considering states that appear at the beginning and at the end of QAOA. In recent work [54] it was shown that after p -th layer of QAOA optimization, for given two local terms H_α and H_β , there is no entanglement between qubits from α and qubits from β if they are further away from each other than $2p$ (on a graph corresponding to interactions present in a Hamiltonian). Therefore, for generic QAOA optimization, one can expect that for low p , i.e., at the beginning of the QAOA, the local Hamiltonian terms will be uncorrelated variables. On the other hand, it is well known that the ground states of local Hamiltonians are product states. Therefore, provided that QAOA worked well and converged to the state close to the ground state of objective Hamiltonian, the same arguments can be applied for the reversed QAOA circuit, and one expects local terms to be uncorrelated as well.

The following Proposition provides a more quantitative description for generic 2-local Hamiltonians corresponding to random graphs.

Proposition 2. *Consider Hamiltonian with connectivity given by Erdős-Rényi random graph in which each edge of the graph is added independently at random with some fixed probability. Assume that the probability of adding edge is chosen so the average degree of a node is equal to $q = \frac{K}{N}$, hence that a random graph has on average N nodes and K edges. For QAOA starting from product state, if the number of layers satisfies*

$$p < \frac{w \log(N)}{8 \log(2q/\ln(2))} - 1 \quad (27)$$

with $w < 1$, then with probability $1 - e^{-N^a}$ the variance of the Hamiltonian is bounded by

$$\text{Var}(\mathcal{H}) \leq f_{\mathcal{H}} q N^{A+1} \quad (28)$$

where

$$\begin{aligned} f_{\mathcal{H}} &= \max_{\alpha, \beta} \|H_\alpha\| \|H_\beta\|, \\ A &= w \frac{(2 + |\log_{2q}(\ln(2))|)}{(1 + |\log_{2q}(\ln(2))|)}, \\ a &= \frac{w}{3(1 + |\log_{2q}(\ln(2))|)}, \end{aligned} \quad (29)$$

where maximization in first definition goes over all two-qubit local terms acting on subsets of

qubits α and β . Since we can always choose $w < 1$ such that $A < 1$, the variance thus scales sub-quadratically for shallow depth QAOA. Importantly, the parameter w can be chosen in such a way that A goes asymptotically to 0 (an exemplary choice is $w = O(\log(N)^{-\frac{1}{2}})$, which in turn means that in the N regime the variance $\text{Var}(\mathcal{H})$ will scale almost linearly with the number of terms.

The proof of the above Proposition 2 uses insights from Ref. [54] and is delegated to Appendix D.2.

6.7 Random quantum states

A simple argument can be made to show that for generic Haar-random pure states, as well as random states appearing in random local quantum circuits, the variance of a local Hamiltonian \mathcal{H} behaves as if different energy terms were independent. Let $|\psi\rangle$ be a pure state on $(\mathbb{C}^2)^{\otimes N}$. For a subset of qubits γ , let ρ_γ denote the reduced density matrix $|\psi\rangle\langle\psi|$ on qubits in γ , and let id_γ denote the normalized maximally mixed state on qubits in γ . Assume now that local Hamiltonian terms H_α, H_β have disjoint supports. It can be then shown (see Proposition 4 in Appendices for the proof) that

$$\text{Cov}(H_\alpha, H_\beta) \leq 3 \|H_\alpha\| \|H_\beta\| \|\rho_{\alpha\cup\beta} - \text{id}_{\alpha\cup\beta}\|_1. \quad (30)$$

Now, it is a well-known fact [55] that with overwhelming probability *all* few-body marginals ρ_γ of a Haar-random multiqubit states $|\psi\rangle$ are exponentially close to maximally mixed states. Therefore, assuming that $\max_{\alpha,\beta}$ does not scale with the system size, we have that high probability over the choice of $|\psi\rangle$, for every disjoint terms H_α, H_β in a local Hamiltonian \mathcal{H}

$$\text{Cov}(H_\alpha, H_\beta) \approx 0. \quad (31)$$

The above reasoning mimics the computation done in Theorem 1 of [56], where it was used to establish that generic Haar-random pure states attain only the so-called standard quantum limit in the paradigmatic interferometric scenarios (again the underlying argument was based on the fact that *all* few-body reduced density matrices of a generic pure state $|\psi\rangle$ are very close to maximally mixed states with overwhelming probability).

Analogous analysis can be carried out for typical states generated by local random quantum

circuits. Such circuits are known to form approximate t -designs, i.e., capture properties of typical Haar-random unitaries captured by low-degree moments [57]. Specifically, a recent paper [58] considered evolution of local entropies for pure states $|\psi\rangle$ generated by shallow local random quantum circuits. From Theorem 1 of that work it directly follows that with probability greater than $1 - \delta$ over the choice random states $|\psi\rangle$ generated by random local circuits in the brick-wall architecture of depth r , all marginals ρ_γ of size $|\gamma| = k$, satisfy $\|\rho_\gamma - \text{id}_\gamma\|_1 \leq \epsilon$, where

$$\delta \leq \frac{\binom{N}{k}^2}{\epsilon^2} \left(2^{2k-N} + 2^k \left(\frac{4}{5} \right)^{2(r-1)} \right). \quad (32)$$

Clearly, if the size of the marginals k is fixed, setting $r = c \log(N/\epsilon)$, for a suitable constant c , allows us to conclude that for all γ such that $|\gamma| \leq k$ one has $\|\rho_\gamma - \text{id}_\gamma\|_1 \leq \epsilon$ with probability approaching 1 with the increasing system size.

6.8 Effects of measurement noise

To conclude, let us provide some analysis of the effects of measurement noise mitigation on the above considerations. First, let us note that it is straightforward to generalize all of the above arguments to include the *uncorrelated* readout noise. Intuitively, if the measurement noise is not correlated, it cannot increase the level of correlations in the energy estimators, and the same holds for noise-mitigation for such model (see Appendix D.3 for derivations).

For the correlated noise model, it is hard to provide analytical results, however, one can expect that if correlations in measurement noise are mild, then it should not drastically increase sample complexity. To test this hypothesis for small system sizes, we performed numerical simulations in the following way. Consider confidence intervals of the energy estimation. The most pessimistic bound on the error, as already explained, is given by Eq. (20) and is additive in the number of Hamiltonian terms K (if one considers the situation without measurement noise, then it suffices to set $\delta_\alpha = 0$ and $\|C_{av}^{S_\alpha}\|_{1 \rightarrow 1} = 1$ and Eq. (20) still holds). To obtain confidence intervals expected for uncorrelated variables, we

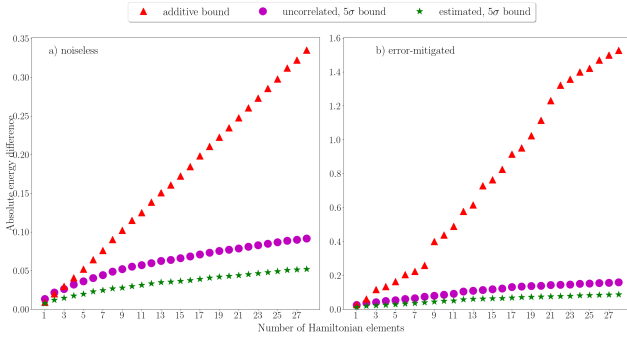


Figure 5: Numerical results illustrating low statistical errors in the estimation of the energy of local Hamiltonians. We show the results for the energy of 8-qubit random Sherrington-Kirkpatrick Hamiltonians, calculated on quantum states obtained *after the third layer* of QAOA optimization. The horizontal axis shows the number of Hamiltonian elements (in a sense that for a given Hamiltonian one estimates only some local terms). The Y-axis shows the absolute difference energy error. Each data point is an average over 92 Hamiltonians. Red triangles correspond to additive error bound from Eq. (20) (with correction matrix norm set to 1 and approximation errors to 0 for noiseless scenario). Magenta circles show the confidence intervals on energy estimation corresponding to 5σ -confidence (via Chebyshev’s bound) assuming *uncorrelated* variables for the sample size $s \approx 10^4$ (see Eq. (26) and description in the text). Green stars are analogous confidence intervals *estimated* in numerical simulations (for each Hamiltonian empirical standard deviations were obtained by repeating simulation 100 times). Plot a) shows results for the noiseless scenario, while b) shows results for the error-mitigated scenario with noise mitigation performed on marginals and noise model inspired by IBM’s device’s characterization.

simply set $\text{Var}(\mathcal{H}) = \sum_{\alpha} \text{Var}(H_{\alpha})$, and provide corresponding confidence interval by calculating LHS of Eq. (26). To test whether resulting bounds are close to what happens in practice, we numerically *estimate* the variance of \mathcal{H} and calculate the resulting confidence intervals. Such estimation of variance can be done, for example, by performing multiple numerical experiments, each giving an empirical estimate of $\text{Var}(\mathcal{H})$ (for a fixed number of samples s), and taking the mean of those estimates. Such comparison is plotted on Fig. 5 for the system of $N = 8$ qubits and Hamiltonians estimated on the states coming from 3-layer QAOA. Bound P_f on the probability (of energy estimator being outside the calculated confidence interval) in Chebyshev’s inequality is set to $P_f = 0.04$ which corresponds to 5σ -confidence. Shown are results for the noiseless scenario, and

for the error-mitigated estimators with the noise model inspired by IBM’s device characterization. It is clear that for tested Hamiltonians the confidence intervals in the estimation behave roughly like for uncorrelated variables.

7 Effects of measurement noise on QAOA – numerical study

In this section we apply our noise characterization and mitigation strategies to numerically study the effects of correlated measurement errors on QAOA and how they can be reduced with our techniques. As test Hamiltonians, we choose those encoding random MAX-2SAT instances with clause density 4, the Hamiltonians corresponding to fully-connected graphs with random interactions and local fields with magnitude from $[-1, 1]$, and the Sherrington-Kirkpatrick (SK) model in 2D (i.e. random Gaussian ZZ-interactions on a square lattice). For all models, we classically simulate a QAOA algorithm on an 8-qubit device with a number of layers ranging from $p = 3$ to 30. The parameter optimization is performed using Simultaneous perturbation stochastic approximation (SPSA) [39, 59–61] (see Appendix E.1 for details of optimization). As a correlated noise model, we adopt one inspired by our previous characterization of the IBM 15-qubit device.

We first simulated the result obtained by a QAOA algorithm where both the energy estimation and the gradient estimation used to guide the state evolution were affected by readout noise. For different numbers of gate layers, we compared the resulting energy estimators between an optimization guided by noisy estimators (“noisy-guided” in Fig. 6) and the optimization guided by estimators on which error-mitigation was performed (“corrected-guided” in Fig. 6). The results of our numerical studies are presented in the first column of Fig. 6, together with the noiseless case as a reference. Note that to make differences more visible, we set offset on the vertical axis. It is clear that for the considered Hamiltonians, the noise-mitigated estimators are much better than noisy ones. This suggests that our noise-mitigation scheme can be used to obtain an overall more reliable QAOA algorithm.

For our second analysis, we wanted to closely analyze the effects of noise (and its mitigation) on the parameter optimization only. To do so,

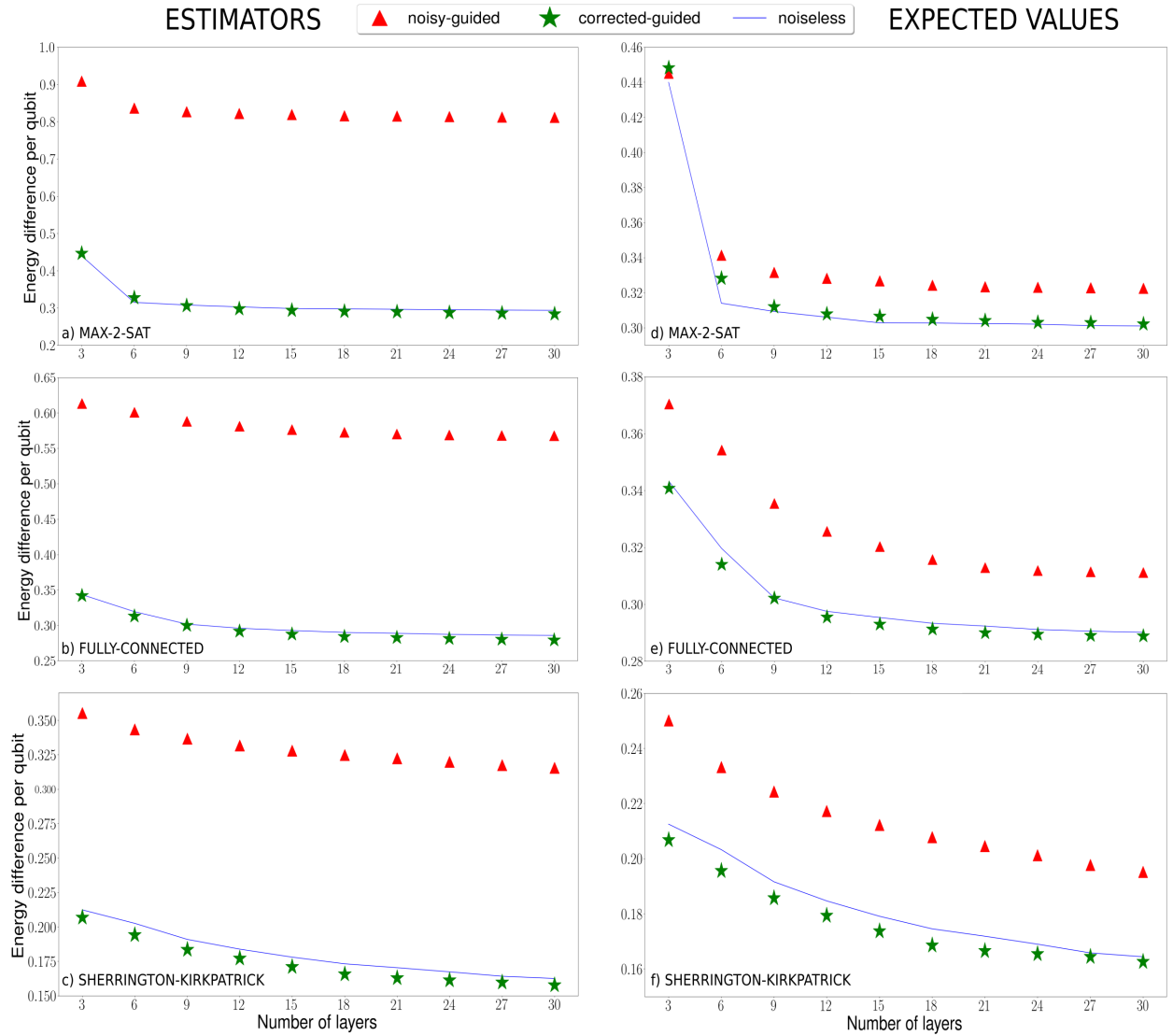


Figure 6: Numerical study of the effects of readout noise on energy estimation for QAOA on an 8-qubit system. The algorithm is used to prepare the ground state of a,d) Hamiltonians encoding random MAX 2-SAT problems, b,e) Hamiltonian corresponding to a fully-connected graph with uniformly random (from $[-1, 1]$) interactions and local fields, and c,f) the 2D Sherrington-Kirkpatrick model. In each plot, the horizontal axis shows a number of layers in the QAOA optimization, while the vertical axis shows the absolute difference between obtained and theoretical energies per qubit. Each data point is the *average* over 96 Hamiltonians. Note that we made an offset on the y-axis in order to make differences visible. Red data-points indicate optimization guided by noisy function evaluations, green points indicate noise-mitigated evaluations, while blue lines correspond to noiseless optimization given for reference. The estimators were obtained from $\approx 10^4$ samples. Due to differences in spectra of various Hamiltonians over which the mean is calculated the fluctuations around the presented means are very high, therefore for clarity we decided not to include error bars.

we still compare the results of QAOA between the cases where the optimization is guided by a noiseless, noisy (“noisy-guided” in Fig. 6) and noise-mitigated (“corrected-guided” in Fig. 6) energy estimators. However, in order to isolate the effect of noise on the optimization procedure, instead of sampling the energy from the noisy probability distributions, we calculated its expectation value directly on the quantum state obtained at each layer of the QAOA circuit (the same circuits as in left column). As a theoretical comparison between the optimal parameters found by QAOA, we took the distance between the resulting average energies on the state, divided by the number of qubits. The plots in the second column of Fig. 6 show our numerical results. For both of the considered models, the noise-mitigated optimization leads to better parameters regions. We note that the difference is not high, yet it is systematic. Since this is the case already for 8 qubits, one can expect that for bigger systems, the relative improvement will be higher.

To conclude, let us stress that the purpose of this section was to illustrate the possible effects of mildly correlated, realistic readout noise on QAOA algorithm. The above results clearly indicate that correlated readout noise can potentially influence the optimization, as opposed to an identical and uncorrelated one (see Ref. [35]). We find that in a number of instances the error-mitigation strategy helps to land in better parameters regions, however, we also note that the effects of noise on the optimization are not dramatic in the studied cases. From the point of view of near-term applications, this should be viewed as a positive result – the noise does not seem to strongly affect optimization, yet even those mild effects can be reduced by performing our error-mitigation. Furthermore, we note that due to the stochastic nature of the SPSA optimizer, the results might differ from run to run (each data point presented in the plots comes from the results of the optimization run which was the better one amongst two performed independent optimizations, see Appendix E.1 for details). Finally, we note that to obtain accurate estimates of energy, the noise-mitigation, unsurprisingly, remains highly beneficial.

8 Discussion

8.1 Conclusions

In the first part of this work, we proposed an efficiently describable model of correlated measurement noise in quantum detectors. The basic idea of the model is to group qubits into strongly-correlated clusters that are mildly affected by their neighborhoods, which, provided that the size of those groups is bounded by a constant, allows to describe a global noise model by much smaller number of parameters compared to the most generic situation. To characterize our noise model, we have introduced Diagonal Detector Overlapping Tomography, which is a procedure inspired by recently introduced Quantum Overlapping Tomography [33], tailored to the efficient characterization of the proposed measurement noise model. Similarly to the [33], proposed method can estimate k -qubit correlations in measurement noise affecting N -qubit device using $O(k2^k \log(N))$ quantum circuits. We have shown that the measurement noise can be efficiently mitigated in problems that require estimation of multiple marginal probability distributions, an example of which is Quantum Approximate Optimization Algorithm [24]. Importantly, from the fact that noise-mitigation is performed on marginal distributions it follows that sampling-complexity of noise mitigation is similar to that of the original problem, provided that cross-talk in readout noise is of bounded locality. We proposed a benchmark of the noise model and error reduction, which we implemented in experiments on up to 15 qubits on IBM’s *Melbourne* device and on 23 qubits on Rigetti’s *Aspen-8* device, and concluded significant improvements compared to simple, uncorrelated noise model. Interestingly, additional experimental data have pointed at previously unreported memory effects in IBM’s device that were demonstrated to not-negligibly change the results of experiments.

In the second part of the work, we provided an analysis of the statistical errors one may expect when performing the simultaneous estimation of multiple local terms of Hamiltonian on various classes of states. We provided simple arguments why low sampling complexity (i.e., scaling of the variance as the square-root of a number of local Hamiltonian terms) should be expected from Haar-random quantum states, and for states gen-

erated by shallow random quantum circuits. Similarly, we gave some arguments based on [54], why for states appearing at the beginning and at the end of the QAOA, one may expect that estimated local terms will effectively behave as uncorrelated variables, reducing the sampling complexity. Furthermore, we have provided analytical results for Hamiltonians encoding random MAX-2-SAT instances.

In the last part of the work, we have presented numerical results and extended discussion of the effects that correlated measurement noise can have on the performance of QAOA. We have demonstrated that already for 8 qubits the correlated measurement noise can alter the energy landscape in such a way, that the quantum-classical optimization leads to sub-optimal energy regions (compared to reference runs without noise). At the same time, we demonstrated that our noise-mitigation procedure can reduce those effects, improving optimization.

8.2 Future research

We believe that our findings will prove useful in both near- and long-term applications as efficient methods of characterizing and reducing measurement noise. At the same time, we find that a number of future research directions opens. It is natural to ask how well the error mitigation will perform in actual multi-qubit experiments of QAOA, which can be tested only with more access to the quantum devices than is available for the public via cloud services. Similarly, testing noise mitigation in more general, Variational Quantum Eigensolver (VQE) scenarios is of great interest [62, 63]. The analysis of statistical errors in the VQE setup is a natural extension – we note that since the local terms in VQE Hamiltonians do not commute, it is less straightforward than in the QAOA scenario. Another interesting problem is to design more benchmarks for the noise model and mitigation than the one method described in this work. A natural extension of techniques presented in this work would be to develop more methods for inference of correlations structure and construction of noise model from DDOT data. Having a noise model for marginal probability distributions, it is desirable to test and compare techniques of noise mitigation that go beyond a standard noise matrix inversion analyzed in this work. In particular, methods

based on Bayesian inference seem promising [18]. Moreover, it is tempting to ask whether error-mitigation performance and demonstrated memory effects are stable over time, a question which is often omitted (see however recent work [64] where systematic methods for studying time instabilities were developed and [65] which made an important contribution by studying the stability of various types of noise over time for IBM’s devices). Another open problem is to find out whether further generalizations of Quantum Overlapping Tomography can be used to perform reliable characterization of more general types of noise – coherent measurement noise and the noise affecting quantum gates. For example, in the work [66] the authors consider an ansatz for the generic noisy quantum process which uses only 2-local noise processes – a setup which seems natural to benefit from similar techniques. In Refs. [67, 68] the authors develop methods of estimating generic Pauli channels, and it would be of great interest to assess whether those methods can benefit from using measurement error-mitigation techniques such as ours. A very important research problem is that of mitigating measurement noise in scenarios which include estimators obtained from a very few samples, such as those in Refs. [69, 70]. We note that in its current form our methods cannot be directly implemented in such scenarios because they operate on marginal probability distributions. Finally, it would be extremely interesting to investigate whether our model of measurement noise is accurate for quantum devices based on architectures different than transmon qubits, such as flux qubits [71], trapped ions [72] or photonic quantum devices [73]. We intend to investigate some of the listed problems in future works.

Data availability

The Python code for creating DDOT perfect collections and for characterization of readout noise (in qiskit [74] and pyquill) will be available in our open-source Quantum Readout-Error Mitigation (QREM) GitHub repository [75]. We intend to further develop the package to contain practically relevant techniques of noise characterization and mitigation, including modifications and refinements of algorithms presented in this work. The results of experiments used in this work will be available as exemplary data-sets in the QREM

repository. Any requests for more experimental data or codes used in the creation of this work are welcome.

Acknowledgements

We are grateful to Tomek Rybotycki for providing help with running experiments on Rigetti’s device. We acknowledge the use of the IBM Quantum Experience and Amazon AWS Braket for this work. The views expressed are those of the authors and do not reflect the official policy or position of IBM, or the Rigetti, or the Amazon.

Funding

FBM and MO acknowledge support by the Foundation for Polish Science through IRAP project co-financed by EU within Smart Growth Operational Programme (contract no. 2018/MAB/5) and through TEAM-NET project (contract no. POIR.04.04.00-00-17C1/18-00). FB acknowledges support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project number 414325145 in the framework of the Austrian Science Fund (FWF): SFB F71. ZZ was supported by the NKFIH through the Quantum Technology National Excellence Program (Project No. 2017-1.2.1-NKP-2017-00001) and the Quantum Information National Laboratory of Hungary, and by the grants K124152, K124176, KH129601, K120569, FK 135220.

References

[1] Frank Arute et al. Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510, Oct 2019. ISSN 1476-4687. DOI: [10.1038/s41586-019-1666-5](https://doi.org/10.1038/s41586-019-1666-5). URL <https://doi.org/10.1038/s41586-019-1666-5>.

[2] Benjamin Villalonga, Dmitry Lyakh, Sergio Boixo, Hartmut Neven, Travis S Humble, Rupak Biswas, Eleanor G Rieffel, Alan Ho, and Salvatore Mandrà. Establishing the quantum supremacy frontier with a 281 pflop/s simulation. *Quantum Science and Technology*, 5(3):034003, Apr 2020. ISSN 2058-9565. DOI: [10.1088/2058-9565/ab7eeb](https://doi.org/10.1088/2058-9565/ab7eeb). URL <http://dx.doi.org/10.1088/2058-9565/ab7eeb>.

[3] Edward Farhi and Aram W Harrow. Quantum Supremacy through the Quan-

tum Approximate Optimization Algorithm. *arXiv e-prints*, art. arXiv:1602.07674, Feb 2016. URL <https://arxiv.org/abs/1602.07674>.

[4] Nikolaj Moll, Panagiotis Barkoutsos, Lev S. Bishop, Jerry M. Chow, Andrew Cross, Daniel J. Egger, Stefan Filipp, Andreas Fuhrer, Jay M. Gambetta, Marc Ganzhorn, Abhinav Kandala, Antonio Mezzacapo, Peter Müller, Walter Riess, Gian Salis, John Smolin, Ivano Tavernelli, and Kristan Temme. Quantum optimization using variational algorithms on near-term quantum devices. *Quantum Science and Technology*, 3(3):030503, Jul 2018. DOI: [10.1088/2058-9565/aab822](https://doi.org/10.1088/2058-9565/aab822). URL <https://arxiv.org/abs/1710.01022v2>.

[5] John Preskill. Quantum Computing in the NISQ era and beyond. *Quantum*, 2:79, August 2018. ISSN 2521-327X. DOI: [10.22331/q-2018-08-06-79](https://doi.org/10.22331/q-2018-08-06-79). URL <https://doi.org/10.22331/q-2018-08-06-79>.

[6] Joel J. Wallman and Joseph Emerson. Noise tailoring for scalable quantum computation via randomized compiling. *Physical Review A*, 94(5), Nov 2016. ISSN 2469-9934. DOI: [10.1103/physreva.94.052325](https://doi.org/10.1103/physreva.94.052325). URL <http://dx.doi.org/10.1103/PhysRevA.94.052325>.

[7] Ying Li and Simon C. Benjamin. Efficient variational quantum simulator incorporating active error minimization. *Physical Review X*, 7(2), Jun 2017. ISSN 2160-3308. DOI: [10.1103/physrevx.7.021050](https://doi.org/10.1103/physrevx.7.021050). URL <http://dx.doi.org/10.1103/PhysRevX.7.021050>.

[8] Kristan Temme, Sergey Bravyi, and Jay M. Gambetta. Error mitigation for short-depth quantum circuits. *Physical Review Letters*, 119(18), Nov 2017. ISSN 1079-7114. DOI: [10.1103/physrevlett.119.180509](https://doi.org/10.1103/physrevlett.119.180509). URL <http://dx.doi.org/10.1103/PhysRevLett.119.180509>.

[9] Suguru Endo, Simon C. Benjamin, and Ying Li. Practical Quantum Error Mitigation for Near-Future Applications. *Physical Review X*, 8:031027, Jul 2018. DOI: [10.1103/PhysRevX.8.031027](https://doi.org/10.1103/PhysRevX.8.031027).

[10] Abhinav Kandala, Kristan Temme, Antonio D. Córcoles, Antonio Mezzacapo, Jerry M. Chow, and Jay M. Gambetta. Er-

- ror mitigation extends the computational reach of a noisy quantum processor. *Nature*, 567(7749):491–495, Mar 2019. ISSN 1476-4687. DOI: [10.1038/s41586-019-1040-7](https://doi.org/10.1038/s41586-019-1040-7). URL <http://dx.doi.org/10.1038/s41586-019-1040-7>.
- [11] Jinzhao Sun, Xiao Yuan, Takahiro Tsunoda, Vlatko Vedral, Simon C. Benjamin, and Suguru Endo. Mitigating realistic noise in practical noisy intermediate-scale quantum devices. *Physical Review Applied*, 15(3), Mar 2021. ISSN 2331-7019. DOI: [10.1103/physrevapplied.15.034026](https://doi.org/10.1103/physrevapplied.15.034026). URL <http://dx.doi.org/10.1103/PhysRevApplied.15.034026>.
- [12] William J. Huggins, Sam McArdle, Thomas E. O’Brien, Joonho Lee, Nicholas C. Rubin, Sergio Boixo, K. Birgitta Whaley, Ryan Babbush, and Jarrod R. McClean. Virtual distillation for quantum error mitigation. 2020. URL <https://arxiv.org/abs/2011.07064>.
- [13] Filip B. Maciejewski, Zoltán Zimborás, and Michał Oszmaniec. Mitigation of readout noise in near-term quantum devices by classical post-processing based on detector tomography. *Quantum*, 4:257, April 2020. ISSN 2521-327X. DOI: [10.22331/q-2020-04-24-257](https://doi.org/10.22331/q-2020-04-24-257). URL <https://doi.org/10.22331/q-2020-04-24-257>.
- [14] Yanzhu Chen, Maziar Farahzad, Shinjae Yoo, and Tzu-Chieh Wei. Detector tomography on ibm quantum computers and mitigation of an imperfect measurement. *Physical Review A*, 100(5), Nov 2019. ISSN 2469-9934. DOI: [10.1103/physreva.100.052315](https://doi.org/10.1103/physreva.100.052315). URL <http://dx.doi.org/10.1103/PhysRevA.100.052315>.
- [15] Sergey Bravyi, Sarah Sheldon, Abhinav Kandala, David C. McKay, and Jay M. Gambetta. Mitigating measurement errors in multiqubit experiments. *Physical Review A*, 103(4), Apr 2021. ISSN 2469-9934. DOI: [10.1103/physreva.103.042605](https://doi.org/10.1103/physreva.103.042605). URL <http://dx.doi.org/10.1103/PhysRevA.103.042605>.
- [16] Michael R Geller and Mingyu Sun. Toward efficient correction of multiqubit measurement errors: pair correlation method. *Quantum Science and Technology*, 6(2):025009, feb 2021. DOI: [10.1088/2058-9565/abd5c9](https://doi.org/10.1088/2058-9565/abd5c9). URL <https://doi.org/10.1088/2058-9565/abd5c9>.
- [17] Michael R Geller. Rigorous measurement error correction. *Quantum Science and Technology*, 5(3):03LT01, Jun 2020. ISSN 2058-9565. DOI: [10.1088/2058-9565/ab9591](https://doi.org/10.1088/2058-9565/ab9591). URL <http://dx.doi.org/10.1088/2058-9565/ab9591>.
- [18] Benjamin Nachman, Miroslav Urbanek, Wibe A. de Jong, and Christian W. Bauer. Unfolding quantum computer readout noise. *npj Quantum Information*, 6(1):84, Sep 2020. ISSN 2056-6387. DOI: [10.1038/s41534-020-00309-7](https://doi.org/10.1038/s41534-020-00309-7). URL <https://doi.org/10.1038/s41534-020-00309-7>.
- [19] Hyeokjea Kwon and Joonwoo Bae. A hybrid quantum-classical approach to mitigating measurement errors in quantum algorithms. *IEEE Transactions on Computers*, page 1–1, 2020. ISSN 2326-3814. DOI: [10.1109/tc.2020.3009664](https://doi.org/10.1109/tc.2020.3009664). URL <http://dx.doi.org/10.1109/TC.2020.3009664>.
- [20] Kathleen E. Hamilton, Tyler Kharazi, Titus Morris, Alexander J. McCaskey, Ryan S. Bennink, and Raphael C. Pooser. Scalable quantum processor noise characterization. In *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)*, pages 430–440, 2020. DOI: [10.1109/QCE49297.2020.00060](https://doi.org/10.1109/QCE49297.2020.00060). URL <https://arxiv.org/abs/2006.01805>.
- [21] Megan L. Dahlhauser and Travis S. Humble. Modeling noisy quantum circuits using experimental characterization. *Phys. Rev. A*, 103:042603, Apr 2021. DOI: [10.1103/PhysRevA.103.042603](https://doi.org/10.1103/PhysRevA.103.042603). URL <https://link.aps.org/doi/10.1103/PhysRevA.103.042603>.
- [22] Lena Funcke, Tobias Hartung, Karl Jansen, Stefan Kühn, Paolo Stornati, and Xiaoyang Wang. Measurement error mitigation in quantum computers through classical bit-flip correction. 2020. URL <https://arxiv.org/abs/2007.03663>.
- [23] Muqing Zheng, Ang Li, Tamás Terlaky, and Xiu Yang. A bayesian approach for characterizing and mitigating gate and measurement errors. 2020. URL <https://arxiv.org/abs/2010.09188>.
- [24] Edward Farhi, Jeffrey Goldstone, and Sam

- Gutmann. A quantum approximate optimization algorithm. 2014. URL <https://arxiv.org/abs/1411.4028>.
- [25] M Cerezo, Andrew Arrasmith, Ryan Babush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and P. J. Coles. Variational quantum algorithms. 2020. URL <https://arxiv.org/abs/2012.09265>.
- [26] Stuart Hadfield, Zhihui Wang, Bryan O’Gorman, Eleanor Rieffel, Davide Venturelli, and Rupak Biswas. From the quantum approximate optimization algorithm to a quantum alternating operator ansatz. *Algorithms*, 12(2):34, Feb 2019. ISSN 1999-4893. DOI: [10.3390/a12020034](https://doi.org/10.3390/a12020034). URL <http://dx.doi.org/10.3390/a12020034>.
- [27] Seth Lloyd. Quantum approximate optimization is computationally universal. 2018. URL <https://arxiv.org/abs/1812.11075>.
- [28] M. E. S. Morales, J. D. Biamonte, and Z. Zimborás. On the universality of the quantum approximate optimization algorithm. *Quantum Information Processing*, 19(9):291, Aug 2020. ISSN 1573-1332. DOI: [10.1007/s11128-020-02748-9](https://doi.org/10.1007/s11128-020-02748-9). URL <https://doi.org/10.1007/s11128-020-02748-9>.
- [29] Pierre Hansen and Brigitte Jaumard. Algorithms for the maximum satisfiability problem. *Computing*, 44(4):279–303, Dec 1990. ISSN 1436-5057. DOI: [10.1007/BF02241270](https://doi.org/10.1007/BF02241270). URL <https://doi.org/10.1007/BF02241270>.
- [30] G. G. Guerreschi and A. Y. Matsuura. Qaoa for max-cut requires hundreds of qubits for quantum speed-up. *Scientific Reports*, 9(1):6903, May 2019. ISSN 2045-2322. DOI: [10.1038/s41598-019-43176-9](https://doi.org/10.1038/s41598-019-43176-9). URL <https://doi.org/10.1038/s41598-019-43176-9>.
- [31] Dmitry Panchenko. The sherrington-kirkpatrick model: An overview. *Journal of Statistical Physics*, 149(2):362–383, Sep 2012. ISSN 1572-9613. DOI: [10.1007/s10955-012-0586-7](https://doi.org/10.1007/s10955-012-0586-7). URL <http://dx.doi.org/10.1007/s10955-012-0586-7>.
- [32] Edward Farhi, Jeffrey Goldstone, Sam Gutmann, and Leo Zhou. The quantum approximate optimization algorithm and the sherrington-kirkpatrick model at infinite size. 2019. URL <https://arxiv.org/abs/1910.08187>.
- [33] Jordan Cotler and Frank Wilczek. Quantum overlapping tomography. *Physical Review Letters*, 124(10), Mar 2020. ISSN 1079-7114. DOI: [10.1103/PhysRevLett.124.100401](https://doi.org/10.1103/PhysRevLett.124.100401). URL <http://dx.doi.org/10.1103/PhysRevLett.124.100401>.
- [34] Jens Koch, Terri M. Yu, Jay Gambetta, A. A. Houck, D. I. Schuster, J. Majer, Alexandre Blais, M. H. Devoret, S. M. Girvin, and R. J. Schoelkopf. Charge-insensitive qubit design derived from the cooper pair box. *Phys. Rev. A*, 76:042319, Oct 2007. DOI: [10.1103/PhysRevA.76.042319](https://doi.org/10.1103/PhysRevA.76.042319). URL <https://link.aps.org/doi/10.1103/PhysRevA.76.042319>.
- [35] Cheng Xue, Zhao-Yun Chen, Yu-Chun Wu, and Guo-Ping Guo. Effects of quantum noise on quantum approximate optimization algorithm. *Chinese Physics Letters*, 38(3):030302, mar 2021. DOI: [10.1088/0256-307x/38/3/030302](https://doi.org/10.1088/0256-307x/38/3/030302). URL <https://doi.org/10.1088/0256-307x/38/3/030302>.
- [36] Jeffrey Marshall, Filip Wudarski, Stuart Hadfield, and Tad Hogg. Characterizing local noise in qaoa circuits. *IOP SciNotes*, 1(2):025208, Aug 2020. ISSN 2633-1357. DOI: [10.1088/2633-1357/abb0d7](https://doi.org/10.1088/2633-1357/abb0d7). URL <http://dx.doi.org/10.1088/2633-1357/abb0d7>.
- [37] Mahabubul Alam, Abdullah Ash-Saki, and Swaroop Ghosh. Analysis of quantum approximate optimization algorithm under realistic noise in superconducting qubits. 2019. URL <https://arxiv.org/abs/1907.09631>.
- [38] Matthew P. Harrigan et al. Quantum approximate optimization of non-planar graph problems on a planar superconducting processor. *Nature Physics*, 17(3):332–336, Mar 2021. ISSN 1745-2481. DOI: [10.1038/s41567-020-01105-y](https://doi.org/10.1038/s41567-020-01105-y). URL <https://doi.org/10.1038/s41567-020-01105-y>.
- [39] Ashley Montanaro and Stasja Stanisic. Compressed variational quantum eigensolver for the fermi-hubbard model. 2020. URL <https://arxiv.org/abs/2006.01179>.

- [40] Pranav Gokhale, Ali Javadi-Abhari, Nathan Earnest, Yunong Shi, and Frederic T. Chong. Optimized quantum compilation for near-term algorithms with openpulse. 2020. URL <https://arxiv.org/abs/2004.11205>.
- [41] Asher Peres. *Quantum theory: Concepts and methods*, volume 57. Springer Science & Business Media, 2006. DOI: <https://doi.org/10.1007/0-306-47120-5>.
- [42] J. S. Lundeen, A. Feito, H. Coldenstrodt-Ronge, K. L. Pregnell, Ch. Silberhorn, T. C. Ralph, J. Eisert, M. B. Plenio, and I. A. Walmsley. Tomography of quantum detectors. *Nature Physics*, 5:27, November 2008. URL <http://dx.doi.org/10.1038/nphys1133>.
- [43] Zdeněk Hradil, Jaroslav Řeháček, Jaromír Fiurášek, and Miroslav Ježek. *3 Maximum-Likelihood Methods in Quantum Mechanics*, pages 59–112. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-44481-7. DOI: [10.1007/978-3-540-44481-7_3](https://doi.org/10.1007/978-3-540-44481-7_3). URL https://doi.org/10.1007/978-3-540-44481-7_3.
- [44] Jaromír Fiurášek. Maximum-likelihood estimation of quantum measurement. *Physical Review A*, 64:024102, August 2001. DOI: [10.1103/PhysRevA.64.024102](https://doi.org/10.1103/PhysRevA.64.024102).
- [45] I. Gianani, Y.S. Teo, V. Cimini, H. Jeong, G. Leuchs, M. Barbieri, and L.L. Sánchez-Soto. Compressively certifying quantum measurements. *PRX Quantum*, 1(2), Oct 2020. ISSN 2691-3399. DOI: [10.1103/prxquantum.1.020307](https://doi.org/10.1103/prxquantum.1.020307). URL <http://dx.doi.org/10.1103/PRXQuantum.1.020307>.
- [46] Tim J. Evans, Robin Harper, and Steven T. Flammia. Scalable bayesian hamiltonian learning. 2019. URL <https://arxiv.org/abs/1912.07636>.
- [47] Nengkun Yu. Sample efficient tomography via Pauli measurements. 2020. URL <https://arxiv.org/abs/2009.04610>.
- [48] B. S. Majewski, N. C. Wormald, G. Havas, and Z. J. Czech. A Family of Perfect Hashing Methods. *The Computer Journal*, 39(6): 547–554, 01 1996. ISSN 0010-4620. DOI: [10.1093/comjnl/39.6.547](https://doi.org/10.1093/comjnl/39.6.547). URL <https://doi.org/10.1093/comjnl/39.6.547>.
- [49] D. R. Stinson, R. Wei, and L. Zhu. New constructions for perfect hash families and related structures using combinatorial designs and codes. *Journal of Combinatorial Designs*, 8(3):189–200, 2000. DOI: [https://doi.org/10.1002/\(SICI\)1520-6610\(2000\)8:3<189::AID-JCD4>3.0.CO;2-A](https://doi.org/10.1002/(SICI)1520-6610(2000)8:3<189::AID-JCD4>3.0.CO;2-A).
- [50] Simon R. Blackburn. Perfect hash families: Probabilistic methods and explicit constructions. *Journal of Combinatorial Theory, Series A*, 92(1):54 – 60, 2000. ISSN 0097-3165. DOI: <https://doi.org/10.1006/jcta.1999.3050>. URL <https://www.sciencedirect.com/science/article/pii/S0097316599930509>.
- [51] Noga Alon and Shai Gutner. Balanced Families of Perfect Hash Functions and Their Applications. May 2008. URL <https://arxiv.org/abs/0805.4300>.
- [52] Siddhartha Santra, Gregory Quiroz, Greg Ver Steeg, and Daniel A Lidar. Max 2-SAT with up to 108 qubits. *New Journal of Physics*, 16(4):045006, apr 2014. DOI: [10.1088/1367-2630/16/4/045006](https://doi.org/10.1088/1367-2630/16/4/045006). URL <https://doi.org/10.1088/1367-2630/16/4/045006>.
- [53] Kenneth Rudinger, Timothy Proctor, Dylan Langharst, Mohan Sarovar, Kevin Young, and Robin Blume-Kohout. Probing context-dependent errors in quantum processors. *Physical Review X*, 9(2), Jun 2019. ISSN 2160-3308. DOI: [10.1103/physrevx.9.021045](https://doi.org/10.1103/physrevx.9.021045). URL <http://dx.doi.org/10.1103/PhysRevX.9.021045>.
- [54] Edward Farhi, David Gamarnik, and Sam Gutmann. The quantum approximate optimization algorithm needs to see the whole graph: A typical case. 2020. URL <https://arxiv.org/abs/2004.09002>.
- [55] Sandu Popescu, Anthony J. Short, and Andreas Winter. Entanglement and the foundations of statistical mechanics. *Nature Physics*, 2(11):754–758, November 2006. DOI: [10.1038/nphys444](https://doi.org/10.1038/nphys444).
- [56] M. Oszmaniec, R. Augusiak, C. Gogolin, J. Kołodyński, A. Acín, and M. Lewenstein. Random bosonic states for robust quantum metrology. *Phys. Rev. X*, 6: 041044, Dec 2016. DOI: [10.1103/PhysRevX.6.041044](https://doi.org/10.1103/PhysRevX.6.041044).

- RevX.6.041044. URL <https://link.aps.org/doi/10.1103/PhysRevX.6.041044>.
- [57] Fernando G. S. L. Brandão, Aram W. Harrow, and Michał Horodecki. Local Random Quantum Circuits are Approximate Polynomial-Designs. *Communications in Mathematical Physics*, 346(2):397–434, September 2016. DOI: [10.1007/s00220-016-2706-8](https://doi.org/10.1007/s00220-016-2706-8).
- [58] Jordan Cotler, Nicholas Hunter-Jones, and Daniel Ranard. Fluctuations of subsystem entropies at late times. October 2020. URL <https://arxiv.org/abs/2010.11922>.
- [59] J. Spall. An overview of the simultaneous perturbation method for efficient optimization. *Johns Hopkins Apl Technical Digest*, 19:482–492, 1998. URL <https://www.jhuapl.edu/Content/techdigest/pdf/V19-N04/19-04-Spall.pdf>.
- [60] Chris Cade, Lana Minch, Ashley Montanaro, and Stasja Stanisic. Strategies for solving the fermi-hubbard model on near-term quantum computers. *Phys. Rev. B*, 102:235122, Dec 2020. DOI: [10.1103/PhysRevB.102.235122](https://doi.org/10.1103/PhysRevB.102.235122). URL <https://link.aps.org/doi/10.1103/PhysRevB.102.235122>.
- [61] Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M. Chow, and Jay M. Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature*, 549:242–246, Sep 2017. DOI: [10.1038/nature23879](https://doi.org/10.1038/nature23879).
- [62] William J. Huggins, Jarrod R. McClean, Nicholas C. Rubin, Zhang Jiang, Nathan Wiebe, K. Birgitta Whaley, and Ryan Babush. Efficient and noise resilient measurements for quantum chemistry on near-term quantum computers. *npj Quantum Information*, 7(1):23, Feb 2021. ISSN 2056-6387. DOI: [10.1038/s41534-020-00341-7](https://doi.org/10.1038/s41534-020-00341-7). URL <https://doi.org/10.1038/s41534-020-00341-7>.
- [63] George S. Barron and Christopher J. Wood. Measurement error mitigation for variational quantum algorithms. 2020. URL <https://arxiv.org/abs/2010.08520>.
- [64] Timothy Proctor, Melissa Revelle, Erik Nielsen, Kenneth Rudinger, Daniel Lobser, Peter Maunz, Robin Blume-Kohout, and Kevin Young. Detecting and tracking drift in quantum information processors. *Nature Communications*, 11(1):5396, Oct 2020. ISSN 2041-1723. DOI: [10.1038/s41467-020-19074-4](https://doi.org/10.1038/s41467-020-19074-4). URL <https://doi.org/10.1038/s41467-020-19074-4>.
- [65] Samudra Dasgupta and Travis S. Humble. Characterizing the stability of nisq devices. In *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)*, pages 419–429, 2020. DOI: [10.1109/QCE49297.2020.00059](https://doi.org/10.1109/QCE49297.2020.00059).
- [66] L. C. G. Govia, G. J. Ribeill, D. Ristè, M. Ware, and H. Krovi. Bootstrapping quantum process tomography via a perturbative ansatz. *Nature Communications*, 11(1):1084, Feb 2020. ISSN 2041-1723. DOI: [10.1038/s41467-020-14873-1](https://doi.org/10.1038/s41467-020-14873-1). URL <https://doi.org/10.1038/s41467-020-14873-1>.
- [67] Steven T. Flammia and Joel J. Wallman. Efficient estimation of Pauli channels. *ACM Transactions on Quantum Computing*, 1(1):1–32, Dec 2020. ISSN 2643-6817. DOI: [10.1145/3408039](https://doi.org/10.1145/3408039). URL <http://dx.doi.org/10.1145/3408039>.
- [68] Robin Harper, Steven T. Flammia, and Joel J. Wallman. Efficient learning of quantum noise. *Nature Physics*, 16(12):1184–1188, Aug 2020. ISSN 1745-2481. DOI: [10.1038/s41567-020-0992-8](https://doi.org/10.1038/s41567-020-0992-8). URL <http://dx.doi.org/10.1038/s41567-020-0992-8>.
- [69] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, Jun 2020. ISSN 1745-2481. DOI: [10.1038/s41567-020-0932-7](https://doi.org/10.1038/s41567-020-0932-7). URL <http://dx.doi.org/10.1038/s41567-020-0932-7>.
- [70] Senrui Chen, Wenjun Yu, Pei Zeng, and Steven T. Flammia. Robust shadow estimation. 2020. URL <https://arxiv.org/abs/2011.09636>.
- [71] Kelly Boothby, Paul Bunyk, Jack Raymond, and Aidan Roy. Next-generation topology of d-wave quantum processors. 2020. URL <https://arxiv.org/abs/2003.00133>.
- [72] Colin D. Bruzewicz, John Chiaverini, Robert McConnell, and Jeremy M. Sage. Trapped-

- ion quantum computing: Progress and challenges. *Applied Physics Reviews*, 6(2):021314, Jun 2019. ISSN 1931-9401. DOI: [10.1063/1.5088164](https://doi.org/10.1063/1.5088164). URL <http://dx.doi.org/10.1063/1.5088164>.
- [73] Jianwei Wang, Fabio Sciarrino, Anthony Laing, and Mark G. Thompson. Integrated photonic quantum technologies. *Nature Photonics*, 14(5):273–284, Oct 2019. ISSN 1749-4893. DOI: [10.1038/s41566-019-0532-1](https://doi.org/10.1038/s41566-019-0532-1). URL <http://dx.doi.org/10.1038/s41566-019-0532-1>.
- [74] Héctor Abraham et al. Qiskit: An open-source framework for quantum computing, 2019. URL <https://qiskit.org/documentation/>.
- [75] F. B. Maciejewski, T. Rybotycki, and M. Oszmaniec. Quantum readout errors mitigation (qrem) – open source github repository, 2020. URL <https://github.com/fbm2718/QREM>.
- [76] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo J. Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Technical Report HPL-2003-97R1*, Hewlett-Packard Labs, 08 2003. URL <https://www.hpl.hp.com/techreports/2003/HPL-2003-97R1.pdf?origin=publicationDetail>.
- [77] Zbigniew Puchała, Łukasz Paweła, Aleksandra Krawiec, and Ryszard Kukulski. Strategies for optimal single-shot discrimination of quantum measurements. *Physical Review A*, 98(4), Oct 2018. ISSN 2469-9934. DOI: [10.1103/PhysRevA.98.042103](https://doi.org/10.1103/PhysRevA.98.042103). URL <http://dx.doi.org/10.1103/PhysRevA.98.042103>.
- [78] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, 2010. DOI: [10.1017/CBO9780511976667](https://doi.org/10.1017/CBO9780511976667).
- [79] V. Akshay, H. Philathong, M. E. S. Morales, and J. D. Biamonte. Reachability deficits in quantum approximate optimization. *Phys. Rev. Lett.*, 124:090504, Mar 2020. DOI: [10.1103/PhysRevLett.124.090504](https://doi.org/10.1103/PhysRevLett.124.090504). URL <https://link.aps.org/doi/10.1103/PhysRevLett.124.090504>.

Appendices

We collect here technical results that are used in the main part of the paper. Some of the results stated here can be of independent interest for further works on quantum error mitigation. In Sec. [A](#) we discuss the details of error mitigation on marginals for correlated readout noise models, while Sec. [B](#) provides details of our noise characterization procedure based on the Diagonal Detector Overlapping Tomography technique. In Sec. [C](#) we give a short overview of the whole noise characterization scheme in a step-by-step manner. Results concerning the sample complexity of energy estimation are discussed in Sec. [D](#), and finally in Sec. [E](#) we provide some additional experimental data and details of the numerical simulations.

A Correlated readout noise model and its usage for error-mitigation on marginals

In this section, we provide some details on how correlated measurement noise affects marginal probability distributions. We start by providing in Appendix [A.1](#) explicit relation between ideal and noisy marginals, when the global probability distribution is affected by readout noise given by our model. Then in Appendix [A.2](#), we discuss the error-mitigation on the level of marginals, including proof of Proposition I from the main text. We finish by analyzing errors in simultaneous estimation of multiple marginal distributions (Appendix [A.3](#)) and estimation of multiple expected values of local terms (Appendix [A.4](#)).

A.1 Noise model for marginal distributions

In this subsection, we show how to translate a noise model for the full probability distribution given in Eq. [\(5\)](#) into a simple noise model for marginal distributions. Our results can be summarised in the

following Proposition 3.

Proposition 3. *Let $p(Y_1 Y_2 \dots Y_N)$ and $\tilde{p}(X_1 X_2, \dots, X_N)$ be the ideal distribution and the one obtained on noisy detector respectively. Define by $p(\mathbf{Y}_S)$ and $\tilde{p}(\mathbf{X}_S)$ their corresponding marginal distributions for qubits in some subset S . If the full distributions are connected by a stochastic matrix of the form Eq. (5) by means of*

$$\tilde{p}(X_1 \dots X_N) = \sum_{Y_1 Y_2 \dots Y_N} p(Y_1 Y_2 \dots Y_N) \Lambda_{X_1 X_2 \dots X_N | Y_1 Y_2 \dots Y_N}, \quad (33)$$

then there exists a left-stochastic matrix $\Lambda^{(S)}$ that connects the noisy marginal to the ideal one, namely

$$\tilde{p}(\mathbf{X}_S) = \sum_{\mathbf{Y}_S} \Lambda_{\mathbf{X}_S | \mathbf{Y}_S}^{(S)} p(\mathbf{Y}_S). \quad (34)$$

Moreover, the form of the marginal noise matrix can be explicitly derived from the decomposition Eq. (5).

We will now proceed to prove the above Proposition by deriving the concrete expression of $\Lambda_{\mathbf{X}_S | \mathbf{Y}_S}^{(S)}$ as a function of the matrices in the given noise model. Let us start by computing the marginal of the noisy distribution on general S . Let us denote by $C(S) = \{C_\chi\}_\chi$ set of clusters which contain qubits from S (note that we might be interested in taking marginal over some qubits from clusters to which qubits from S belong), and by $L(S) = \{\chi\}_\chi$ the corresponding set of labels of those clusters. Following from Eq. (33) with global noise map given by Eq. (5) we have

$$\begin{aligned} \tilde{p}(\mathbf{X}_S) &= \sum_{X \notin \mathbf{X}_S} \tilde{p}(X_1 X_2 \dots X_N) \\ &= \sum_{\substack{X \in C(S) \\ X \notin \mathbf{X}_S}} \sum_{X \notin C(S)} \sum_{Y_1 Y_2 \dots Y_N} p(Y_1 Y_2 \dots Y_N) \prod_X \Lambda_{\mathbf{X}_{C_\chi} | \mathbf{Y}_{C_\chi}}^{\mathbf{Y}_{N_\chi}} = \\ &= \sum_{\substack{X \in C(S) \\ X \notin \mathbf{X}_S}} \sum_{Y_1 Y_2 \dots Y_N} p(Y_1 Y_2 \dots Y_N) \prod_{\chi \in L(S)} \Lambda_{\mathbf{X}_{C_\chi} | \mathbf{Y}_{C_\chi}}^{\mathbf{Y}_{N_\chi}} \prod_{\chi \notin L(S)} \sum_{X \notin C(S)} \Lambda_{\mathbf{X}_{C_\chi} | \mathbf{Y}_{C_\chi}}^{\mathbf{Y}_{N_\chi}} = \\ &= \sum_{\substack{X \in C(S) \\ X \notin \mathbf{X}_S}} \sum_{Y_1 Y_2 \dots Y_N} p(Y_1 Y_2 \dots Y_N) \prod_{\chi \in L(S)} \Lambda_{\mathbf{X}_{C_\chi} | \mathbf{Y}_{C_\chi}}^{\mathbf{Y}_{N_\chi}}, = \\ &= \sum_{\substack{X \in C(S) \\ X \notin \mathbf{X}_S}} \sum_{Y \in \mathbf{Y}_{\mathcal{N}(S) \cup C(S)}} p(\mathbf{Y}_{\mathcal{N}(S) \cup C(S)}) \prod_{\chi \in L(S)} \Lambda_{\mathbf{X}_{C_\chi} | \mathbf{Y}_{C_\chi}}^{\mathbf{Y}_{N_\chi}}, \end{aligned} \quad (35)$$

where to obtain the above simplifications, we have exploited the fact that the noise matrices $\Lambda^{\mathbf{Y}_{N_\chi}}$ are all left-stochastic (for any fixed \mathbf{Y}_{N_χ}) and we have defined $p(\mathbf{Y}_{\mathcal{N}(S) \cup C(S)})$ as the marginal of the ideal distribution on the qubits belonging both the clusters and the neighborhoods of qubits from S . By $\mathcal{N}(S)$ we denote set of qubits from neighbourhoods of clusters $C(S)$ but without including the qubits from S , i.e., $\mathcal{N}(S) = \cup_{\chi \in L(S)} (\mathcal{N}_\chi / S)$. Note that this additional requirement is necessary since qubits which are neighbors of some clusters in $C(S)$ might belong to some other clusters in $C(S)$.

Now by means of the chain rule for probability we decompose

$$p(\mathbf{Y}_{\mathcal{N}(S) \cup C(S)}) = p(\mathbf{Y}_S) p(\mathbf{Y}_{\{\mathcal{N}(S) \cup C(S)\} / \{S\}} | \mathbf{Y}_S) \quad (36)$$

which after substituting to the Eq. (35) gives

$$\tilde{p}(\mathbf{X}_S) = \sum_{Y \in \mathbf{Y}_S} p(\mathbf{Y}_S) \sum_{\substack{X \in C(S) \\ X \notin \mathbf{X}_S}} \sum_{\substack{Y \in \mathbf{Y}_{\mathcal{N}(S) \cup C(S)} \\ Y \notin \mathbf{Y}_S}} p(\mathbf{Y}_{\{\mathcal{N}(S) \cup C(S)\} / \{S\}} | \mathbf{Y}_S) \prod_{\chi \in L(S)} \Lambda_{\mathbf{X}_{C_\chi} | \mathbf{Y}_{C_\chi}}^{\mathbf{Y}_{N_\chi}}. \quad (37)$$

Now notice that Eq. (37) coincides exactly with Eq. (34) if we identify the marginal noise matrix elements as

$$\Lambda_{\mathbf{X}_S|Y_S}^{(\mathcal{S})} := \sum_{\substack{X \in C(\mathcal{S}) \\ X \notin \mathbf{X}_S}} \sum_{\substack{Y \in \mathbf{Y}_{\mathcal{N}(\mathcal{S}) \cup C(\mathcal{S})} \\ Y \notin \mathbf{Y}_S}} p\left(\mathbf{Y}_{\{\mathcal{N}(\mathcal{S}) \cup C(\mathcal{S})\} / \{\mathcal{S}\}} | \mathbf{Y}_S\right) \prod_{\chi \in L(\mathcal{S})} \Lambda_{\mathbf{X}_{C_\chi} | \mathbf{Y}_{C_\chi}}^{\mathbf{Y}_{\mathcal{N}_\chi}}, \quad (38)$$

which shows, unsurprisingly, that the noise matrix acting on the marginal is state-dependent (via marginal distribution $p\left(\mathbf{Y}_{\{\mathcal{N}(\mathcal{S}) \cup C(\mathcal{S})\} / \{\mathcal{S}\}} | \mathbf{Y}_S\right)$). Note that the above equation *does not* include simple matrix multiplication of noise matrices for neighborhoods $\mathcal{N}(\mathcal{S})$. In other words, the generic marginal noise matrix *is not* a convex combination of the matrices from the set

$$\cup_{\chi \in L(\mathcal{S})} \left\{ \Lambda_{\mathbf{Y}_{\mathcal{N}_\chi}}^{\mathbf{Y}_{\mathcal{N}_\chi}} \right\} \quad (39)$$

even if the neighborhoods do not overlap.

It is now left to show that the matrix defined in Eq. (38) is left-stochastic. To this aim, let us sum over string \mathbf{X}_S and obtain

$$\begin{aligned} \sum_{\mathbf{X}_S} \Lambda_{\mathbf{X}_S|Y_S}^{(\mathcal{S})} &:= \sum_{\mathbf{X}_S} \sum_{\substack{X \in C(\mathcal{S}) \\ X \notin \mathbf{X}_S}} \sum_{\substack{Y \in \mathbf{Y}_{\mathcal{N}(\mathcal{S}) \cup C(\mathcal{S})} \\ Y \notin \mathbf{Y}_S}} p\left(\mathbf{Y}_{\{\mathcal{N}(\mathcal{S}) \cup C(\mathcal{S})\} / \{\mathcal{S}\}} | \mathbf{Y}_S\right) \prod_{\chi \in L(\mathcal{S})} \Lambda_{\mathbf{X}_{C_\chi} | \mathbf{Y}_{C_\chi}}^{\mathbf{Y}_{\mathcal{N}_\chi}} = \\ &= \sum_{\substack{Y \in \mathbf{Y}_{\mathcal{N}(\mathcal{S}) \cup C(\mathcal{S})} \\ Y \notin \mathbf{Y}_S}} p\left(\mathbf{Y}_{\{\mathcal{N}(\mathcal{S}) \cup C(\mathcal{S})\} / \{\mathcal{S}\}} | \mathbf{Y}_S\right) \sum_{\mathbf{X}_S} \prod_{\chi \in L(\mathcal{S})} \Lambda_{\mathbf{X}_{C_\chi} | \mathbf{Y}_{C_\chi}}^{\mathbf{Y}_{\mathcal{N}_\chi}} = \\ &= \sum_{\substack{Y \in \mathbf{Y}_{\mathcal{N}(\mathcal{S}) \cup C(\mathcal{S})} \\ Y \notin \mathbf{Y}_S}} p\left(\mathbf{Y}_{\{\mathcal{N}(\mathcal{S}) \cup C(\mathcal{S})\} / \{\mathcal{S}\}} | \mathbf{Y}_S\right) = 1, \end{aligned}$$

where we simply used the fact that the cluster noise matrices are left-stochastic and that the conditional distribution $p\left(\mathbf{Y}_{\{\mathcal{N}(\mathcal{S}) \cup C(\mathcal{S})\} / \{\mathcal{S}\}} | \mathbf{Y}_S\right)$ is normalised.

A.2 Noise mitigation for marginal distributions – proof of Proposition 1

Here we analyze in detail the noise mitigation strategy outlined in the main text, focusing on its scalability and effectiveness to recover the ideal marginal distribution. Again let us assume that we are interested in marginal on qubits from some subset \mathcal{S} .

Let us start by outlining the reasoning behind our choice of the mitigation strategy. Recall that, as proved in the previous section, the noisy marginals can always be related to ideal ones by means of a marginal stochastic matrix as for Eq. (34). Hence, taking the inverse of the marginal noise matrix would be the natural strategy for correct noise mitigation. However, recall that, as shown in Eq. (38), the form of the marginal noise matrix depends on the ideal distribution itself, which is generally unknown. Therefore one needs an ansatz for such a distribution, which hopefully will work in the mitigation generic case of arbitrary conditional distribution. As indicated in the main text, the natural choice for such ansatz is a uniform distribution (inserted in place of $p\left(\mathbf{Y}_{\{\mathcal{N}(\mathcal{S}) \cup C(\mathcal{S})\} / \{\mathcal{S}\}} | \mathbf{Y}_S\right)$ in Eq. (38)).

Now, let us make the following important observation. We are interested in mitigation of the noise of a $|\mathcal{S}|$ -qubit marginal distribution of qubits generally belonging to some set of distinct clusters $C(\mathcal{S})$. Note that, by definition, a cluster is a set of qubits with correlations in errors so big, that one can not consider the measurement outcomes on them separately. We argue that, if one is interested in correcting the marginal distribution of only *parts* of the clusters, say $\tilde{p}(\mathbf{X}_S)$, it is still a better idea *first* correct the marginal distribution on the whole variables in clusters, namely

$$\tilde{p}(\mathbf{X}_{C(\mathcal{S})}) = \sum_{\mathbf{Y}_{C(\mathcal{S})}} p(\mathbf{Y}_{C(\mathcal{S})}) \Lambda_{\mathbf{X}_{C(\mathcal{S})} | \mathbf{Y}_{C(\mathcal{S})}}^{(C(\mathcal{S}))}. \quad (40)$$

Notice that the corrected cluster marginal distribution can always be post-processed to obtain the two-body marginal distribution of interest. By proceeding in a similar manner as in the previous section, one can obtain the following expression for the cluster noise matrix

$$\Lambda_{\mathbf{x}_{C(\mathcal{S})}|\mathbf{Y}_{C(\mathcal{S})}}^{(C(\mathcal{S}))} = \sum_{\mathbf{Y}_{\mathcal{N}(\mathcal{S})}} p(\mathbf{Y}_{\mathcal{N}(\mathcal{S})}|\mathbf{Y}_{C(\mathcal{S})}) \prod_{\chi \in L(\mathcal{S})} \Lambda_{\mathbf{x}_{C_\chi}|\mathbf{Y}_{C_\chi}}^{\mathbf{Y}_{\mathcal{N}_\chi}}, \quad (41)$$

where $L(\mathcal{S})$ is a set of labels of the clusters $\mathcal{C}(\mathcal{S})$ (note that those are also labels for neighbours of those clusters). Recall that $\mathcal{N}(\mathcal{S})$ was defined as set of qubits from neighbourhoods of clusters $C(\mathcal{S})$ but without including the qubits from \mathcal{S} , i.e., $\mathcal{N}(\mathcal{S}) = \cup_{\chi \in L(\mathcal{S})} (\mathcal{N}_\chi \setminus \mathcal{S})$, so in the above definition we do not average over some qubits in clusters. As mentioned before, since the exact form of such a matrix requires access to the unknown information of $p(\mathbf{Y}_{\mathcal{N}(\mathcal{S})}|\mathbf{Y}_{C(\mathcal{S})})$, we propose instead to invert the average cluster matrix

$$\Lambda_{av}^{C(\mathcal{S})} := \frac{1}{2^{|\mathcal{N}(\mathcal{S})|}} \sum_{\mathbf{Y}_{\mathcal{N}(\mathcal{S})}} \Lambda^{\mathbf{Y}_{\mathcal{N}(\mathcal{S})}}, \quad (42)$$

and use it to perform the error mitigation. From our numerical analysis, the replacement of the cluster noise matrix with its average version yields a much better correction than making a similar replacement at the level of the $|\mathcal{S}|$ -body marginals (however we tested in only for the case of 2-qubit marginal distributions). This provides a clear indication in favor of performing the error mitigation at the level of the clusters instead of the single variables. Moreover, notice that the computational cost of computing the two noise matrices in Eq. (41) and (38) is comparable. Indeed, both require access to the collections of matrices $\{\Lambda^{\mathbf{Y}_{\mathcal{N}(\mathcal{S})}}\}_{\mathbf{Y}_{\mathcal{N}(\mathcal{S})}}$. Hence, as long as the size of the involved clusters and neighborhoods is reasonably small (particularly, it does not scale with the system size), computing the matrix in Eq. (42) is efficient.

Due to the above, the following discussion will concern the mitigation on the level of marginals on the level of clusters $C(\mathcal{S})$. For clarity, let us from now on use slightly changed notation

$$\mathcal{S} \rightarrow \mathcal{S} = \cup_{\chi} \mathcal{C}_\chi, \quad (43)$$

to indicate that we are interested in error-mitigation on the subset of qubits \mathcal{S} which consists of full clusters, i.e., prior to performing error-mitigation we do not wish to take marginal over qubits belonging to the same cluster (while later one can of course marginalize the corrected distributions).

Having obtained $\Lambda_{av}^{\mathcal{S}}$, one can use its inverse as a correction matrix which can reduce the noise on the marginal distribution, by defining the corrected distribution as

$$\mathbf{p}_{\mathcal{S}}^{\text{corr}} = \left(\Lambda_{av}^{\mathcal{S}}\right)^{-1} \tilde{\mathbf{p}}_{\mathcal{S}}, \quad (44)$$

where we used vector notation for clarity. $\tilde{\mathbf{p}}_{\mathcal{S}}$ is noisy distribution on qubits from \mathcal{S} . Similarly, the vector without tilde symbol $\mathbf{p}_{\mathcal{S}}$ will denote corresponding ideal distribution, i.e., we have $\tilde{\mathbf{p}}_{\mathcal{S}} = \Lambda^{(\mathcal{S})} \mathbf{p}_{\mathcal{S}}$. Error reduction via average matrix is perfect only in the case ¹ of infinite-statistics and under the assumption that the actual conditional distribution in Eq. (41) is uniform (because then $\Lambda^{\mathcal{S}} = \Lambda_{av}^{\mathcal{S}}$). Since this scenario is not realistic, in practice one can hope only for partial noise mitigation, and not its complete reverse. The errors related to statistical noise when correcting probability distribution were thoroughly analyzed in work [13] (we also repeat a similar analysis in the context of expected values of Hamiltonians in next sections). Here we will analyze the errors which arise due to the fact that we use the inverse of the average cluster matrix $\Lambda_{av}^{(\mathcal{S})}$ instead of inverse of the exact matrix $\Lambda^{(\mathcal{S})}$, while keeping the assumption of infinite statistics. We will denote effective correction matrix as $A_{av}^{(\mathcal{S})} := \left(\Lambda_{av}^{(\mathcal{S})}\right)^{-1}$ (as in Eq. (11)).

¹Clearly, the implicit assumption is that our model of noise is exact.

To quantify the errors, we make use the Total-Variation-Distance (see Eq. (89)) to measure the distance between probability distributions. For later purposes, let us also recall that the operator norm induced by the vector L1 norm is

$$\|A\|_{1 \rightarrow 1} = \sup_{\|v\|_1=1} \|A|v\rangle\|_1, \quad (45)$$

where A is any linear operator acting on given vector space.

Now we will proceed to proving Proposition 1 which states that the maximum error due to using average correction matrix is given by

$$\text{TVD}(\mathbf{p}_S^{\text{corr}}, \mathbf{p}^S) \leq \frac{1}{2} \|A_{av}^S\|_{1 \rightarrow 1} \max_{\mathbf{Y}_{\mathcal{N}(S)}} \|\Lambda_{av}^S - \Lambda^{\mathbf{Y}_{\mathcal{N}(S)}}\|_{1 \rightarrow 1}, \quad (46)$$

As a starting point for the proof, let us make following decomposition $\Lambda^{(S)} = \Lambda_{av}^{(S)} + (\Lambda^{(S)} - \Lambda_{av}^{(S)})$ so to replace the left-hand-side of Eq. (46) with

$$\begin{aligned} \|A_{av}^{(S)} (\Lambda^{(S)} \tilde{\mathbf{p}}_S) - \mathbf{p}_S\|_1 &= \|A_{av}^{(S)} (\Lambda^{(S)} - \Lambda_{av}^{(S)}) \mathbf{p}_S\|_1 \\ &\leq \sup_{\mathbf{p}_S} \|A_{av}^{(S)} (\Lambda^{(S)} - \Lambda_{av}^{(S)}) \mathbf{p}_S\|_1 \\ &\leq \|A_{av}^{(S)}\|_{1 \rightarrow 1} \|\Lambda^{(S)} - \Lambda_{av}^{(S)}\|_{1 \rightarrow 1}, \end{aligned} \quad (47)$$

where the first inequality indicates the maximization over all possible probability distributions over the set \mathcal{S} variables and the second inequality follows from Eq. (45) and the sub-multiplicativity of the L1 norm. Now, let us bound the second term of the above equation as follows

$$\begin{aligned} \|\Lambda^{(S)} - \Lambda_{av}^{(S)}\|_{1 \rightarrow 1} &= \left\| \left(\sum_{\mathbf{Y}_{\mathcal{N}(S)}} p(\mathbf{Y}_{\mathcal{N}(S)} | \mathbf{Y}_S) \prod_{\chi \in L(S)} \Lambda_{\mathbf{x}_{c_\chi}}^{\mathbf{Y}_{\mathcal{N}_\chi}} \right) - \Lambda_{av}^{(S)} \right\|_{1 \rightarrow 1} \\ &\leq \sup_{\{p(\mathbf{Y}_{\mathcal{N}(S)} | \mathbf{Y}_S)\}} \left\| \left(\sum_{\mathbf{Y}_{\mathcal{N}(S)}} p(\mathbf{Y}_{\mathcal{N}(S)} | \mathbf{Y}_S) \prod_{\chi \in L(S)} \Lambda_{\mathbf{x}_{c_\chi}}^{\mathbf{Y}_{\mathcal{N}_\chi}} \right) - \Lambda_{av}^{(S)} \right\|_{1 \rightarrow 1}, \end{aligned} \quad (48)$$

where the above sequence of inequalities removes the dependence on the unknown conditional distribution $p(\mathbf{Y}_{\mathcal{N}(S)} | \mathbf{Y}_S)$.

The maximization in Eq. (48) corresponds to the maximal distance of the average matrix $\Lambda_{av}^{(S)}$ from the convex hull of the set of matrices $\{\Lambda^{\mathbf{Y}_{\mathcal{N}(S)}}\}_{\mathbf{Y}_{\mathcal{N}(S)}}$ (hence over a polytope with extremal points being each of the matrices $\Lambda^{\mathbf{Y}_{\mathcal{N}(S)}}$). It is straightforward to see that this function is convex, due to the triangle inequality and absolute homogeneity of the L1 norm. From this fact, it follows that the maximum in Eq. (48) is attained for one of the extremal points of the convex hull, i.e., a particular $\Lambda^{\mathbf{Y}_{\mathcal{N}_i}} \Lambda^{\mathbf{Y}_{\mathcal{N}(S)}}$. Hence, combining this last bound with Eq. (47) one obtains exactly Eq. (46), which proves Proposition 1.

A.3 Statistical error bounds

The idea behind the energy estimation routine in variational algorithms such as QAOA is to estimate Hamiltonian with K -terms by estimating those terms separately and then adding them up. However, if one provides statistical error bounds for each of those terms, one needs to take into account the probabilistic nature of such bounds. Let us say that we want to estimate some distribution \mathbf{p} with N outcomes by sampling from it s number of times. It is well known [76] that the probability of estimated distribution being ϵ -close to the true distribution \mathbf{p} in terms of TVD (Eq. (9)) is vanishing exponentially in number of sample

$$\Pr(\text{TVD}(\mathbf{p}, \mathbf{p}^{\text{est}}) \geq \epsilon) \leq (2^n - 2) \exp(-2s\epsilon^2). \quad (49)$$

Now it is often convenient to set the probability fixed and consider the confidence intervals, i.e., the bound for TVD $(\mathbf{p}, \mathbf{p}^{est})$. Then we can rewrite the above equation as a function $\epsilon^*(n, s, P_{1,err}, 1)$ of three fixed parameters – number of outcomes n , number of samples s and probability of the upper bound being incorrect $P_{1,err}$. Then the basic manipulations of Eq.(49) give

$$\text{TVD}(\mathbf{p}, \mathbf{p}^{est}) \leq \epsilon^*(n, s, P_{1,err}, 1) = \sqrt{\frac{\log(2^n - 2) - \log(P_{1,err})}{2s}}. \quad (50)$$

However, since in estimation of the K-term Hamiltonian one combines the upper bounds of the form (50) for particular terms into upper bound $\epsilon^*(n, s, P_{1,err}, K)$ for the whole Hamiltonian, one needs to make sure that all upper bounds for particular terms are true at the same time with the desired probability. The union bound states that the probability of at least one event from some set occurring is no greater than the sum of probabilities of particular events. In our case, the interesting set consists of events of the type “one of the bounds of type (50) is not satisfied”. Hence the probability of at least one bound being wrong is upper bounded by

$$P_{K,err} \leq K P_{1,err}. \quad (51)$$

Therefore, if we wish to ensure that probability of all the bounds being right at the same time is fixed and equal to $P_{1,err}$ for fixed number of samples s , we need to effectively increase the upper bound to

$$\begin{aligned} \epsilon^*(n, s, P_{1,err}, K) &= \\ &= \sqrt{\frac{\log(2^n - 2) - \log\left(\frac{P_{1,err}}{K}\right)}{2s}} = \\ &= \sqrt{\frac{\log(2^n - 2) - \log(P_{1,err}) + \log(K)}{2s}} \end{aligned} \quad (52)$$

with additional term $\log K$ under the square root. In other words, we effectively lower the probability of error occurring in estimation of particular marginal distributions by a factor of K , which ensures that simultaneous estimation of K marginal distributions has the precision from Eq. (52) with probability not lower than the initial $P_{1,err}$ (by the virtue of union bound).

It is instructive to look at this from the perspective of sampling complexity. Let’s say that we would like to increase number of samples

$$s \rightarrow \tilde{s} =: s(1 + f_{oh})$$

in such a way, that upper bound remains fixed, i.e.,

$$\epsilon^*(n, \tilde{s}, P_{1,err}, K) = \epsilon^*(n, s, P_{1,err}, 1).$$

Simple calculations show that in this case we would need to (multiplicatively) increase the number of samples s by the overhead equal to

$$1 + f_{oh} = 1 + C \log(K)$$

where parameter C is equal to

$$\left(\log \left(\frac{2^n - 2}{P_{1,err}} \right) \right)^{-1}. \quad (53)$$

Hence we see that for fixed dimension (fixed number of outcomes) simultaneous estimation of K Hamiltonian terms leads to sampling overhead logarithmic in K .

Note that if one fixes the initial sampling size s , probability of error $P_{1,err}$ and number of simultaneously estimated marginals K , the sampling overhead is, perhaps counterintuitively, decaying in the number of outcomes. This dependence is roughly linear in the number of outcomes, hence it is exponential in the number of qubits.

A.4 Proofs of energy error bounds

Here we will provide proofs of worst-case error bounds for energy estimation given in the main text. Note that in the following discussion, for clarity of notation we won't add a special index to marginal probability distributions indicating that they are marginals. This should be clear from the context. On the contrary, local Hamiltonian terms, noise matrices, and correction matrices will have an additional label (α) indicating their locality.

A.4.1 Approximation errors

We will start by providing proof of Eq. (19) which gives additive bound for possible deviations of error-mitigated energy from ideal one when the source of deviations is an approximation used in constructing noise model (recall Eq. (10)). Consider estimation of the local term of Hamiltonian H_α . Recall that locality of the terms means that H_α acts non-trivially only on the subset of qubits. Now let us denote by \mathbf{p} the marginal probability distribution on the qubits belonging to the relevant subset. Then expected value of the local term can be written as

$$\langle H_\alpha \rangle = \sum_t \lambda_t^\alpha p_t = \langle \lambda^\alpha | \mathbf{p} \rangle, \quad (54)$$

where by λ^α we denoted vector of eigenvalues of local term H_α , and we used convenient bracket notation to denote scalar product. For example, if $H_\alpha = \sigma^z \otimes \sigma^z$, then $\lambda^\alpha = (1, -1, -1, 1)$. Now we want to consider two different estimators of the same local term – one from ideal distribution $\mathbf{p}^{\text{ideal}}$ and second from the error-mitigated noisy distribution $\mathbf{p}^{\text{corr}} = C_{av} \mathbf{p}^{\text{noisy}}$ (recall Proposition 1). Now we want to upper bound the difference between energy estimators based on those two marginal distributions. From Eq. (54) we can write

$$|\langle H_\alpha^{\text{corr}} \rangle - \langle H_\alpha^{\text{ideal}} \rangle| = |\langle \lambda | (|\mathbf{p}^{\text{corr}}\rangle - |\mathbf{p}^{\text{ideal}}\rangle) \rangle| \leq \underbrace{\max_t |\lambda_t|}_{\|H_\alpha\|} \underbrace{\langle \mathbb{1} | (|\mathbf{p}^{\text{corr}}\rangle - |\mathbf{p}^{\text{ideal}}\rangle) \rangle}_{2\text{TVD}(\mathbf{p}^{\text{corr}}, \mathbf{p}^{\text{ideal}})} \leq 2\|H_\alpha\|\delta_\alpha, \quad (55)$$

where δ^α is approximation error defined in Eq. (16), $\langle \mathbb{1} |$ is a vector of ones and the last inequality follows from Lemma 1. Factor two comes from the fact that Total-Variation Distance is defined with $\frac{1}{2}$ factor. The additive error bound from Eq. (19) is just a multiple application of triangle inequality

$$|\langle \mathcal{H}^{\text{corr}} \rangle - \langle \mathcal{H} \rangle| = \left| \sum_\alpha \langle H_\alpha^{\text{corr}} \rangle - \langle H_\alpha^{\text{ideal}} \rangle \right| \leq \sum_\alpha |\langle H_\alpha^{\text{corr}} \rangle - \langle H_\alpha^{\text{ideal}} \rangle| \leq 2 \sum_\alpha \delta^\alpha \|H_\alpha\|. \quad (56)$$

A.4.2 Statistical errors

Now we will proceed to prove Eq. (19) which bounds the effects of statistical noise on the error-mitigation. First, let us assume that the noise matrix acting on the marginal of interest is known exactly (not approximately as in the above derivations). In that case, in the lack of presence of the statistical errors, the correction is done exactly. However, in reality the statistics are finite, and the *estimator* of the noisy marginal distribution $\mathbf{p}^{\text{noisy}}$ from which we sample on imperfect detector can be formally written as

$$\underbrace{\mathbf{p}^{\text{noisy}}}_{\text{true}} \rightarrow \underbrace{\mathbf{p}^{\text{noisy}} + \epsilon}_{\text{estimated}}. \quad (57)$$

Now we want to bound the error in the energy estimated after acting by correction matrix Λ_α^{-1} on the estimated distribution above. In analogy to Eq. (55) we can write

$$\begin{aligned}
|H_\alpha^{\text{corr, est}} - \langle H_\alpha^{\text{ideal}} \rangle| &= |\langle \lambda | \left(\underbrace{\Lambda_\alpha^{-1} | \mathbf{p}^{\text{noisy}} \rangle}_{| \mathbf{p}^{\text{ideal}} \rangle} + \Lambda_\alpha^{-1} | \epsilon \rangle - | \mathbf{p}^{\text{ideal}} \rangle \right) | = |\langle \lambda | \Lambda_\alpha^{-1} | \epsilon \rangle| \leq \\
&\leq \|H_\alpha\| |\langle \mathbb{1} | \Lambda_\alpha^{-1} | \epsilon \rangle| \leq \|H_\alpha\| \|\Lambda_\alpha^{-1}\|_{1 \rightarrow 1} \|\epsilon\|_1 \leq \\
&\leq 2\|H_\alpha\| \|\Lambda_\alpha^{-1}\|_{1 \rightarrow 1} \epsilon^* ,
\end{aligned} \tag{58}$$

where first inequality follows from definition of operator norm, second inequality follows from a definition of any induced operator norm and the last inequality follows from analysis in previous section with ϵ^* being bound given by Eq. (52). In analogy to Eq. (56) by applying multiple times triangle inequality we obtain Eq. (19).

A.4.3 Approximation and statistical errors

Now we will combine two previous bounds by using a triangle inequality. Let us denote by $\mathbf{p}^{\text{noisy, est}}$ the estimator of noisy probability distribution $\mathbf{p}^{\text{noisy}}$ on the marginal α (hence the RHS of Eq. (57)), and by C_{av}^α the average correction matrix used to correct that marginal. We want to bound the distance between corrected estimator of noisy distribution $C_{av}^\alpha \mathbf{p}^{\text{noisy, est}}$ and the ideal distribution $\mathbf{p}^{\text{ideal}}$. For particular marginal we have triangle inequality for Total-Variation Distance

$$\begin{aligned}
\text{TVD} \left(C_{av}^\alpha \mathbf{p}^{\text{noisy, est}}, \mathbf{p}^{\text{ideal}} \right) &\leq \underbrace{\text{TVD} \left(C_{av}^\alpha \mathbf{p}^{\text{noisy}}, C_{av}^\alpha \mathbf{p}^{\text{noisy, est}} \right)}_{\leq 2\|C_{av}^\alpha\|_{1 \rightarrow 1} \epsilon^*} + \underbrace{\text{TVD} \left(C_{av}^\alpha \mathbf{p}^{\text{noisy}}, \mathbf{p}^{\text{ideal}} \right)}_{2\delta_{\alpha\text{ph}\alpha}} \\
&\leq 2\|C_{av}^\alpha\|_{1 \rightarrow 1} \epsilon^* + 2\delta_\alpha ,
\end{aligned} \tag{59}$$

where ϵ^* is statistical bound given by Eq. (52), and δ_α is approximation error defined in Eq. (16). First underlined inequality follows from properties of induced operator norm and from statistical errors bound proved in the previous section (this step is analogous to the one in the second line of Eq. (58)). The second underlined inequality follows from Proposition 1. To translate the above result to expected values of local Hamiltonian one just repeats the reasoning given in proofs of Eq. (56) and Eq. (58) for marginal distributions of the form $C_{av}^\alpha \mathbf{p}^{\text{noisy, est}}$ for which we have bounds of the form given by Eq. (59).

To finish this section, let us note that the reasoning presented in this section is analogous to the one given in Ref. [13] where analysis of effects of statistical noise and non-classical noise on the error-mitigation performed on global distributions was presented.

B Details of Diagonal Detector Overlapping Tomography

In this section, we give more details regarding our noise characterization procedure using DDOT. We start by providing efficient way of construction of DDOT circuits in Section. B.1. In Appendix B.2 we show how to use results of DDOT to infer the correlations structure of readout noise in a device. Then we discuss the construction of Diagonal Detector Overlapping Tomography circuits which are balanced (Appendix B.3) and perfect (Appendix B.4), together with proofs for scaling of required number of random circuits. In Appendix B.5, we discuss heuristic procedures of making DDOT circuits more balanced. Finally, in Appendix B.6 we explain the effect of overestimating correlations which can happen if the implemented DDOT collection is not balanced.

B.1 Construction of Diagonal Detector Overlapping Tomography circuits

Here we will present in detail our building block for efficient characterization of correlations in a measurement device: Diagonal Detector Overlapping Tomography (DDOT). In analogy to Quantum Overlapping Tomography (QOT) [33], a collection of (N, k) DDOT circuits allows one to reconstruct the noise matrices for the readout process of any group of k qubits in an N -qubit device.

More precisely, we define a (N, k) perfect collection of DDOT circuits as a collection of N -qubit quantum circuits constructed only from $\mathbb{1}$ and X gates, with the property that for every subset of qubits of the size k , each of the computational basis states on that subset is prepared at least once. For example, we will call a collection of circuits $(N, 2)$ perfect if for all pairs of qubits, each of the states from $\{|00\rangle, |01\rangle, |10\rangle, |11\rangle\}$ is prepared at least once in the whole collection.

One way to construct DDOT collection is to follow Ref. [33] and make use of the notion of hash functions (here by a hash function we mean every function $[N] \rightarrow [k]$ with $k < N$). To construct (N, k) perfect collection of DDOT circuits using hash functions one can use Algorithm 1, which encapsulates the idea of Quantum Overlapping Tomography from Ref. [33] translated to the construction of DDOT circuits. Specifically, this method corresponds to QOT with two “bases” which are preparation of state $|0\rangle$ or $|1\rangle$. Each hash function assigns each qubit a label from $[k]$. For a given function, qubits are assigned to k disjoint batches based on the value of the function value. For fixed assignment of batches, qubits belonging to a batch are initialized in the same state ($|0\rangle$ or $|1\rangle$), independently from the qubits in other batches. In this way, 2^k circuits are specified, which independently implement all computational basis states on all the batches. For example, in the case of $k = 2$ and $N = 6$ qubits, some specific hash functions could result in initializing the following states – $\{000000, 000111, 111000, 111111\}$, which implement all two-qubit computational basis states on the pairs of qubits from left and right parts of the register (in this example the two batches are $\{Q0, Q1, Q2\}$ and $\{Q3, Q4, Q5\}$). The DDOT circuits collection constructed in this way is perfect if the underlying collection of hash functions is perfect. This means that for *each* k -qubit subset in N -qubit device, there exists at least one hash function in the collection which assigns each qubit from that subset a distinct number from $[k]$. Note that if a given function is indeed injective on some k -qubit subset, this means that qubits from those subsets belong to distinct batches and therefore all computational-basis states will be implemented on them. Hence if this holds for all subsets, the collection is perfect. In Algorithm 1 we generate *random* hash functions to create a DDOT collection, therefore there is no deterministic guarantee that the collection constructed in this way will indeed be perfect. However, it follows directly from arguments in [33] (specifically, Section III) that if we use Algorithm 1 to generate DDOT collection, then if we wish the collection to be perfect with probability at least $1 - \delta$, the needed number circuits κ is at least of the order

$$\kappa > (2e)^k \left(\log(N) + \frac{1}{k} \log\left(\frac{1}{\delta}\right) \right), \quad (60)$$

see Appendix B.4 for detailed bounds.

On the other hand, one can also consider construction which uses random circuits without referring to the notion of hash functions. Generation of random bitstrings and using them as definitions of quantum circuits is used to construct DDOT collection in Algorithm 2. Similarly to Algorithm 2, since we use randomness to generate the collection, there is no deterministic guarantee that the resulting collection will be perfect. However, in Appendix B.4 we show that if we wish the collection to be perfect with probability at least $1 - \delta$, then the needed number of circuits κ is at least of the order

$$\kappa > 2^k \left(k \log(2N) + \log\left(\frac{1}{\delta}\right) \right), \quad (61)$$

which looks similar to Eq. (60), however in practice exhibits better scaling due to specific factors (see Appendix B.4 for more detailed bounds). Hence we expect that for higher system sizes the random circuits algorithm should perform better in practice compared to the one which uses hash functions. Of course, when the collection with desired properties has been constructed, it does not matter what method was used to create it.

As mentioned in the main text, during the implementation of DDOT, different circuits will in general be sampled a different number of times, this may cause some correlations to be overestimated (this effect can be reduced by proper post-processing of the data – see Appendix B.6). Hence it is beneficial for the perfect collection of DDOT circuits to be *approximately balanced*. Here by 'balanced' we mean that all k -qubit computational basis states are sampled the same number of times, and by 'approximately balanced' that they are sampled approximately the same number of times (see related notion for hash functions, for example in Ref [51]). For the construction using random circuits, we prove in Appendix B.3 that one can expect that with high probability the collection will also have this property with the required number of circuits scaling similarly to Eq. (61).

To conclude, let us point out that for a given pair of numbers (N, k) it suffices to generate the DDOT collection only once, and it can be used in the design of future experiments. We will make a number of pre-generated collections publicly available in our GitHub repository QREM (Quantum Readout Errors Mitigation) [75]. We note that the described techniques are suitable for noise characterization not only for the noise model we proposed but also for other models with bounded locality of correlations, for example with two-qubit correlations considered in Ref. [15].

B.2 Inferring the structure of clusters and neighborhoods

After the implementation of a (N, k) perfect collection of DDOT circuits, one has potential access to a lot of information about the measurement noise in a device. Here we provide a method of using data from such implementation to infer the correlations in measurement noise. As a starting point, let us note that one can use the output of the DDOT circuits to construct all the possible two-qubit noise matrices $\Lambda_{X_i X_j | Y_i Y_j}$. In particular, whenever $k \geq 2$, there are subsets of circuits that implement all computational-basis states for each two-qubit subsystem. This allows to gather all two-qubit marginal probability distributions of the form

$$\{p(X_i X_j | Y_i Y_j)\}_{i,j} \quad (62)$$

where X_i (X_j) is the measured state of i th (j th) qubit, and the Y_i (Y_j) is the input state of the corresponding qubit (i.e., the state that is supposed to be implemented by the quantum circuit from the DDOT collection). In the following Example 1 we give explicitly calculated single-qubit noise-matrices for exemplary pair of qubits $i = 1$ and $j = 2$.

Example 1. For two-qubit set $S = \{1, 2\}$, assuming no dependence on the state of neighbors, generic left-stochastic map acting on its detector has a form

$$\Lambda^S = \begin{pmatrix} p(00|00) & p(00|01) & p(00|10) & p(00|11) \\ p(01|00) & p(01|01) & p(01|10) & p(01|11) \\ p(10|00) & p(10|01) & p(10|10) & p(10|11) \\ p(11|00) & p(11|01) & p(11|10) & p(11|11) \end{pmatrix}. \quad (63)$$

It follows that single-qubit noise matrices from Eq. (8) for first qubit have form

$$\Lambda^{Y_2='0'} = \frac{1}{2} \begin{pmatrix} p(00|00) + p(01|00) & p(00|10) + p(01|10) \\ p(10|00) + p(11|00) & p(10|10) + p(11|10) \end{pmatrix}, \quad (64)$$

$$\Lambda^{Y_2='1'} = \frac{1}{2} \begin{pmatrix} p(00|01) + p(01|01) & p(00|11) + p(01|11) \\ p(10|01) + p(11|01) & p(10|11) + p(11|11) \end{pmatrix} \quad (65)$$

$$(66)$$

while for second qubit they are

$$\Lambda^{Y_1='0'} = \frac{1}{2} \begin{pmatrix} p(00|00) + p(10|00) & p(00|01) + p(10|01) \\ p(01|00) + p(11|00) & p(01|01) + p(11|01) \end{pmatrix}, \quad (67)$$

$$\Lambda^{Y_1='1'} = \frac{1}{2} \begin{pmatrix} p(00|10) + p(10|10) & p(00|11) + p(10|11) \\ p(01|10) + p(11|10) & p(01|11) + p(11|11) \end{pmatrix}. \quad (68)$$

Note that the idea is to simply fix the state of 'neighbouring' qubit to be either '0' or '1' and calculate according conditional marginals.

Now, we propose to use the information about the two-qubit noise matrices to calculate the correlations (Eq. (8)) between a given pair of qubits. For the i -th qubit, it's the dependence from the j -th qubit can be calculated by constructing two single-qubit matrices on i -th qubit – the first one with the condition that j -th qubit was initialized in $|0\rangle$ state, and the second in $|1\rangle$ state.

Those matrices can be used to calculate the parameter $c_{j \rightarrow i}$ as the norm of a difference of those matrices (Eq. (8)). We then propose to infer the structure of the readout correlations according to the magnitude of the parameters $c_{j \rightarrow i}$ as follows. One specifies threshold parameters δ_{clust} and δ_{neighb} for the level of correlations between qubits, and then assigns qubit j to the neighborhood or to the cluster of qubit i , if the parameter $c_{j \rightarrow i}$ is greater than the respective threshold. In general, we advise to set those thresholds to be higher than the likely effects of statistical deviations. We outline the whole procedure in Algorithm 3, which takes as input the conditional single-qubit noise matrices $\{\Lambda_{Q_i}^{Y_j}\}_{i \neq j}$ together with a set of thresholds, and outputs the structure of the clusters and neighborhoods in a device.

Finally, let us note that the above-described inference of correlations from two-qubit marginal distributions works under the assumption that the correlations do not vanish under taking marginals over other (than given pair) qubits. This does not need to be true in practice, and one can consider generalizations of Algorithm 3. The analogous set of single-qubit matrices depending on states of t neighboring qubits would be created in a fully analogous manner to that of Example 1 but now one would need to fix the state of t qubits. Note that if one implemented DDOT (N, k) collection, the data to create such matrices for $t = k - 1$ is available from the experiments. In the future, we intend to investigate more elaborate methods of inferring correlations structure using DDOT, and we will accordingly expand our repository [75].

Once a model for the above structure is obtained, one can use the rest of the data obtained in the DDOT procedure to reconstruct the cluster noise matrices as a function of the state of the neighbors and consequently construct a global noise model (Eq. (5)), as well as the correction matrices for the marginals (Eqs. (10),(11)). From the definition of a perfect (N, k) collection, it follows that one can reconstruct only cluster noise matrices involving a number $t = |\mathcal{C}_\chi| + |\mathcal{N}_\chi|$ of qubits which does not exceed k . If it happens that in fact $t > k$, we propose to implement one of two following solutions:

1. Neglect correlations between some qubits in order to enforce that $t = k$. For example, in Algorithm 3 if the number of qubits assigned to the neighborhood of some cluster exceeds the limit, one could decide to *not assign* to that neighborhood the qubits with the lowest values of $c_{j \rightarrow i}$ parameters. This will result in an imperfect model, which might nevertheless be accurate enough for the purposes of error mitigation.
2. Refine the noise model by performing additional experiments of standard Diagonal Detector Tomography on a chosen subset.

We note that for the refinement of the noise model, there also exists the alternative possibility of constructing a *restricted* DDOT collection that implements the characterization on a *specific* set of t -length subsets (as opposed to all such subsets). We will analyze this problem in more detail in future works, as well as during the development of our GitHub repository [75].

To finish this subsection, we note that Algorithm 3, while being straightforward and efficient, is not a flawless method. For example, it might happen that the noise on qubit i highly depends on the joint

Algorithm1 Generation of a perfect collection of (N, k) DDOT quantum circuits using random hash functions

Input:

L : number of hash functions

N : number of all qubits

k : size of the marginal

1. Start collection by creating two circuits which prepare all qubits in $|0\rangle$ state and in $|1\rangle$ state.
2. Generate L random hash functions $\{f_1, \dots, f_L\}$, i.e., random mappings $f_i : [N] \rightarrow [k]$.
3. Define k -qubit sub-register as a set of all bit-strings $\{\mathbf{X}^1, \dots, \mathbf{X}^{2^k}\}$ of length k .
4. **For** each function f_i , **do**:
For each bitstring \mathbf{X}^i in k -qubit register, **do**:
Define string \mathbf{Y} of length N in the following way:

$$(\mathbf{Y})_j = (\mathbf{X}^i)_{f_i(j)}$$

Save \mathbf{Y} as definition of one quantum circuit in the DDOT collection – each symbol '0' corresponds to identity gate, each '1' symbol corresponds to NOT gate.

5. Check if generated set of $s = 2 + L(2^k - 2)$ bitstrings $\{\mathbf{Y}\}$ contains all combinations of k -length subsets of symbols '0' and '1'.
If yes:
family $\{\mathbf{Y}\}$ defines perfect DDOT collection,
If not:
set $L \rightarrow L + 1$, generate new random hash function, add it to the collection and perform steps 2, 3 and 4.
-

Algorithm2 Generation of (N, k) perfect collection of DDOT quantum circuits using random circuits

Input:

s : number of circuits

N : number of all qubits

k : size of the marginal

1. Start collection by creating two circuits which prepare all qubits in $|0\rangle$ state and in $|1\rangle$ state.
 2. Generate s random bitstrings of size N . Each bitstring is a definition of quantum circuits – symbol '0' corresponds to identity gate, and symbol '1' symbol corresponds to NOT gate.
 3. Check if generated set of s bitstrings $\{\mathbf{Y}\}$ contains all combinations of k -length subsets of symbols '0' and '1'.
If yes:
family $\{\mathbf{Y}\}$ defines perfect DDOT collection,
If not:
set $s \rightarrow s + 1$, generate new random bitstring, add it to the collection and check new collection.
-

Algorithm3 Assignment of qubits to clusters and neighborhoods

Input: $\mathcal{S}_{\text{pairs}}$: set of all pair indices without repetitions of indices δ_{clust} : threshold of correlations for assignment to clusters δ_{neighb} : threshold of correlations for assignment to neighborhoods $\left\{ \Lambda_{X_i|Y_i}^{Y_j} \right\}_{i,j}$: collection of single-qubit noise matrices depending on state of single neighbour for all pairs.**for** (i, j) in $\mathcal{S}_{\text{pairs}}$ **do** Calculate $c_{j \rightarrow i}$ and $c_{i \rightarrow j}$ as

$$c_{j \rightarrow i} = \frac{1}{2} \|\Lambda_{Q_i}^{Y_j=0'} - \Lambda_{Q_i}^{Y_j=1'}\|_{1 \rightarrow 1}, \quad c_{i \rightarrow j} = \frac{1}{2} \|\Lambda_{Q_j}^{Y_i=0'} - \Lambda_{Q_j}^{Y_i=1'}\|_{1 \rightarrow 1},$$

 where Y_j denotes input state of qubit j (see Eq. (8)). **if** $c_{j \rightarrow i} > \delta_{\text{clust}}$ **or** $c_{i \rightarrow j} > \delta_{\text{clust}}$ **then** Assign qubits i and j to the same cluster. **else** **if** $c_{j \rightarrow i} > \delta_{\text{neighb}}$ **then** Assign qubit j to the neighborhood of qubit i . **if** $c_{i \rightarrow j} > \delta_{\text{neighb}}$ **then** Assign qubit i to the neighborhood of qubit j .

state of j and l , but this dependence is much lower if one considers qubit j and qubit l separately. In that case, Algorithm 3 might not assign the three qubits to the same cluster/neighborhood, even though they are correlated. A way to overcome the above limitation is to consider generalizations of Algorithm 3 that makes use of three-qubit parameters “ $c_{j \rightarrow i}$ ”, hence requiring a $k \geq 3$ DDOT procedure. This might be particularly useful for refining dependencies between clusters which were reported by the original Algorithm 3.

B.3 Constructing balanced DDOT collections

Interestingly, the tools utilized in the previous sections for the analysis of statistical deviations can be used to estimate the probability that a given collection of DDOT circuits will be approximately balanced. Recall that a perfect collection of (N, k) DDOT circuits is a collection of N -qubit quantum circuits consisting of $\mathbb{1}$ and X gates with a property that for each k -qubit subset every k -qubit computational basis state is implemented at least once in the whole collection. If the collection is *balanced*, it means that additionally each basis state is sampled the same number of times.

Consider the randomized construction of such collection – take circuits which are uniformly random combinations of 0 and 1 symbols (‘0’ corresponding to the identity gate and ‘1’ to the NOT gate). This can be constructed efficiently since it suffices to choose each of the N bits at random independently. Such construction can be viewed as sampling from a 2^N -dimensional uniform distribution (corresponding to all possible N -bit strings describing possible circuits). The k -bit marginal distributions obtained from this distribution correspond to the distribution of local circuits on k -qubit subsets (i.e., the noise characterization by implementing computational-basis states on k -qubit subsets). Having this perspective, we can formulate the problem of approximate balancing of DDOT family in the following way: What is the probability that, when sampling from the global 2^N -dimensional uniform distribution, all of the k -bit marginals (corresponding to the subsets of interest in DDOT) will be at most ϵ -distant from the uniform distributions (on 2^k -dimensional space)?

Now we can use tools of statistical analysis presented in the previous sections. Our single marginal distribution has 2^k outcomes, hence from Eq. (49) we get that the probability of a single marginal

being *at least* ϵ -distant in TVD is bounded by

$$P_1 \leq 2^{2^k} \exp(-2s\epsilon^2) , \quad (69)$$

where s is here the number of random circuits in the collection (viewed as samples from the global uniform distribution of bitstrings). Since there is $\binom{N}{k}$ number of k -bit marginals, applying the union bound (as in the derivations in the previous sections) gives the upper bound for the probability $P_{N,k}$ that at least one of the $\binom{N}{k}$ marginals is at least ϵ -distant from uniform distributions as

$$P_{N,k} \leq \binom{N}{k} 2^{2^k} \exp(-2s\epsilon^2) . \quad (70)$$

Hence with probability *at least* $1 - \binom{N}{k} 2^{2^k} \exp(-2s\epsilon^2)$ all marginals are *at most* ϵ -distant from uniform distribution. Note that $\epsilon = 0$ corresponds to a perfectly balanced family of circuits, and small but non-zero ϵ will correspond to the approximately balanced family.

Let us now choose some parameter δ as an upper bound for probability Eq. (70) and find the bound for the required number of random circuits s (viewed as samples from a global uniform distribution) which are needed to obtain a family for which each marginal is distant from the uniform distribution by at most ϵ . After basic manipulations of Eq. (70) we obtain that

$$s \geq \frac{2^k \log 2 + \log \binom{N}{k} + \log \left(\frac{1}{\delta}\right)}{2\epsilon^2} \approx \frac{2^k + k \log N + \log \frac{1}{\delta}}{2\epsilon^2} . \quad (71)$$

B.4 The efficiency of the random construction of DDOT collection

B.4.1 Random circuits

Adopting the perspective from the previous section, we can in a simple manner tackle the problem of bounding the required number of random circuits which are needed to obtain a perfect collection of $\binom{N}{k}$ DDOT circuits. Consider randomly sampling s number of bit-string of length N . Now for a fixed k -element subset, the probability of a particular k -element combination *not appearing* is $1 - \frac{1}{2^k}$. Hence after s samples, there is $\left(1 - \frac{1}{2^k}\right)^s$ probability that this particular combination did not appear. Since there are $2^k \binom{N}{k}$ combinations of interest (i.e., 2^k small k -length bitstrings for all $\binom{N}{k}$ subsets), we can use the union bound to obtain

$$2^k \binom{N}{k} \left(1 - \frac{1}{2^k}\right)^s \approx 2^k \binom{N}{k} \exp\left(-\frac{s}{2^k}\right) \quad (72)$$

as the upper bound for the probability that *at least* one k -length bit-string did not appear after s samples. This means that with probability of at least $1 - 2^k \binom{N}{k} \exp\left(-\frac{s}{2^k}\right)$, all of the k -length bitstrings appeared.

Let us now choose some parameter δ as the upper bound for Eq. (72) and calculate the resulting bound for the number of samples (i.e., random circuits). After basic manipulations we obtain

$$s > 2^k \left(k \log 2 + \log \binom{N}{k} + \log \frac{1}{\delta} \right) \approx 2^k \left(k \log 2N + \log \frac{1}{\delta} \right) . \quad (73)$$

B.4.2 Random hash functions

Using simple combinatorial arguments, in Ref. [33] it is shown that the probability of the collection of $\binom{N}{k}$ *random hash functions* not being perfect (for our purposes perfect hash function collection means that it can be used to construct a perfect DDOT collection) is upper bounded by

$$\binom{N}{k} \left(1 - \frac{k!}{k^k}\right)^L , \quad (74)$$

where L is the number of generated random hash functions. Then the authors simplify the use of the above fact to derive a bound on the needed number of hash functions for the collection to be perfect. Let us repeat the derivation from Ref. [33] with paying attention to specific factors. From Algorithm 1 it follows that the number of circuits s obtained from a given hash function collection is equal to $s = 2 + (2^k - 2)L$. Now by choosing the parameter δ as upper bound on Eq. (74), we obtain that required number of circuits is lower bounded by

$$\begin{aligned}
s &\geq 2 + (2^k - 2) \frac{-\log \binom{N}{k} + \log \delta}{\log(1 - \sqrt{2\pi k} e^{-k})} \\
&\geq 2 + (2^k - 2) \frac{-\log \binom{N}{k} + \log \delta}{-\sqrt{2\pi k} e^{-k} - 2\pi k e^{-2k}} \\
&\geq 2 + (2^k - 2) \frac{-\log \binom{N}{k} + \log \delta}{-\sqrt{2\pi} k e^{-k}} \\
&\approx 2^k e^k \left(\frac{1}{k} \log \frac{1}{\delta} + \log N \right), \tag{75}
\end{aligned}$$

where we used that $\log(1-x) \geq -x-x^2$. Let us note that Eq. (75) is a reiterated result from Ref. [33].

In above derivations we assumed that $(-\log \binom{N}{k} + \log \delta) < 0$, then we utilized the fact that

$$\sqrt{2\pi} k^{k+\frac{1}{2}} e^{-k} \leq k! \leq e k^{k+\frac{1}{2}} e^{-k}, \tag{76}$$

and we used approximation $\binom{N}{k} \approx N^k$.

B.5 Heuristic balancing of DDOT collection

After generating a perfect DDOT collection, one can be interested in making it more balanced. There are various possible figures of merit that can be used to quantify the “balancing” of the family. For example, in previous sections, we used TVD between a uniform distribution and generated sample, when viewing obtained circuits in the collection as samples from the uniform distribution. Another possibility is to calculate a number of appearances of each marginal term in the whole collection (for example, for 2-qubit subsets the number that each of 00, 01, 10, and 11 appeared, for every 2-qubit subset), and then take an empirical standard deviation $\sigma_{n,k}$ of this quantity. A perfectly balanced family would have 0 standard deviation defined in that way.

Now we discuss a simple heuristic method to improve balancing. The starting point is a perfect (N, k) DDOT collection. Now we apply the following steps in a loop.

1. Calculate a number of appearances of all k -qubit terms in the collection.
2. Find the set of non-overlapping k -bit marginals which appear in the collection the least number of times compared to the whole population. If k does not divide N , choose $\lfloor \frac{N}{k} \rfloor$ subsets and add random gates to the remaining bits.
3. Add circuit which implements those least-appearing marginals to the collection.

For example, say that in the last step of the above procedure for $k = 2$ and $N = 5$ we added circuit

$$01100. \tag{77}$$

This might mean that we found that the least appearing marginal is state 01 on qubits Q_0 and Q_1 , the marginal 10 on qubits Q_2 and Q_3 appeared the same or second-least number of times in the whole collection, while the 0 on last qubit Q_4 was random. In this way, we are adding missing marginal terms “by hand” at each step of the loop. Clearly, this procedure is heuristic and it is not guaranteed to succeed since by adding certain circuits that implement desired marginals we also implement other marginals

(in the example above, we add, e.g., marginal 00 on qubits $Q0$ and $Q3$). If it happens that those other marginals are in the opposite “tail” of the whole distribution (i.e., they are the most abundant ones), then the collection will not become more balanced (it can actually become less balanced). However, on average it is more likely that by doing so we add marginals that are closer to the “average marginals” (i.e., those which appear a number of times close to the mean appearance number).

Clearly, a lot of practical refinements of the described method are possible – for example, in the second step of the method, instead of adding the least appearing marginals one can focus on maximizing the number of low-appearance marginals. In practice, we found that the described method without modifications usually reduced $\sigma_{N,k}$ with a growing number of circuits added in the loop.

To conclude, let us note that the special case of the above technique can be also used to add circuits to the non-perfect collection of DDOT circuits in order to make it closer to being perfect. Namely, if the collection is not perfect, this means that it does not implement some computational-basis states on some k -qubit subsets. Hence in that case the above procedure would report “the least appearing marginals” to be those which are missing in the collection and it would keep adding them in a loop until there are no missing terms – a strategy which might turn out to be better than adding random circuits if the number of missing terms is not too big.

B.6 Overestimation of correlations

In the main text, we explained that when considering the implementation of DDOT circuits, some two-qubit correlations might be overestimated due to the fact the collection of circuits is not balanced. To understand this effect, let us now consider the following illustrative (but rather unrealistic) example.

B.6.1 Explanation of the effect and post-processing strategy

Say that we want marginal probability distribution on qubits $Q0$ and $Q2$ when input state was $|00\rangle$. Then if noise is *uncorrelated* it does not matter whether global input state on three qubits was $|000\rangle$ or $|010\rangle$ ($|001\rangle$ or $|011\rangle$). However, when there are correlations, it might happen that those two distributions *will be different* depending on the input state of $Q1$. When we marginalize over $Q1$ we forgot about “where the data came from”. Normally, we would just add the marginal probability from circuits implementing both global states and then normalized it. But if global state $|000\rangle$ was implemented a different number of times than $|010\rangle$, it can cause that effectively correlations which are caused by $Q1$ are wrongly identified as correlations between $Q0$ and $Q2$, because some global probability distributions contribute to the marginal with higher weights (i.e., are effectively counted more times when calculating marginal distributions). As mentioned in the main text, the natural way to reduce such effects is to create a collection that is balanced, hence it samples from all two-qubit states the same number of times. The other thing one can do is to perform post-processing of the experimental data in such a way, that all contributions to the given marginal are weighted by the inverse of a number of times they were implemented. Note that they come from different global distributions, so what is important is the number of times the given marginal state was implemented together with some specific state on all the other qubits. Importantly, this method is not perfect because for big systems the “state on all the other qubits” will likely be different each time anyway (this is due to the fact that collection of DDOT circuits will be random, hence it becomes quite unlikely to obtain two times the same bitstring if the number of qubits is high). Another thing one can do is to change the weights of the given contribution to the marginal distribution depending on the state of some particular subset of qubits in order to assess whether inferred correlations were correct. Specifically, one might perform a recursive procedure in which the structure of clusters and neighborhoods inferred from non-post-processed data is validated on particular subsets using this type of post-processing. There are certainly a lot of practical possibilities to improve the post-processing scheme and we intend to investigate them in future research.

B.6.2 Illustration of the effect

To have some idea how big a described effect can be, let us consider implementation of the following collection of circuits on three qubits $\{Q_0, Q_1, Q_2\}$

$$\{000, 001, 001, 010, 011, 100, 101, 110, 111\} , \quad (78)$$

where each 0 corresponds to implementation of identity gate and each 1 to the NOT gate. Note that considered collection is clearly abundant for 3-qubit characterization since it contains more than $2^3 = 8$ circuits. One can think of the above circuits as a part of a collection of DDOT circuits on a higher number of qubits, while we only look at a specific triple of qubits. Now let us assume that the qubits Q_0 and Q_2 are completely *uncorrelated* in terms of readout noise. On the other hand, Q_1 and Q_2 are highly correlated in a following way

$$Q_1 = |0\rangle \implies \text{do nothing} , \quad (79)$$

$$Q_1 = |1\rangle \implies \text{apply bitflip to } Q_2 . \quad (80)$$

Now if one takes outcomes from circuits from the collection, add them up, normalize and consider resulting estimators of probabilities of obtaining different outcomes after implementation of collection from Eq. (78) *marginalized* over qubit 1, it follows from direct computation that the noise matrix on Q_0 and Q_2 is of the form

$$\Lambda^{Q_0 Q_2} = \begin{pmatrix} p(00|00) & p(00|01) & p(00|10) & p(00|11) \\ p(01|00) & p(01|01) & p(01|10) & p(01|11) \\ p(10|00) & p(10|01) & p(10|10) & p(10|11) \\ p(11|00) & p(11|01) & p(11|10) & p(11|11) \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{3} & 0 & 0 \\ \frac{1}{2} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{3} & \frac{1}{2} \end{pmatrix} , \quad (81)$$

from which we obtain that the noise matrices on Q_0 depending on state of Q_2 are

$$\Lambda^{Y_2='0'} = \begin{pmatrix} p(00|00) + p(01|00) & p(00|10) + p(01|10) \\ p(10|00) + p(11|00) & p(10|10) + p(11|10) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad (82)$$

$$\Lambda^{Y_2='1'} = \begin{pmatrix} p(00|01) + p(01|01) & p(00|11) + p(01|11) \\ p(10|01) + p(11|01) & p(10|11) + p(11|11) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad (83)$$

$$(84)$$

and on Q_2 depending on state of Q_0

$$\Lambda^{Y_0='0'} = \begin{pmatrix} p(00|00) + p(10|00) & p(00|01) + p(10|01) \\ p(01|00) + p(11|00) & p(01|01) + p(11|01) \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \quad (85)$$

$$\Lambda^{Y_0='1'} = \begin{pmatrix} p(00|10) + p(10|10) & p(00|11) + p(10|11) \\ p(01|10) + p(11|10) & p(01|11) + p(11|11) \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{1}{2} \\ \frac{2}{3} & \frac{1}{2} \end{pmatrix} . \quad (86)$$

The above matrices give correlation factors (Eq. (8))

$$c_{2 \rightarrow 0} = 0 \quad (87)$$

$$c_{0 \rightarrow 2} = \frac{1}{6} . \quad (88)$$

Hence the Algorithm 3 would report that the noise on Q_2 significantly depends on the state of Q_0 even though physically there is no dependence. The correlations between Q_1 and Q_2 give rise to false correlations between Q_0 and Q_2 after marginalizing over Q_1 . This is solely due to the fact that in Eq. (78) different three-qubit states are sampled different numbers of times, which gives some contributions to the marginals have higher weight when constructing effective noise matrix on Q_0 and Q_2 .

C Noise characterization scheme overview

Here we provide a step-by-step description of our noise characterization procedure. We note that stages 0 and 1 were only briefly mentioned in the main text, and they correspond to the verification of the undertaken assumptions: stage 0 verifies the quality of the single-qubit gates, and stage 1 the assumption about classical nature of the noise in the measurement device. Stages 2 and 3 describe the proper characterization scheme of Diagonal Detector Overlapping Tomography which was discussed in detail in the main part of the work and in Appendix B.

C.1 Stage 0 – single-qubit gate fidelities

To begin, let us note that our characterization procedure relies on the assumption of perfect state preparation. However, in practice, this assumption might be significantly violated. Therefore we propose not to use qubits with single-qubit gate infidelities above some threshold – in our experiments we arbitrarily chose this threshold to be 0.01.

In experiments on IBM’s *Melbourne* backend, the single-qubit gate fidelities were good enough (fidelities above 99%) to use all of the qubits. In the case of Rigetti’s *Aspen-8*, we discarded 8 qubits which had fidelities below 98%, while still using qubits 5 qubits which had fidelity in range [98%, 99%].

C.2 Stage 1 – assessing classical form of the noise

In order to perform simultaneous estimation of single-qubit detectors with an overcomplete operator basis, one needs to implement 6 different circuits – each implementing eigenstate of a different Pauli matrix on every qubit at the same time. After this implementation, one needs to post-process data to obtain marginal single-qubit distributions and use standard detector tomography algorithms, for example, those described in [44] (and implemented in Python in online repository [75]).

Having reconstructed POVMs describing each single-qubit detector in a device, one can assess the classicality of the noise using methods described in Ref. [13] – for the sake of completeness, we will recall the main notions of that procedure here. First, we will need a notion of distance between quantum measurements. Such distances are usually related to the probability distributions that those measurements generate via Born’s rule. Recall from Eq. (9) that the Total Variation Distance between probability distributions \mathbf{p} and \mathbf{q} as L1 norm of difference of those vectors

$$D_{\text{TV}}(\mathbf{p}, \mathbf{q}) = \frac{1}{2} \|\mathbf{p} - \mathbf{q}\|_1 = \frac{1}{2} \sum_i |p_i - q_i|. \quad (89)$$

Now, the distance between quantum measurements related to TVD is the operational distance D_{op} defined for two POVMs \mathbf{M} and \mathbf{P} as [77]

$$D_{\text{op}}(\mathbf{M}, \mathbf{P}) = \max D_{\text{TV}}(\mathbf{p}_{\mathbf{M}}, \mathbf{q}_{\mathbf{P}}), \quad (90)$$

where $\mathbf{p}_{\mathbf{M}}$ (or $\mathbf{q}_{\mathbf{P}}$) is a probability distribution generated by measurement \mathbf{M} (or \mathbf{P}) via Born’s rule, and the maximization goes over all quantum states. Hence, the operational distance between two quantum measurements is the worst-case distance between probability distributions they can generate. Now, to quantify readout noise one can simply calculate² operational distance for reconstructed POVM \mathbf{M} for each qubit and the ideal measurement \mathbf{P} . Then, to quantify coherent part of the noise, one can make the following decomposition [13]

$$\mathbf{M} = \underbrace{\Delta \mathbf{P}}_{\text{classical part}} + \underbrace{\Delta}_{\text{coherent part}}. \quad (91)$$

For the ideal measurement \mathbf{P} being the standard measurement in the computational basis, this decomposition is straightforward – the classical part of the noise is contained in the diagonal part of the

²See Refs. [13, 77] on practical calculation of RHS of Eq. (90).

measurement operators, while off-diagonal terms are a coherent part. The magnitude of the coherent part of the noise can be quantified as $D_{\text{op}}(\mathbf{M}, \Lambda\mathbf{P})$. The assumption of fully classical noise leads to discarding the coherent part Δ in the Eq. (91) and performing error-mitigation *as if* POVM $\Lambda\mathbf{P}$ was exact description of the detector. This leads to the propagation of coherent errors under error-mitigation, and it can be quantified via $\|\Lambda^{-1}\|_{1 \rightarrow 1} D_{\text{op}}(\mathbf{M}, \Lambda\mathbf{P})$. Following guidelines in Ref. [13], we propose to *discard* every qubit which fulfills the following inequality

$$\|\Lambda^{-1}\|_{1 \rightarrow 1} D_{\text{op}}(\mathbf{M}, \Lambda\mathbf{P}) \geq D_{\text{op}}(\mathbf{M}, \mathbf{P}) . \quad (92)$$

Fortunately, in experiments on both IBM's and Rigetti's machines, this step did not lead us to discard any qubits – the noise in those devices remains highly classical, as indicated in previous experiments [13, 14].

Before proceeding further, let us note that while assessing classicality of the noise via single-qubit QDTs we make the following implicit assumption – the coherent part of the noise does not scale significantly with growing system size (by this we mean that it is at most additive in the number of qubits). Furthermore, in using the rule Eq. (92) we disregard effects of statistical deviations.

C.3 Stage 2 – Diagonal Detector Overlapping Tomography

The main idea of DDOT was described in the main text. The practical generation of relevant circuits was described in Appendix B.1 and summarized in Algorithm 1 (using random hash functions) and in Algorithm 2 (using random circuits).

C.4 Stage 3 – inferring correlations structure

The procedure of inferring correlations (i.e., structure of clusters and neighbourhoods) was described in detail in Appendix B.2 and summarized in Algorithm 3.

D Sample complexity of energy estimation

In this section, we give some derivations related to the estimation of expectation values of local Hamiltonian terms on various quantum states. We start by discussing correlations in random states in Appendix D.1. Then in Appendix D.2, we prove Proposition 2 from the main text, which concerns states appearing in the QAOA algorithms. Finally, in Appendix D.3 we analyze what happens with the covariances of local Hamiltonian terms if the uncorrelated measurement noise and its mitigation are present.

D.1 Local correlations in random states

Proposition 4. *Let $|\psi\rangle$ be a pure state on $(\mathbb{C}^2)^{\otimes N}$, for a subset $\gamma \subset [N]$ of qubits we denote by ρ_γ and id_γ the marginal of $|\psi\rangle\langle\psi|$ corresponding to γ and the maximally mixed state on γ , respectively. Let H_α, H_β be local hamiltonians term that act on disjoint subsets of qubits. We then have*

$$\text{Cov}(H_A, H_B) \leq 3\|H_A\|\|H_B\|\|\rho_{\alpha\cup\beta} - \text{id}_{\alpha\cup\beta}\|_1. \quad (93)$$

Proof. A simple algebra gives

$$\begin{aligned} \text{Cov}(H_\alpha, H_\beta) &= \text{tr}(H_\alpha \otimes H_\beta(\rho_{\alpha\cup\beta} - \rho_\alpha \otimes \rho_\beta)) \\ &= \text{tr}(H_\alpha \otimes H_\beta(\Delta_{\alpha\cup\beta} - \Delta_\alpha \otimes \rho_\beta + \text{id}_\alpha \otimes \Delta_\beta)) , \end{aligned} \quad (94)$$

where $\Delta_\gamma = \rho_\gamma - \text{id}_\gamma$. We now apply the well-known inequality

$$\text{tr}(X) \leq \|A\|\|X\|_1 , \quad (95)$$

for $A = H_\alpha \otimes H_\beta$ and $X = \Delta_{\alpha\cup\beta} - \Delta_\alpha \otimes \rho_\beta - \text{id}_\alpha \otimes \Delta_\beta$. The 1-norm can be upper bounded as follows:

$$\|\Delta_{\alpha\cup\beta} - \Delta_\alpha \otimes \rho_\beta + \text{id}_\alpha \otimes \Delta_\beta\|_1 \leq \|\Delta_{\alpha\cup\beta}\|_1 + \|\Delta_\alpha\|_1 + \|\Delta_\beta\|_1 \leq 3\|\Delta_{\alpha\cup\beta}\|_1, \quad (96)$$

where we used the following properties of 1-norm: triangle inequality, multiplicativity: $\|A \otimes B\|_1 = \|A\|_1 \|B\|_1$, and data-processing inequality [78]. We conclude the proof by inserting (96) into (95) and using $\|H_\alpha \otimes H_\beta\| = \|H_\alpha\| \|H_\beta\|$. \square

D.2 Proof of Proposition 2

Consider Hamiltonian with connectivity given by Erdős-Rényi random graph with N nodes and K edges with average degree equal to $q = \frac{K}{N}$. We will show that if the number of levels satisfies $p = c \log(N) + 1$ with $c \leq \frac{1}{2 \log(2q/\ln 2)}$, then the variance of the energy $\text{Var}(\mathcal{H})$ will scale as $\mathcal{O}(N^{2-x})$, with $x > 0$ depending on c .

Let $G = (V, E)$ be a graph (with vertex set V and edge set E), we denote by $B(i, r)$ the set of vertices that are in graph distance r or less from vertex i . Similarly, for an edge $\alpha = (i, j) \in E$, we define $C(\alpha, r)$ the set of vertices that are in graph distance r or less away from α . For $\alpha = (i, j)$ and $\beta = (v, w)$, we have that if $v \notin B(i, r+2)$, then $\beta \notin C(\alpha, r)$.

The random Hamiltonian corresponding to a graph can be written as sum of 2-qubit terms (with also single qubit terms incorporated into to these) corresponding to the edges E of the random graph $G(N, q)$, i.e., $H = \sum_{(i,j) \in E} H_{(i,j)}$. The variance is then

$$\text{Var}(\mathcal{H}) = \sum_{\alpha, \beta \in E} \text{Cov}(H_\alpha, H_\beta). \quad (97)$$

To bound this quantity we will utilize the following two facts. First, we notice that than any non-zero term in Eq. (97) can be bounded by

$$\text{Cov}(H_\alpha, H_\beta) \leq \text{Var}(H_\alpha) \text{Var}(H_\beta) \leq \|H_\alpha\| \|H_\beta\|, \quad (98)$$

where we used known covariance-variance inequality together with Popoviciu's inequality.

Second, we reiterate an important observation from Ref. [54] about these types of QAOAs: Consider two operators O_1 and O_2 acting non-trivially only on the sets of qubits $A_1 \subset V$ and $A_2 \subset V$, respectively. If U is a unitary corresponding to a p -level QAOA (with any parameter setting), the set of nodes A_1 and A_2 is in graph distance at least $2p$ distance from each other and $|\psi\rangle$ is product state, then

$$\langle \psi | U^\dagger O_1 O_2 U | \psi \rangle = \langle \psi | U^\dagger O_1 U | \psi \rangle \langle \psi | U^\dagger O_2 U | \psi \rangle. \quad (99)$$

With these two ingredients, we can bound the variance as

$$\begin{aligned} \text{Var}(\mathcal{H}) &= \sum_{\alpha, \beta \in E} \text{Cov}(H_\alpha, H_\beta) \\ &= \sum_{\alpha \in E} \sum_{\beta \in C(\alpha, 2p)} \text{Cov}(H_\alpha, H_\beta) \\ &\leq \sum_{\alpha \in E} \sum_{\beta \in C(\alpha, 2p)} \|H_\alpha\| \|H_\beta\|. \end{aligned} \quad (100)$$

An even more rough upper bound can be given by

$$\text{Var}(\mathcal{H}) \leq f_{\mathcal{H}} \sum_{\alpha \in E} \max_{\alpha \in E} |C(\alpha, 2p)| = f_{\mathcal{H}} K \max_{\alpha \in E} |C(\alpha, 2p)|, \quad (101)$$

with setting

$$f_{\mathcal{H}} = \max_{\alpha, \beta} \|H_\alpha\| \|H_\beta\| \quad (102)$$

Now we can turn to the concrete case of Hamiltonians corresponding to graph sampled from the Erdős-Rényi graphs with average degree q . For such graphs the *Neighborhood Size Theorem* states the following [54]:

For any

$$r < \frac{w \log(N)}{4 \log(2q/\ln(2))}, \quad (103)$$

where $0 < w > 1$, there exists a constants $a > 0$ and $A < 1$ such that

$$\text{Prob} \left[\max_{i \in V} |\mathcal{B}(i, r)| \geq N^{A/2} \right] \leq e^{-N^{a/2}}. \quad (104)$$

To be more specific, we can give the expression of a and A in terms of r and N :

$$A = w \frac{(2 + |\log_{2q}(\ln(2))|)}{(1 + |\log_{2q}(\ln(2))|)}, \quad (105)$$

$$a = \frac{w}{3(1 + |\log_{2q}(\ln(2))|)}. \quad (106)$$

The proof of the above comes from the proof of Neighborhood Size Theorem given in [54], page 13. Specifically our inequalities correspond to setting $s = 1$.

Now note that since for any $\alpha = (i, j) \in E$ and $\beta = (v, w) \in E$, we have that if $\beta \in \mathcal{C}(\alpha, r)$ then $v \in \mathcal{B}(i, r+2)$. This immediately implies that generally for $\alpha = (i, j) \in E$ we have $|\mathcal{C}(\alpha, r)| \leq |\mathcal{B}(i, r+2)|^2$, and thus also

$$\text{Prob} \left[\max_{\alpha \in E} |\mathcal{C}(\alpha, r-2)| \geq N^A \right] \leq \text{Prob} \left[\max_{\alpha \in E} |\mathcal{B}(\alpha, r)|^2 \geq N^A \right] = \text{Prob} \left[\max_{\alpha \in E} |\mathcal{B}(\alpha, r)| \geq N^{A/2} \right] \leq e^{-N^{a/2}}. \quad (107)$$

If we now choose $2p = r - 2$, and thus

$$p < \frac{w \log(N)}{8 \log(2q/\ln(2))} - 1, \quad (108)$$

then by combining Eq. (107) with bound in Eq. (101), we get that with probability at least $1 - e^{-N^{a/2}}$ variance of the Hamiltonian is bounded by

$$\text{Var}(\mathcal{H}) \leq f_{\mathcal{H}} K \max_{\alpha \in E} |\mathcal{C}(\alpha, r)| \leq f_{\mathcal{H}} K N^A = f_{\mathcal{H}} q N^{A+1} \quad (109)$$

which is statement of Proposition 2.

D.3 Covariances of local terms in presence of uncorrelated readout noise

D.3.1 Presence of uncorrelated readout noise

As a starting point, let us assume that we have a state with correlations bounded in trace norm

$$\|\rho_{\alpha\beta} - \rho_{\alpha} \otimes \rho_{\beta}\|_1 \leq \epsilon_{\alpha\beta}, \quad (110)$$

where $\rho_{\alpha\beta}$ is marginal quantum state on subsystems α and β which is close to a product state of those subsystems. The special instance $\epsilon = 0$ corresponds to the product state case. We are interested in what happens with the covariances of local Hamiltonian terms if the measurement is affected by uncorrelated classical noise of the form $\Lambda = \otimes_i \Lambda_{Q_i}$, where Λ_{Q_i} is noise matrix acting on qubit Q_i (see Eq. (16)). Similarly, as in the main text, the Hamiltonian we consider is classical (i.e., diagonal), and local terms can be decomposed into sums of products of Pauli σ_z terms. Let us denote by $\mathbf{p}_{\alpha}^{\text{noisy}} = \Lambda_{\alpha} \mathbf{p}_{\alpha}^{\text{ideal}}$ a noisy marginal distribution on the qubit subset α corresponding to local Hamiltonian term H_{α} . Note that

in the uncorrelated noise model $\Lambda_\alpha = \bigotimes_{Q_i \in \alpha} \Lambda_{Q_i}$. Similarly to Eq. (54), we can write the expectation value of the noisy local Hamiltonian term as

$$\langle \tilde{H}_\alpha \rangle_{\rho_\alpha} := \langle \lambda_\alpha | \mathbf{P}_\alpha^{\text{noisy}} \rangle = \langle \lambda_\alpha | \Lambda_\alpha \mathbf{P}_\alpha^{\text{ideal}} \rangle, \quad (111)$$

where λ^α is a vectorized spectrum of term H_α , the subscript ρ_α indicates the marginal state on which the expectation value is calculated and we used a convenient bracket notation to denote scalar product.

We are now interested in bounding the covariance $\text{Cov}(H_\alpha^{\text{noisy}}, H_\beta^{\text{noisy}})$ between two local Hamiltonian terms when measured in a quantum state from Eq. (110). Let us decompose a marginal quantum state as

$$\begin{aligned} \rho_{\alpha\beta} &= \rho_\alpha \otimes \rho_\beta + \Delta_{\alpha\beta} \\ \Delta_{\alpha\beta} &:= \rho_{\alpha\beta} - \rho_\alpha \otimes \rho_\beta. \end{aligned} \quad (112)$$

Now can write

$$\begin{aligned} \text{Cov}(\tilde{H}_\alpha, \tilde{H}_\beta) &= \langle \tilde{H}_{\alpha\beta} \rangle_{\rho_{\alpha\beta}} - \langle \tilde{H}_\alpha \rangle_{\rho_\alpha} \langle \tilde{H}_\beta \rangle_{\rho_\beta} = \\ &= \langle \tilde{H}_{\alpha\beta} \rangle_{\Delta_{\alpha\beta}} + \langle \tilde{H}_{\alpha\beta} \rangle_{\rho_\alpha \otimes \rho_\beta} - \langle \tilde{H}_\alpha \rangle_{\rho_\alpha} \langle \tilde{H}_\beta \rangle_{\rho_\beta} = \\ &= \underbrace{\langle \lambda_{\alpha\beta} | \Lambda_{\alpha\beta} \mathbf{P}_{\Delta_{\alpha\beta}}^{\text{ideal}} \rangle}_{\text{correlated part}} + \underbrace{\langle \lambda_{\alpha\beta} | \Lambda_{\alpha \otimes \beta} \mathbf{P}_{\alpha \otimes \beta}^{\text{ideal}} \rangle - \langle \lambda_\alpha | \Lambda_\alpha \mathbf{P}_\alpha^{\text{ideal}} \rangle \langle \lambda_\beta | \Lambda_\beta \mathbf{P}_\beta^{\text{ideal}} \rangle}_{\text{uncorrelated part}}. \end{aligned} \quad (113)$$

In the above we slightly abused the notation – in general, α and β can overlap, in which case one needs to insert proper identities and accordingly redefine product state $\rho_\alpha \otimes \rho_\beta$ (together with bound in Eq. (110) which will now correspond to the different, more refined division of qubits) and corresponding noise matrices. We also denoted by $\mathbf{P}_{\Delta_{\alpha\beta}}^{\text{ideal}}$ a formal vector given by diagonal elements of $\Delta_{\alpha\beta}$ (this corresponds to measurement in computational basis). Now note that in the last line of Eq. (113) the part underlined as “uncorrelated part” contains terms without any correlations except those which can appear if α and β overlap. In particular, if $\alpha \cap \beta = \emptyset$, then since the noise is uncorrelated, we have $\langle \lambda_{\alpha\beta} | = \langle \lambda_\alpha | \otimes \langle \lambda_\beta |$ and $\Lambda_{\alpha \otimes \beta} \mathbf{P}_{\alpha \otimes \beta}^{\text{ideal}} = (\Lambda_\alpha \mathbf{P}_\alpha^{\text{ideal}}) \otimes (\Lambda_\beta \mathbf{P}_\beta^{\text{ideal}})$, therefore then this part is equal 0. Otherwise, it gives a non-zero contribution, which however would be present even without any measurement noise.

The only part which adds non-trivial correlations is therefore $\langle \lambda_{\alpha\beta} | \Lambda_{\alpha\beta} \mathbf{P}_{\Delta_{\alpha\beta}}^{\text{ideal}} \rangle$ which using Eq. (110) and elementary transformations can be bounded as

$$|\langle \lambda_{\alpha\beta} | \Lambda_{\alpha\beta} \mathbf{P}_{\Delta_{\alpha\beta}}^{\text{ideal}} \rangle| \leq \|H_\alpha H_\beta\| \|\Lambda_{\alpha\beta}\|_{1 \rightarrow 1} \|\Delta_{\alpha\beta}\|_1 \leq \|H_\alpha H_\beta\| \epsilon_{\alpha\beta}. \quad (114)$$

In the last inequality we made use of the following facts. If α and β do not overlap, we have $\|\Lambda_{\alpha\beta}\|_{1 \rightarrow 1} = \|\Lambda_\alpha \otimes \Lambda_\beta\|_{1 \rightarrow 1} = \|\Lambda_\alpha\|_{1 \rightarrow 1} \|\Lambda_\beta\|_{1 \rightarrow 1} = 1$. In case α and β overlap, the non-overlapping parts will simply give stochastic matrices (and they have $1 \rightarrow 1$ norm equal to 1), while overlapping parts will be squared. We therefore obtain $\|\Lambda_{\alpha\beta}\|_{1 \rightarrow 1} = \prod_{i \in \alpha \cap \beta} \|\Lambda_i^2\|_{1 \rightarrow 1} \leq \prod_{i \in \alpha \cap \beta} \|\Lambda_i\|_{1 \rightarrow 1}^2 = 1$, where we used submultiplicity of norm. Of course, if the noise model is known, the $\|\Lambda_{\alpha\beta}\|_{1 \rightarrow 1}$ could be also calculated explicitly.

From the above, it follows that if $\epsilon_{\alpha\beta}$ is small, then under uncorrelated measurement noise the covariances between local terms will be, unsurprisingly, small as well.

D.3.2 Effects of error-mitigation

To include error-mitigation on the above considerations let us note that since error-mitigation is operation performed classically, it can be incorporated into the spectrum of the Hamiltonian by defining spectrum $\langle \lambda_\alpha A_\alpha^\dagger |$ of “error-mitigated Hamiltonian” term H_α^{corr} which energy is estimated on the noisy

parameter	α	γ	A	a	c
starting value	0.602	0.101	200	0.06	0.12

Table 1: Starting values of hyperparameters used in optimization. The meaning of the parameters is in agreement with standard conventions (see for example Refs. [59, 60]).

probability distribution $|\mathbf{p}_\alpha^{\text{noisy}}\rangle$. Here we define a dual of correction matrix A_α^\dagger acting on the spectrum of the local Hamiltonian. Note that if the correction matrix is exact (i.e., not approximate), it immediately follows that expected value of such defined Hamiltonian coincides with the true value of the energy

$$\langle H_\alpha^{\text{corr}} \rangle = \langle \lambda_\alpha A_\alpha^\dagger | \mathbf{p}_\alpha^{\text{noisy}} \rangle = \langle \lambda_\alpha A_\alpha^\dagger | \Lambda_\alpha \mathbf{p}_\alpha^{\text{ideal}} \rangle = \langle \lambda_\alpha | \underbrace{A_\alpha \Lambda_\alpha}_1 | \mathbf{p}_\alpha^{\text{ideal}} \rangle = \langle \lambda_\alpha | \mathbf{p}_\alpha^{\text{ideal}} \rangle = \langle H_\alpha \rangle . \quad (115)$$

Using this perspective we can derive the bounds on covariances between error-mitigated local Hamiltonian terms in a manner fully analogous to previous derivations. The ‘‘correlated part’’ this time is bounded as

$$|\langle \lambda_{\alpha\beta} A_{\alpha\beta}^\dagger | \Lambda_{\alpha\beta} \mathbf{p}_{\Delta_{\alpha\beta}}^{\text{ideal}} \rangle| \leq \|H_\alpha H_\beta\| \|A_{\alpha\beta}^\dagger\|_{1 \rightarrow 1} \|\Lambda_{\alpha\beta}\|_{1 \rightarrow 1} \|\Delta_{\alpha\beta}\|_1 \leq \|H_\alpha H_\beta\| \|A_{\alpha\beta}^\dagger\|_{1 \rightarrow 1} \epsilon_{\alpha\beta} . \quad (116)$$

The rest of discussion is identical to the previous analysis of the effects of uncorrelated noise, the difference being that now we have additional terms coming from the duals of correction matrices. Note that those duals in special cases can have huge values – indeed, if the noise matrix is barely invertible in the first place, one can expect that error-mitigation will highly increase the uncertainty in the estimation of energy.

E Details of numerical simulations and additional experimental data

In this short section, we provide some additional information on numerical simulations and experiments. This includes detailed discussion of methods used to simulate QAOA in Appendix E.1, and additional experimental results on characterization of correlations in Appendix E.2.

E.1 Simulation of the Quantum Approximate Optimization Algorithm

Here we provide some details of the performed numerical simulations. Simulation of QAOA was performed on the system of $N = 8$ qubits. In the main text, we described how the algorithm works, however, let us now repeat it for the sake of completeness. In standard implementation [24], one initializes quantum system to be in $|+\rangle^{\otimes N}$ state, where $|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$. Then p -layer QAOA is performed via implementation of unitaries of the form

$$U_p = \Pi_j^p \exp(-i \alpha_j \mathcal{H}_D) \exp(-i \beta_j \mathcal{H}_O), \quad (117)$$

where \mathcal{H}_D is driver Hamiltonian, which we take to be

$$\mathcal{H}_D = \sum_k^N \sigma_x^k, \quad (118)$$

and \mathcal{H}_O is objective Hamiltonian that one wishes to optimize (i.e., to find approximation for its ground state energy), and $\{\alpha_j\}, \{\beta_j\}$ are the angles to-be-optimized. The quantum state after p -th layer is

$$|\psi_p\rangle = U_p |+\rangle^{\otimes N}, \quad (119)$$

and the function which is passed to classical optimizer is the estimator of the expected value $\langle \psi_p | \mathcal{H}_O | \psi_p \rangle$ (note that this makes those estimators to effectively be a function of parameters $\{\alpha_j\}, \{\beta_j\}$). The estimator is obtained by sampling from the distribution defined by the measurement of $|\psi_p\rangle$ in the computational basis, taking the relevant marginals, and calculating the expected value of \mathcal{H}_O using values of those estimated marginals.

Theoretically one could optimize p -layer QAOA by simultaneous optimization of all $2p$ angles. However, this is hard in practice, since the number of optimized parameters increases the complexity of classical optimization. In our optimizations, we therefore modified the optimization to be divided into steps in the following way. In each step of optimization, we optimized over the set of 6 angles (i.e., 3 QAOA layers). Then the input to the next step was the optimized state obtained in the previous steps. For example, input to second optimization step was the quantum state $U_3^{(1)} |+\rangle^{\otimes N} = U_3 |+\rangle^{\otimes N}$, where we used superscript to denote optimization step. Then input to third optimization step was the state $U_3^{(2)} U_3^{(1)} |+\rangle^{\otimes N} = U_6 |+\rangle^{\otimes N}$, etc.

In summary, for p -layer QAOA, the optimization was effectively divided into $\frac{p}{3}$ steps, and in each step, we optimized over 6 angles, i.e., 3 QAOA layers. Each step used $2 * 800 = 1600$ function evaluations (the factor 2 comes from two evaluations needed for gradient evaluation which was done 800 times) plus a single final function evaluation. We further performed each such procedure additional time and chose the better run (out of 2). Each energy estimator was obtained using 10^4 energy measurements. Therefore the total number of function evaluations was $\approx \frac{p}{3} \times 3.2 \times 10^7$.

As a classical optimizer, we used Simultaneous perturbation stochastic approximation (SPSA) [39, 59–61]. Then with a growing number of optimization steps, we gradually changed the parameters to (heuristically) make the optimization more adaptive. Parameters α and γ were not changed, while other parameters were changed according to prescription

$$\begin{aligned} a_p &= a_0 0.9^p, \\ c_p &= c_0 0.9^p, \\ A_p &= A_0 1.1^p, \end{aligned} \tag{120}$$

where 0 subscript denotes starting values of parameters given in Table 1.

We note that the results presented in Fig. 6 exhibit rather poor convergence of the algorithm in a sense that adding more layers above $p = 9$ changed the resulting energies only slightly. This can be explained by the fact that we did not perform an exhaustive search over hyperparameters, but can also be a manifestation of the recently reported fact that QAOA might have problems with reaching global minima for relatively complicated Hamiltonians (like high-density MAX-2-SAT used in our work) [79]. Clearly, in the context of our work, only the comparison of noisy and noise-mitigated optimization to the noiseless run was of significance.

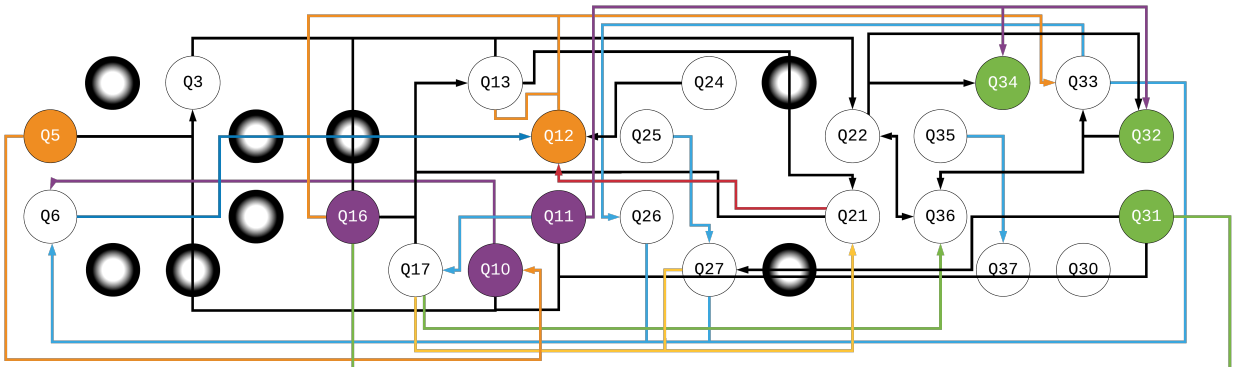


Figure 7: Depiction of correlations in Rigetti device. In Fig. 2 we presented the above image splitted into two parts for clarity.

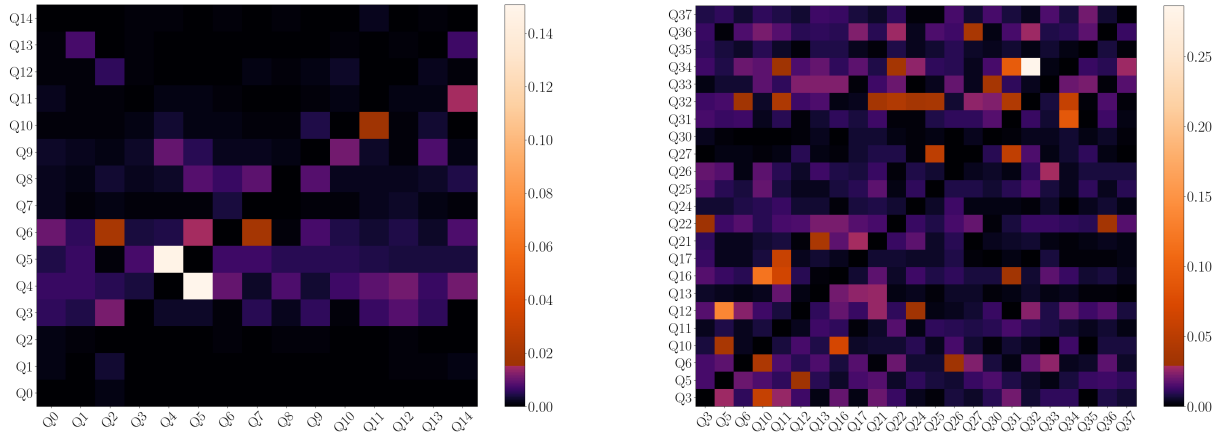


Figure 8: Heatmap of the correlations (Eq. (8)) in IBM’s 15-qubit *Melbourne* device (left) and a 23-qubit subset of Rigetti’s *Aspen-8* device (right). The convention is that “row is affected by column”, i.e., the measurement noise on the qubit with the label given by row index depends on the state of the qubit with the label given by column index, and the magnitude of the dependence is indicated by colors.

E.2 Additional experimental data

Here we present additional data concerning correlations reconstructed in DDOT characterization of IBM’s 15-qubit *Melbourne* device and a 23-qubit subset of Rigetti’s *Aspen-8* device. The full depiction of correlations in Rigetti’s device is presented in Fig. 7. The heatmaps showing the reported correlations in both devices are presented in Fig. 8. In Fig. 9 we show how many parameters are needed to describe various noise models, and table in Fig. 10 shows how much time data-processing took. The data was processed using laptop with 32GB DDR4 RAM (speed 2667MT/s) and Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz. We note that no multi-threading was implemented – in principle, one likely can reduce the run-time further by exploiting parallel calculations. We intend to optimize the code used for data-processing during development of our online repository [75].

IBM's Melbourne - 15 qubits						
noise model	uncorrelated	only neighbors	clusters & neighbors	CTMP [15]	full classical	full quantum
parameters	30	72	76	450	$\approx 1.0 * 10^9$	$\approx 3.5 * 10^{13}$

Rigetti's Aspen-8 - 25 qubits						
noise model	uncorrelated	only neighbors	clusters & neighbors	CTMP [15]	full classical	full quantum
parameters	50	342	756	1058	$\approx 7.0 * 10^{13}$	$\approx 5.9 * 10^{20}$

Figure 9: Total number of parameters needed to describe a noise model. The uncorrelated noise model is tensor product of single-qubit stochastic noise matrices. The "only neighbors" model corresponds to considering only single-qubit (trivial) clusters and their neighborhoods estimated in our experiments. The "clusters & neighbors" model corresponds to our full noise model containing both non-trivial clusters and their neighborhoods. CTMP is a number of parameters needed to describe 2-local classical noise model from Ref. [15] *without any assumptions* on the correlations structure. We note that combining our DDOT characterization with CTMP model could reduce the number of parameters in CTMP by pointing to negligible correlations that can be disregarded. For comparison, the "full classical" noise model refers to generic stochastic map, and "full quantum" to a generic d-outcome POVM.

	DDOT circuits	Shots per DDOT circuit	Pre-processing time	
IBM - 15q	749	8192	≈ 2.5 min	
Rigetti - 23q	504	1000	≈ 4.5 min	

	Error-mitigated experiments	Shots per experiment	Total post-processing time	Error-mitigation time
IBM - 15q	1200	40960	≈ 5.5 min	≈ 9 s
Rigetti - 23q	399	1000	≈ 2 min	≈ 6 s

Figure 10: Time of data processing. In first table, the "pre-processing time" includes calculation of marginal noise matrices on 2-qubit subsets, calculation of pairwise correlations, reconstruction of the noise model and calculation of inverse noise matrices needed for corrections of all possible two-qubit marginal probability distributions (note that this in general includes also higher-dimensional corrections, as explained in Fig. 3). In the second table, the "Total post-processing time" includes both calculation of marginal distributions needed to estimate energies of investigated 2-local Hamiltonians and performed error-mitigation on them. The "Error-mitigation time" shows only the latter.