nature computational science

Supplementary information

https://doi.org/10.1038/s43588-024-00709-1

Provable bounds for noise-free expectation values computed from noisy samples

In the format provided by the authors and unedited

I. PAULI TWIRLING

Within the theory of the paper we made the simplifying assumption of Pauli noise. This assumption is not given in general. Suppose a Clifford quantum circuit layer $\mathcal{U}(\cdot) = U \cdot U^\dagger$ on n qubits and its noisy version $\widetilde{\mathcal{U}} = \mathcal{U} \circ \Lambda$. A more realistic description of the noise is given by

$$\Lambda(\rho) \;=\; \sum_i A_i \rho A_i^\dagger \,, \tag{1}$$

where the A_i are Kraus operators [1], which leads to

$$\widetilde{\mathcal{U}}(\rho) \;=\; \sum_i A_i U \rho U^\dagger A_i^\dagger \,. \tag{2}$$

Applying Pauli twirling [2–4], i.e., averaging over $\mathcal{U}$ conjugated by each element of the Pauli group on n qubits yields

$$\widetilde{\mathcal{U}}_{\text{twirled}}(\rho) \;=\; \frac{1}{4^n} \sum_{i,j} Q_j A_i U P_j \rho P_j U^\dagger A_i^\dagger Q_j \,, \tag{3}$$

for Paulis P_j, Q_j with $Q_j U P_j = U$ for all $j = 1, \ldots, 4^n$. This is known to translate the more general noise given in (2) on average to a Pauli noise model as given in Main-Eq. 1. In practice, we do not enumerate all 4^n Paulis, but uniformly sample from them and apply a certain number of random Paulis to approximate the average.

Suppose now we have a noise model that-on average-looks like Pauli noise. Then, expectation values $\text{tr}(\rho H)$ will have the same value in case of a true Pauli noise model as well as in case of a twirled general model. That also holds if we set $H = |x\rangle\langle x|$, i.e., we evaluate the probability of sampling $|x\rangle$. If we estimate the same sampling probability for the actual Pauli noise model and the twirled noise model, then the sampling probabilities also must be the same.

II. PROBABILISTIC ERROR CANCELLATION & SAMPLING

In this section, we discuss how applying PEC [5] to quantum circuits affects the resulting sampling probabilities. PEC consists of two steps: learning the noise when running a quantum circuit on a particular quantum device, and then, mitigating the noise to get an unbiased estimator of an expectation value. Here, we assume we have learned the noise already and focus on the error mitigation. Given a noise model Λ, PEC constructs a *Quasiprobability Decomposition* (QPD) to implement the inverse noise by combining multiple weighted quantum circuits.

In a QPD, a quantum operation $\mathcal{U}$ is implemented as a linear combination of other (possibly noisy) operations $\mathcal{E}_i$, $i = 1, \ldots, M$,

$$\mathcal{U}(\cdot) \;=\; \sum_{i=1}^{M} a_i \mathcal{E}_i(\cdot) \,, \tag{4}$$

where $a_i \in \mathbb{R}$, $\mathcal{U}(X) = UXU^\dagger$, $\mathcal{E}_i$ denote (noisy) operations, and $\sum_{i=1}^M a_i = 1$. This has been proposed in the context of error mitigation [6], where $\mathcal{U}$ is assumed to be a noise-free operation and $\mathcal{E}_i$ are noisy operations that can be implemented on a noisy device. If this is being applied to multiple gates and qubits, the number of necessary operations M explodes exponentially. Thus, instead of enumerating all of them, one rewrites (4) as

$$\mathcal{U}(\cdot) \;=\; \gamma \sum_{i=1}^{M} p_i s_i \mathcal{E}_i(\cdot) \,, \tag{5}$$

where $\gamma = \|a\|_1 \geq 1$, $p_i = |a_i|/\gamma$, and $s_i = \text{sign}(a_i)$, and samples from the probability distribution defined through p_i. Suppose we are interested in estimating $\langle H \rangle = \text{tr}(\mathcal{U}(\rho) H)$ for some initial state ρ and observable H. Then, we can use the QPD to write

$$\text{tr}(\mathcal{U}(\rho) H) \;=\; \gamma \sum_{i=1}^{M} p_i s_i \,\text{tr}(\mathcal{E}_i(\rho) H). \tag{6}$$

Thus, instead of enumerating all M circuits, we can sample from p_i, and only evaluate the sampled circuits corresponding to i, to get an unbiased estimator for $\langle H \rangle$. However, the variance of this estimation is amplified by γ^2, i.e., γ^2-times more samples are needed than for the original noise-free circuit to achieve an estimate of the same accuracy. The sampling overhead γ^2 grows exponentially in the number of qubits and depth of the circuit, and thus, can be prohibitively large for circuits beyond a certain circuit size and noise levels.

While PEC has only been considered for the estimation of expectation values, it also generates samples from every random circuit that is measured. However, we will show that this essentially amplifies the noise and increases the sampling overhead compared to the results presented within this paper. To this extent, we introduce the following mixed state introduced by PEC:

$$\rho_{\text{PEC}} \;=\; \sum_{i=1}^{M} p_i \mathcal{E}_i(\rho), \tag{7}$$

for some initial state ρ. The state ρ_{PEC} is achieved by dropping the factor γ as well as the signs s_i from (5). This allows us to state the following lemma.

Lemma 1. *Suppose a n-qubit state $\rho = U|0\rangle\langle 0|U^\dagger$, where U is some unitary, with*

$$\text{tr}(\rho |x\rangle\langle x|) \;=\; p_x \geq 0, \tag{8}$$

for a computational basis state $|x\rangle$, $x \in \{0,1\}^n$.

Further, suppose that U can be error-mitigated on a noisy device by using PEC with corresponding $\gamma \geq 1$ and denote the resulting mixed state introduced in (7) by ρ_{PEC}. Then, the probability of measuring $|x\rangle$ on the noisy devices using PEC is lower bounded by

$$\text{tr}(\rho_{PEC} |x\rangle\langle x|) = p_x^{PEC} \;\geq\; p_x/\gamma. \tag{9}$$

Proof. Consider the QPD resulting from PEC

$$\mathcal{U}(\cdot) = \sum_{i=1}^{M} a_i \mathcal{E}_i(\cdot)\,. \tag{10}$$

Using (10) we can write

$$p_x = \mathrm{tr}(\rho|x\rangle\langle x|) = \sum_{i=1}^{M} a_i \,\mathrm{tr}(\mathcal{E}_i(|0\rangle\langle 0|)|x\rangle\langle x|)\,. \tag{11}$$

By defining $\gamma = \|a\|_1$, $p_i = |a_i|/\gamma$, and $s_i = \mathrm{sign}(a_i)$, we can rewrite (11) as

$$p_x = \gamma \sum_{i=1}^{M} p_i s_i \,\mathrm{tr}(\mathcal{E}_i(|0\rangle\langle 0|)|x\rangle\langle x|)\,. \tag{12}$$

Further, $s_i \,\mathrm{tr}(\mathcal{E}_i(|0\rangle\langle 0|)|x\rangle\langle x|)$ allows us to define a random variable $Y_i \in \{-1, 0, +1\}$ that equals ± 1 if we measure $\mathcal{E}_i(|0\rangle\langle 0|)$ and obtain $|x\rangle$, where the sign is determined by s_i, and 0 otherwise. The random variable Y_i satisfies $\mathbb{E}[Y_i] = s_i \,\mathrm{tr}(\mathcal{E}_i(|0\rangle\langle 0|)|x\rangle\langle x|)$. We denote the probabilities of Y_i taking the values $-1, 0, +1$ by $q_i^{-1}, q_i^0, q_i^{+1} \geq 0$, respectively. Note that by construction, for each i only one of q_i^{-1}, q_i^{+1} can be larger than zero.

In addition, let the probabilities p_i define a random variable $I \in \{1, \ldots, M\}$. Then, by the law of total expectation, we get

$$\gamma\mathbb{E}[Y_I] = \gamma \sum_{i=1}^{M} \mathbb{E}[Y_i|i]\mathbb{P}[i] \tag{13}$$

$$= \gamma \sum_{i=1}^{M} p_i s_i \,\mathrm{tr}(\mathcal{E}_i(|0\rangle\langle 0|)|x\rangle\langle x|) \tag{14}$$

$$= p_x\,. \tag{15}$$

This can be rewritten as

$$\sum_{i=1}^{M} p_i \left(q_i^{+1} - q_i^{-1}\right) \;=\; \frac{p_x}{\gamma}\,. \tag{16}$$

The total probability to measure $|x\rangle$ when applying PEC, independent of the sign of Y_I, is then given by

$$\sum_{i=1}^{M} p_i \left(q_i^{+1} + q_i^{-1}\right) \;\geq\; \frac{p_x}{\gamma}, \tag{17}$$

where the lower bound follows immediately from (16), and the right-hand-side is exactly the probability of measuring $|x\rangle$ for state ρ_{PEC}. □

If we compare the result from Lemma 1 with the lower bound presented in Main-Eq. 2, we see that PEC implies the squared overhead compared to direct sampling. Further, this implies that CVaR-based approaches may substantially reduce the overhead to achieve insightful results, particularly when combined with problem structure to filter noisy samples.

III. VARIANCE OF ESTIMATING THE CVAR

In this section, we present a short exposition on how to estimate CVaR. We will first state the following lemma.

Lemma 2. *Let $X_1, \ldots, X_n$ be i.i.d. copies of X (with X integrable) and let $X_{(1)}, \ldots, X_{(n)}$ be their order statistic. For $\alpha \in (0, 1]$ let $E_n = (X_{(1)} + \cdots + X_{(\lfloor \alpha n \rfloor)})/\lfloor \alpha n \rfloor$. Then*

$$\mathbb{E}[E_n] \to \mathrm{CVaR}_\alpha(X) \quad \text{as } n \to \infty\,.$$

If X is square integrable and $F_X(x_\alpha) = \alpha$,

$$\sqrt{n}(E_n - \mathrm{CVaR}_\alpha(X)) \to N(0, \mathrm{CVaRv}_\alpha(X))$$

in distribution as $n \to \infty$ where here $\mathrm{CVaRv}_\alpha(X) := \alpha^{-1}\,\mathrm{Var}[X \mid X \leq x_\alpha]$ is the limiting variance.

To estimate $\overline{\mathrm{CVaR}}_\alpha(X)$, we use the estimator $\overline{E}_n = (X_{(n-\lfloor \alpha n \rfloor + 1)} + \cdots + X_{(n)})/\lfloor \alpha n \rfloor$ and obtain analogous results.

Proof. Recall $F_X(x) = \mathbb{P}[X \leq x]$ and define $F_X(x-) = \mathbb{P}[X < x]$. We make the following definitions for (left limits) of empirical cumulative distribution functions:

$$\hat{F}_n(x) = \#\{i \leq n \colon X_i \leq x\}/n\,,$$
$$\hat{F}_n(x-) = \#\{i \leq n \colon X_i < x\}/n\,.$$

Also let $\Delta F_X(x) = F_X(x) - F_X(x-)$ and $\Delta\hat{F}_n(x) = \hat{F}_n(x) - \hat{F}_n(x-)$. The key observation is that

$$E_n = \frac{1}{\lfloor \alpha n \rfloor} \sum_{i=1}^{n} X_i \min\left\{ \frac{(\lfloor \alpha n \rfloor - n\hat{F}_n(X_i-))_+}{n\Delta\hat{F}_n(X_i)}, 1 \right\}.$$

Indeed, any $x \in \mathbb{R}$ will appear in the sum defining $\lfloor \alpha n \rfloor E_n$ precisely $\min\{(\lfloor \alpha n \rfloor - n\hat{F}_n(x-))_+, n\Delta\hat{F}_n(x)\}$ times; the $n\Delta\hat{F}_n(X_i)$ in the denominator above takes care of overcounting. Now

$$\mathbb{E}[E_n] = \frac{n}{\lfloor \alpha n \rfloor}\mathbb{E}\left[X_1 \min\left\{ \frac{(\lfloor \alpha n \rfloor - n\hat{F}_n(X_1-))_+}{n\Delta\hat{F}_n(X_1)}, 1 \right\}\right]$$
$$= \frac{n}{\lfloor \alpha n \rfloor}\mathbb{E}\left[A_n(X_1)\right]$$

where

$$A_n(x) := x\mathbb{E}\left[\min\left\{ \frac{(\lfloor \alpha n \rfloor - n\hat{F}_{n-1}(x-))_+}{1 + n\Delta\hat{F}_{n-1}(x)}, 1 \right\}\right].$$

The first equality above follows from the linearity of the expectation and the i.i.d. property of $X_1, \ldots, X_n$ and the second equality follows from conditioning on X_1. Using the strong law of large numbers we have $(\hat{F}_n(x), \hat{F}_n(x-), \Delta\hat{F}_n(x)) \to (F_X(x), F_X(x-), \Delta F_X(x))$ a.s. as $n \to \infty$. By separately considering the $\Delta F_X(x) = 0$ and $\Delta F_X(x) > 0$ cases we get

$$A_n(x) \to x \cdot 1(F_X(x) < \alpha)$$
$$+ x\frac{\alpha - F_X(x-)}{\Delta F_X(x)}\, 1(\alpha \in (F_X(x-), F_X(x)))$$

as $n \to \infty$ unless $\alpha = F_X(x) = F_X(x-)$; however we have $\mathbb{P}[\alpha = F_X(X_1) = F_X(X_1-)] = 0$ so this case does not matter to evaluate the limit of $\mathbb{E}[E_n]$. Thus by dominated convergence

$$\begin{aligned}\mathbb{E}[E_n] &\to \alpha^{-1}\mathbb{E}[X_1 \mid F_X(X_1) < \alpha] \\ &\quad + \sum_{x\colon\, \alpha \in (F_X(x-), F_X(x))} x(1 - \alpha^{-1}F_X(x-)) \\ &= \mathrm{CVaR}_\alpha(X)\end{aligned}$$

as $n \to \infty$. The second claim on the central limit theorem is a special case of [7]. □

Let us make the following remark on monotonicity: If $\phi\colon \mathbb{R} \to \mathbb{R}$ is non-decreasing and $\phi(X)$ is integrable, then

$$\begin{aligned}0 &\geq \mathbb{E}[(\phi(X) - \phi(X'))(1(X \leq x) - 1(X' \leq x))] \\ &= 2 \cdot \mathbb{E}[\phi(X) \mid X \leq x] - 2 \cdot \mathbb{E}[\phi(X)]\mathbb{P}[X \leq x]\,.\end{aligned}$$

By applying this to $\phi(x) = x$ and $x = x_\alpha$ we see that $\mathrm{CVaR}_\alpha(X) \leq \mathbb{E}[X]$. Furthermore, by replacing X by a random variable sampled from the law of X conditioned on $X \leq x_{\alpha'}$ for $\alpha' > \alpha$ we can deduce that $\mathrm{CVaR}_\alpha(X)$ is non-decreasing in α. Much more crudely, we can bound $\mathrm{CVaRv}_\alpha(X) \leq \alpha^{-1}\mathbb{E}[X^2]/\mathbb{P}[X \leq x_\alpha] \leq \mathbb{E}[X^2]/\alpha^2$.

In the following, we analyze behavior of the limiting distribution of the estimator E_n in some concrete cases.

In the case where X has a Bernoulli distribution with success probability p, we observe that E_n has the same distribution as $\min\{B_n/\lfloor \alpha n \rfloor, 1\}$ where B_n is Binomial distributed with parameter (n, p). An application of the central limit theorem thus yields

$$\begin{aligned}&\sqrt{n}(E_n - \min\{p/\alpha, 1\}) \\ &\quad \to \begin{cases} \alpha^{-1}\sqrt{p(1-p)}N & : \alpha > p \\ \sqrt{(1-p)p^{-1}}\, N \cdot 1(N \geq 0) & : \alpha = p, \\ 0 & : \alpha < p\,. \end{cases}\end{aligned}$$

in distribution as $n \to \infty$ where N is a standard normal random variable.

To analyze the case where $N \sim N(0,1)$, it will be useful to recall the following asymptotic expansion [8, (8.11(i))] of incomplete Gamma functions:

$$\begin{aligned}\Gamma(a, y) &:= \int_y^\infty s^{a-1}e^{-s}\, ds \\ &= y^{a-1}e^{-y}\left(\sum_{k=0}^{n-1} \frac{(a-1)\cdots(a-k)}{y^k} + O(y^{-n})\right)\end{aligned}$$

as $y \to \infty$ for any fixed $n \geq 1$ and $a > 0$. In particular as $x \to \infty$,

$$\begin{aligned}\frac{\Gamma(1/2, x^2/2)}{\sqrt{2}} &= \frac{e^{-x^2/2}}{x}\left(1 - \frac{1}{x^2} + \frac{3}{x^4} + O(x^{-6})\right), \\ \frac{\sqrt{2}}{\Gamma(1/2, x^2/2)} &= xe^{x^2/2}\left(1 + \frac{1}{x^2} - \frac{2}{x^4} + O(x^{-6})\right).\end{aligned}$$

Let $x_\alpha = F_N^{-1}(\alpha)$ and write $f_N = F_N'$ for the density of N. By [8, (7.17(iii))] we get the asymptotic relationship

$$x_\alpha \sim -\sqrt{-\log(4\pi\alpha^2 \log(1/(2\alpha)))}$$

as $\alpha \to 0$. We will compute $\mathrm{CVaR}_\alpha(N)$ and $\mathrm{CVaRv}_\alpha(N)$ via the cumulant generating function ϕ of a truncated Gaussian

$$\begin{aligned}\phi(\theta) &= \log \mathbb{E}[e^{\theta N} \mid N \leq x_\alpha] \\ &= -\log F_N(x_\alpha) + \log \int_{-\infty}^{x_\alpha} \frac{e^{-t^2/2+\theta t}}{\sqrt{2\pi}}\, dt \\ &= -\log F_N(x_\alpha) + \log \int_{-\infty}^{x_\alpha} \frac{e^{-(t-\theta)^2/2+\theta^2/2}}{\sqrt{2\pi}}\, dt \\ &= -\log F_N(x_\alpha) + \theta^2/2 + \log F_N(x_\alpha - \theta).\end{aligned}$$

Differentiating at $\theta = 0$ yields the expressions

$$\begin{aligned}\mathrm{CVaR}_\alpha(N) &= \phi'(0) = -\frac{1}{\alpha}f_N(x_\alpha), \\ \alpha\,\mathrm{CVaRv}_\alpha(N) &= \phi''(0) = 1 + \frac{1}{\alpha}f_N'(x_\alpha) - \frac{1}{\alpha^2}f_N(x_\alpha)^2 \\ &= 1 - \frac{1}{\alpha}x_\alpha f_N(x_\alpha) - \frac{1}{\alpha^2}f_N(x_\alpha)^2.\end{aligned}$$

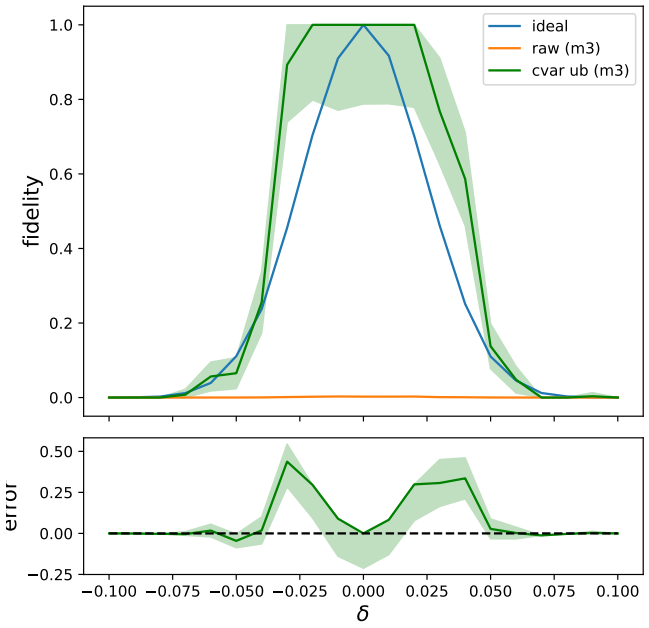
Since $\alpha = F_N(x_\alpha) = \Gamma(1/2, x_\alpha^2/2)/(2\sqrt{\pi})$, it follows that as $\alpha \to 0$,

$$\begin{aligned}\mathrm{CVaR}_\alpha(N) &= x_\alpha\left(1 + \frac{1}{x_\alpha^2} - \frac{2}{x_\alpha^2} + O(x_\alpha^{-6})\right), \\ \mathrm{CVaRv}_\alpha(N) &= \frac{1}{\alpha} + \frac{x_\alpha^2}{\alpha}\left(1 + \frac{1}{x_\alpha^2} - \frac{2}{x_\alpha^2} + O(x_\alpha^{-6})\right) \\ &\quad - \frac{x_\alpha^2}{\alpha}\left(1 + \frac{1}{x_\alpha^2} - \frac{2}{x_\alpha^2} + O(x_\alpha^{-6})\right)^2 \\ &= \frac{1}{\alpha x_\alpha^2} + O(x_\alpha^{-4}).\end{aligned}$$

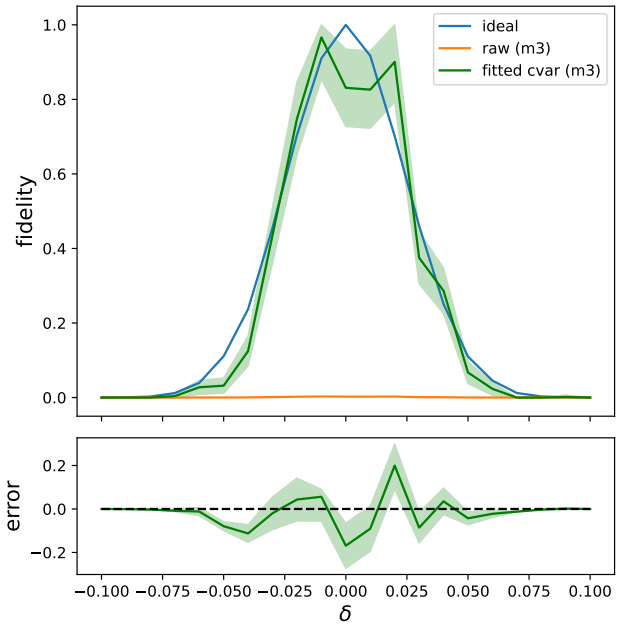
As a final example, we can consider the case where X has density $f_X(x) = \beta x^{-1-\beta}1(x \geq 1)$ where $\beta > 0$ (i.e., we consider a power law tail). Here, one can compute that for $\beta > 2$, $\overline{\mathrm{CVaRv}}_\alpha(X) = \beta(\beta-1)^{-2}(\beta-2)^{-1}\alpha^{-2/\beta}$ which is worse than the decay in the standard normal case and achieves the worst case upper bound on the variance in the $\beta \to 2$ limit.

IV. RELATION TO BRUTE-FORCE SEARCH

A brute-force search enumerates all 2^n candidate solutions and checks which one is optimal. The sampling overhead of $\sqrt{\gamma}$ on noisy devices can thus be related to brute-force search thereby allowing us to derive a hardware requirements for QAOA. Assuming, for simplicity, that the probability p_x to sample the optimal solution is close to 1 we require hardware with $\sqrt{\gamma} < 2^n$. We can relate this to the layer fidelity to obtain a requirement



SUPPLEMENTARY FIG. 1: Fidelity estimates on 100 qubits: (top) Ideal results from noise-free simulation, raw fidelity estimates using only M3 readout error mitigation, and the $\overline{\text{CVaR}}$ upper bounds and corresponding 95% confidence intervals (shaded area). (bottom) Difference between $\overline{\text{CVaR}}$ upper bounds and ideal noise-free results and corresponding 95% confidence intervals (shaded area).



SUPPLEMENTARY FIG. 2: Fidelity fit on 100 qubits: (top) Ideal results from noise-free simulation, raw fidelity estimates using only M3 readout error mitigation, and the fitted $\overline{\text{CVaR}}$ values and corresponding 95% confidence intervals (shaded area). (bottom) Difference between fitted $\overline{\text{CVaR}}$ values and ideal noise-free results and corresponding 95% confidence intervals (shaded area).

on hardware quality necessary for potential quantum advantage. First, we assume that each layer i in a QAOA circuit has the same layer fidelity $\text{LF} = 1/\sqrt{\gamma_i}$. As a result the γ of the circuit is $\gamma = \prod_{i=1}^{d(n)} \gamma_i = 1/\text{LF}^{2d(n)}$ where $d(n)$ is the depth defined as the number of non-overlapping two-qubit gate layers. This assumption is reasonable when transpiling QAOA circuits to a line of qubits which requires layers of CNOT gates applied on every other edge [9]. Therefore, the sampling cost to compensate for noise is $1/\text{LF}^{d(n)}$. For a line of qubits we may assume that to leading order $d(n) \sim 3np$. The factor $3n$ comes from the fact that $n-2$ layers of SWAP gates are needed to implement full connectivity and each SWAP merged with an R_{ZZ} is implemented with three CNOT gates. Here, p is the number of QAOA layers which is sometimes assumed to grow with the logarithm of problem size, i.e., $p \propto \log(n)$ [9, 10]. If the sampling overhead should stay below brute-force search we therefore require $\text{LF}^{-3np} < 2^n$ which implies that the layer fidelity must satisfy

$$\text{LF} > \frac{1}{2^{1/3p}}. \tag{18}$$

This requirement is only dependent on problem size through the relation between p and n. However, the layer fidelity decreases with the number of qubits in the layer [11]. If we further assume that layers are dense, i.e., every layer on n qubits consists of approximately $n/2$ CNOT gates, we can compute a corresponding CNOT fidelity as $\text{LF}^{2/n}$, as well as the corresponding lower bound

$$\text{LF}^{2/n} > \frac{1}{2^{2/3pn}}. \tag{19}$$

V. 100-QUBIT FIDELITY ESTIMATION

In this section, we report the results of the 100-qubit fidelity estimation experiments. The setup is almost the same as in Main-Sec. 'Experiments - Fidelity Estimation' except for the number of shots used and the discretization of δ. Instead of 20,000 shots per data point, we use 100,000 shots per data point. We discretize δ between -0.1 and +0.1 with steps of 0.01. Since the circuits are larger, the resulting layer fidelities are smaller and have been estimated as 0.4620 (even) and 0.4220 (odd), respectively. This implies $\alpha = 0.00144$ and an EPLG of 0.01638. Thus, the sampling overhead for CVaR is given by $\sqrt{\gamma} = 692.4$ while the sampling overhead for PEC would be given by $\gamma^2 = 2.30 \times 10^{11}$. This implies that running PEC would require more than 10^8 times more

samples, which is clearly impractical. The hardware results are shown in Supplementary Fig. 1. The results show again a nice agreement between theory and experiment, although the confidence intervals get larger due to the increased sampling overhead. To further reduce the width of the confidence intervals, we would have to further increase the number of samples.

Fitting the CVaR to the data by adjusting α leads to $\alpha = 0.00295$ and an *effective EPLG* of 0.01461. The results are shown in Supplementary Fig. 2. This provides a very close approximation of the fidelity with substantially smaller overhead than PEC and may be used as a building block in the aforementioned algorithms.

VI. 40-QUBIT QAOA CIRCUITS

The 40-qubit circuits in the main texts are based on those by Sack and Egger [12]. In this work, the authors consider random three-regular graphs transpiled to a line of qubits using a swap network [9]. This results in circuits that alternate only two types of layers of CNOT gates as described in the main text. Furthermore, the authors carefully chose the decision variable to physical qubit mapping to minimize the number of layers of the swap network as described by Matsuo et al. [13]. The optimal parameters resulting from the light-cone optimization are given by $(\gamma_1, \beta_1) = (2.8405, 0.3982)$ for $p = 1$ and $(\gamma_1, \beta_1, \gamma_2, \beta_2) = (1.1506, 0.3288, 0.1941, 0.6582)$ for $p = 2$, respectively.

VII. 127-QUBIT QAOA CIRCUITS

Supplementary Fig. 3 shows an example of the considered random higher-order Ising model. In Supplementary Fig. 4, taken from Refs. [14, 15], we briefly discuss the optimized circuits for the 127-qubit higher-order instances to have a self-contained description. This illustrates that all 2-qubit gates needed for the implementation of $e^{-i\gamma H}$ can be scheduled in just 3 different layers of non-overlapping CNOT gates. In each QAOA round p, each layer is used once to compute and once to uncompute ZZ and ZZZ parity values, for an overall CNOT depth of $6p$. The exact values of the heuristically computed, using parameter transfer, QAOA angles that give a strictly increasing expectation value as p increases up to 5 are given by Pelofske et al. [16].

[1] M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, 2011.

[2] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland. Randomized benchmarking of quantum gates. *Physical Review A*, 77(1), 2008. DOI: 10.1103/PhysRevA.77.012307. Publisher: American Physical Society.

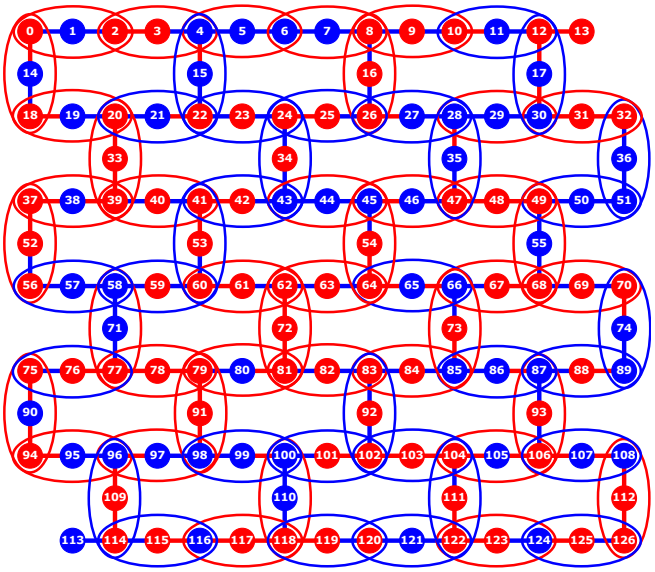
[3] C. Dankert, R. Cleve, J. Emerson, and E. Livine. Exact and approximate unitary 2-designs and their application to fidelity estimation. *Physical Review A*, 80(1):012304, 2009. DOI: 10.1103/PhysRevA.80.012304. Publisher: American Physical Society.

[4] E. Magesan, J. M. Gambetta, and J. Emerson. Scalable and Robust Randomized Benchmarking of Quantum Processes. *Physical Review Letters*, 106(18):180504, 2011. DOI: 10.1103/PhysRevLett.106.180504. Publisher: American Physical Society.

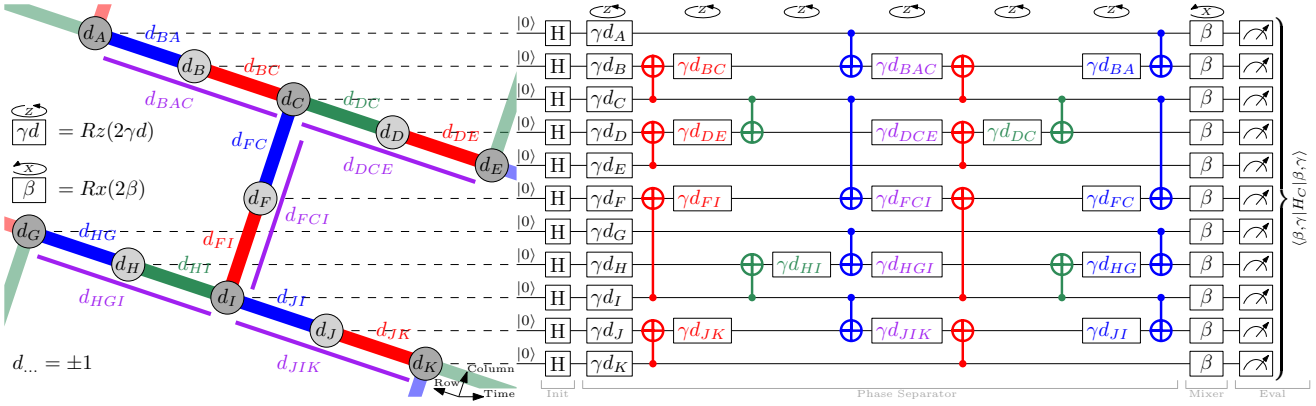
[5] E. van den Berg, Z. K. Minev, A. Kandala, and K. Temme. Probabilistic error cancellation with sparse Pauli-Lindblad models on noisy quantum processors. *Nature Physics*, 19:1116–1121, 2023. DOI: 10.1038/s41567-023-02042-2.

[6] K. Temme, S. Bravyi, and J. M. Gambetta. Error Mitigation for Short-Depth Quantum Circuits. *Physical Review Letters*, 119(18), 2017. DOI: 10.1103/physrevlett.119.180509.

[7] J. Dedecker and F. Merlevède. Central limit theorem and almost sure results for the empirical estimator of superquantiles/CVaR in the stationary case. *Statistics*, 56(1):53–72, 2022. DOI: 10.1080/02331888.2022.2043325.



SUPPLEMENTARY FIG. 3: Example heavy-hex hardware-compatible 127-qubit higher-order Ising model. Nodes denote the linear terms, edges between nodes denote the quadratic terms, and the ovals encircling three neighboring nodes on the hardware graph denote hyper-edges. Polynomial (Ising model) coefficients of -1 are denoted by red, and $+1$ are denoted by blue.



SUPPLEMENTARY FIG. 4: **From Refs.** [14, 15]: Diagram of a heavy-hex graph compatible $p = 1$ QAOA circuit for sampling heavy-hex compatible higher-order Ising models (specifically cubic terms centered on degree 2 nodes). Left hand side of the figure shows the 3-edge-coloring and the bipartition of the graph, and the cubic terms are denoted by the adjacent purple lines next to the hardware graph. The right hand side of the figure shows the corresponding QAOA circuit for this sub-component of the heavy-hex graph (which can be extended arbitrarily to a large heavy-hex graph, and to higher p). The cubic terms are addressed by a single layer of Rx rotation gates (shown in purple), and the total CNOT depth per p is always 6. Following the phase separator, the transverse field mixer is applied and then the state of all qubits are measured after p rounds have been applied.

[8] *NIST Digital Library of Mathematical Functions*. https://dlmf.nist.gov/, Release 1.1.11 of 2023-09-15. Available online: https://dlmf.nist.gov/. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

[9] J. Weidenfeller, L. C. Valor, J. Gacon, C. Tornow, L. Bello, S. Woerner, and D. J. Egger. Scaling of the quantum approximate optimization algorithm on superconducting qubit based hardware. *Quantum*, 6:870, 2022. DOI: 10.22331/q-2022-12-07-870.

[10] S. Bravyi, A. Kliesch, R. Koenig, and E. Tang. Obstacles to Variational Quantum Optimization from Symmetry Protection. *Physical Review Letters*, 125(26):260505, 2020. DOI: 10.1103/PhysRevLett.125.260505.

[11] D. C. McKay, I. Hincks, E. J. Pritchett, M. Carroll, L. C. G. Govia, and S. T. Merkel. Benchmarking Quantum Processor Performance at Scale, 2023. DOI: 10.48550/arXiv.2311.05933.

[12] S. H. Sack and D. J. Egger. Large-scale quantum approximate optimization on nonplanar graphs with machine learning noise mitigation. *Physical Review Research*, 6(1), 2024. DOI: 10.1103/physrevresearch.6.013223.

[13] A. Matsuo, S. Yamashita, and D. J. Egger. A SAT Approach to the Initial Mapping Problem in SWAP Gate Insertion for Commuting Gates. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E106.A(11):1424–1431, 2023. DOI: 10.1587/transfun.2022eap1159.

[14] E. Pelofske, A. Bärtschi, and S. Eidenbenz. Quantum Annealing vs. QAOA: 127 Qubit Higher-Order Ising Problems on NISQ Computers. In *International Conference on High Performance Computing ISC HPC'23*, pages 240–258, 2023. DOI: 10.1007/978-3-031-32041-5_13.

[15] E. Pelofske, A. Bärtschi, and S. Eidenbenz. Short-Depth QAOA circuits and Quantum Annealing on Higher-Order Ising Models. *npj Quantum Information*, 10:30, 2024. DOI: 10.1038/s41534-024-00825-w.

[16] E. Pelofske, A. Bärtschi, L. Cincio, J. Golden, and S. Eidenbenz. Scaling Whole-Chip QAOA for Higher-Order Ising Spin Glass Models on Heavy-Hex Graphs, 2023. DOI: 10.48550/arXiv.2312.00997.