Quantum error mitigation

Zhenyu Cai*

Department of Materials, University of Oxford, Oxford OX1 3PH, United Kingdom and Quantum Motion, 9 Sterling Way, London N7 9HJ, United Kingdom

Ryan Babbush

Google Quantum AI, Venice, California 90291, USA

Simon C. Benjamin

Department of Materials, University of Oxford, Oxford OX1 3PH, United Kingdom and Quantum Motion, 9 Sterling Way, London N7 9HJ, United Kingdom

Suguru Endo

NTT Computer and Data Science Laboratories, NTT Corporation, Musashino 180-8585, Japan

William J. Huggins

Google Quantum AI, Venice, California 90291, USA

Ying Li

Graduate School of China Academy of Engineering Physics, Beijing 100193, China

Jarrod R. McClean and Thomas E. O'Brien

Google Quantum AI, Venice, California 90291, USA

(published 13 December 2023)

For quantum computers to successfully solve real-world problems, it is necessary to tackle the challenge of *noise*: the errors that occur in elementary physical components due to unwanted or imperfect interactions. The theory of quantum fault tolerance can provide an answer in the long term, but in the coming era of noisy intermediate-scale quantum machines one must seek to mitigate errors rather than completely eliminate them. This review surveys the diverse methods that have been proposed for quantum error mitigation, assesses their in-principle efficacy, and describes the hardware demonstrations achieved to date. Commonalities and limitations among the methods are identified, while mention is made of how mitigation methods can be chosen according to the primary type of noise present, including algorithmic errors. Open problems in the field are identified, and the prospects for realizing mitigation-based devices that can deliver a quantum advantage with an impact on science and business are discussed.

DOI: 10.1103/RevModPhys.95.045005

CONTENTS

I. Introduction 2
II. Concepts 3
A. Narrative introduction to concepts and terminology 3
B. Error-mitigated estimators 3
C. Faults in the circuit 4
D. Exponential scaling of the sampling overhead 5
III. Methods 5
A. Zero-noise extrapolation 5
B. Probabilistic error cancellation 7
C. Measurement error mitigation 9
D. Symmetry constraints 10
E. Purity constraints 12
F. Subspace expansions 14
G. *N* representability 16
H. Learning-based methods 17
IV. Comparisons and Combinations 19
A. Comparison among QEM methods 19
1. Noise calibration overhead 19
a. Gate error mitigation 19
b. State error mitigation 20
c. Observable error mitigation 20
2. Mean square errors 21
B. Benchmarking QEM from other perspectives 21
1. State discrimination 21
2. Quantum estimation theory 22

*Corresponding author: cai.zhenyu.physics@gmail.com

3. State extraction 22
C. Combinations of QEM methods 23
D. Comparison to the other error-suppression methods 24
V. Applications 25
A. Coherent errors 25
B. Logical errors in fault-tolerant quantum computation 26
C. Algorithmic (compilation) errors 27
VI. Open Problems 27
A. Overarching problems 27
B. Technical questions 28
VII. Conclusion 28
List of Symbols and Abbreviations 28
Acknowledgments 28
Appendix: Practical Techniques in Implementations 29
1. Monte Carlo sampling 29
a. Exponential sampling overhead 29
2. Pauli twirling 30
3. Measurement techniques 30
References 31

I. INTRODUCTION

The central promise of quantum computing is to enable algorithms that have been shown to provide both polynomial and superpolynomial speedups over the best-known classical algorithms for a special set of problems. These problems range from simulating quantum mechanics (Feynman, 1982) to purely algebraic problems such as factoring integers (Shor, 1999). The strongest challenge to the viability of practical quantum computing has always been its sensitivity to errors and noise. It was realized early on that the coupling of quantum systems to their environment sets an ultimate time limit and size limit for any quantum computation (Unruh, 1995). This constraint poses a formidable challenge to the ambitions of realizing a quantum computer since it set bounds on the scalability of any algorithm. With the advent of quantum error correction (QEC) (Shor, 1995; Calderbank and Shor, 1996; Steane, 1996), this challenge has been solved, at least in theory. The well-known threshold theorem (Aharonov and Ben-Or, 1997; Kitaev, 1997) showed that if errors in the quantum hardware could be reduced below a finite rate, known as the threshold, a fault-tolerant quantum computation could be carried out for an arbitrary length even on noisy hardware. However, besides the technical challenge of building hardware that achieves the threshold, the implementation of a fault-tolerant universal gate set with current codes, such as the surface code (Fowler *et al.*, 2012), generates a qubit overhead that currently seems daunting . For example, recent optimized approaches showed that scientific applications that are classically intractable may require hundreds of thousands of qubits (Kivlichan *et al.*, 2020), while industrial applications will require millions of qubits (Lee *et al.*, 2021). There is ongoing theoretical research to find alternative codes with a more favorable overhead, and recent progress gives reasons for optimism (Gottesman, 2014; Breuckmann and Eberhardt, 2021; Dinur *et al.*, 2022; Panteleev and Kalachev, 2022). Nevertheless, the challenge of realizing full-scale fault-tolerant quantum computing is a considerable one.

This begs the question as to whether other approaches, prior to the era of fully fault-tolerant systems, might achieve quantum advantage with significant practical impacts. One might hope so, given the continual and noteworthy progress that has been made in quantum computational hardware. In recent years, it has become routine to see reports of experiments demonstrating high-quality control over multiple qubits (Asavanant *et al.*, 2019; Madjarov *et al.*, 2020; Ebadi *et al.*, 2021; Jurcevic *et al.*, 2021; Xue *et al.*, 2022), sometimes reaching even beyond 50 qubits (Arute *et al.*, 2019; Wu *et al.*, 2021). Meanwhile, other experiments have indeed demonstrated early-stage fault-tolerant potentials (see Egan *et al.*, 2021; Abobeih *et al.*, 2022; Krinner *et al.*, 2022; Postler *et al.*, 2022; Ryan-Anderson *et al.*, 2022; Takeda *et al.*, 2022; Google Quantum AI, 2023). The works mentioned here are far from exhaustive, as it is impossible to capture all of the breakthroughs on different fronts across the diverse range of platforms. See Acín *et al.* (2018) and Altman *et al.* (2021) and references therein for key milestones in different platforms.

The primary goal of *quantum error mitigation* (QEM) is to translate this continuous progress in quantum hardware into immediate improvements for quantum information processing. While accepting that hardware imperfections will limit the complexity of quantum algorithms, nevertheless we can expect every advance to enable this boundary to be pushed further. As this review demonstrates, the mitigation approach indeed proves to be both practically effective and interesting as an intellectual challenge.

When exploring the prospects for achieving quantum advantage through error mitigation, it is crucial to consider suitable forms of circuits. It is understood that in the era of noisy intermediate-scale quantum (NISQ) devices only certain approaches may be able to achieve meaningful and useful results. Owing to the limited coherence times and the noise floor present in quantum hardware, one typically resorts to the idea of quantum computation with short-depth circuits. Motivating examples include variational quantum circuits in physics simulations (Peruzzo *et al.*, 2014; Wecker, Hastings, and Troyer, 2015; McClean *et al.*, 2016), approximate optimization algorithms (Farhi, Goldstone, and Gutmann, 2014), and even heuristic algorithms for quantum machine learning (Biamonte *et al.*, 2017). In applications of these kinds, the algorithm can typically be understood as applying a short-depth quantum circuit to a simple initial state and then estimating the expectation value of a relevant observable. Such expectation values ultimately lead to the output of the algorithm, which must be accurate enough to be useful in some context [for example, for estimating the energies of molecular states a useful level of chemical accuracy corresponds to 1 kcal/mol (Helgaker, Jorgensen, and Olsen, 2000)]. This leads to the most essential feature of QEM: the ability to minimize the noise-induced bias in expectation values on noisy hardware. However, this can also be achieved by QEC and many other long-established tools like decoherence-free subspaces and dynamical decoupling sequences (derived from optimal quantum control) (Lidar, 2014; Suter and Álvarez, 2016). Therefore, this feature alone is not sufficient to capture the QEM techniques that we cover in this review.

It is challenging to find a universally acceptable definition of quantum error mitigation. For the purposes of this review, we will define the term *quantum error mitigation* as algorithmic schemes that reduce the noise-induced bias in the expectation value by postprocessing outputs from an ensemble

of circuit runs, using circuits at the same noise level as the original unmitigated circuit or above. That is, QEM will only reduce the effective damage due to noise for the entire ensemble of circuit runs (with the help of postprocessing), but when we zoom into each individual circuit run the circuit noise level remains unchanged or even increases. This is in contrast to other techniques like QEC that aim to reduce the effect of noise on the output in every single circuit run.

Since QEM performs postprocessing using data directly from noisy hardware, it will become impractical if the amount of noise in the entire circuit is so large that it completely damages the output. In practice, this usually means that, for a given hardware setup, there is a maximum circuit size (circuit depth times qubit number) beyond which QEM will become impractical, usually due to an infeasible number of circuit repetitions. In contrast to QEC, there is no specific error threshold that one must surpass before QEM can be useful; different qubit operation error rates lead simply to different circuit sizes for which QEM will be practical. In other words, as quantum hardware continues to advance, we will be able to apply QEM to ever larger quantum circuits for more challenging applications without requiring large jumps in technology.

There are certain desirable features for error mitigation; the methods reviewed here meet the following criteria to differing extents. To begin, the mitigation method should ideally only require a modest qubit overhead to remain practical on current and near-term quantum hardware. Nevertheless, error-mitigation techniques should provide an accuracy guarantee for the method. Such a guarantee should ideally provide a formal error bound on the mitigated expectation values that indicates how well the method works at varying levels of noise. The bounds would then indicate which concrete hardware improvements would lead to improved estimates. Methods that are conceptually simple and easy to implement experimentally lead to practically feasible approaches. Last, a reliable error-mitigation method should require few assumptions (or no assumptions) about the final state that is prepared for the computation. Making strong assumptions about the final state, for example, that the state is a product state, may restrict the method to scenarios where a computational advantage over classical approaches may not be given.

We start by introducing the basic notion of QEM in Sec. II, with the details of different QEM techniques presented in Sec. III. The comparison and combinations of these individual techniques are then discussed in Sec. IV. In Sec. V we explore the application of QEM in different noise scenarios. Finally, we discuss the open problems in the field in Sec. VI and offer a conclusion in Sec. VII

II. CONCEPTS

A. Narrative introduction to concepts and terminology

In this section we introduce certain key concepts and terminology that are common to all QEM methods. However, note that the approach used in this section might not be the native way for introducing individual techniques in Sec. III. In those cases, we keep terminology that is unique to the given technique self-contained in the respective section.

Near-term quantum devices have imperfections that degrade the desired output information. A QEM protocol will aim to minimize this degradation. We use the term *primary circuit* here to describe the process that, ideally, would produce the perfect output state ρ_0 whose properties we are interested in. In practice, owing to the noise present in the primary circuit, the actual output state is some noisy state ρ instead.

Typically the ideal output information that we seek is the expectation value of some observable of interest O of the ideal output ρ_0. Commonly we would obtain this information simply by averaging the measurement results of repeated execution, as opposed to, say, some phase estimation techniques that can obtain the result through single-shot measurements but require deeper circuits that are more relevant to the fault-tolerant computing era. Therefore, even if we had ideal hardware, we would still need to perform repeated executions to determine the average. We use N_{cir} to denote the number of circuit executions, or "shots," that we employ: this includes any executions of variant circuits called for in the QEM protocol. Even in the noiseless limit, the finite N_{cir} usually implies a finite inaccuracy in our estimated average, often called shot noise. However, with perfect noiseless hardware, there would be zero bias. In other words, there would be no systematic shift to the estimated mean versus the true value (the infinite sampling limit). Given that our hardware is not perfect, there generally is finite bias. QEM protocols aim to reduce this bias, but this often means an increase in the variance (for a fixed number of circuit executions N_{cir}). One could increase N_{cir} to compensate but this cost should be acknowledged; the cost is the sampling overhead of that error-mitigation method versus the ideal noiseless case. In Secs. II.B–II.D we make these terms and concepts more precise.

B. Error-mitigated estimators

Our goal is to estimate the expectation value $\mathrm{Tr}[O\rho_0]$ of some observable of interest O. Using the outputs of the primary circuit and its variants, we can construct an estimator $\hat{O}$ for our target parameter $\mathrm{Tr}[O\rho_0]$. The quality of a given estimator can be assessed in different ways. One way is to use prediction intervals, which calculate the interval within which the outcome of the estimator will fall with a given probability, offering a rigorous bound on the worst-case deviation of the estimator. Here, however, in order to see the different factors that contribute to the deviation of the estimator more clearly, we instead focus on the expected (average-case) square deviation of our estimator $\hat{O}$ from the true value $\mathrm{Tr}[O\rho_0]$, which is called the mean square error,

$$\mathrm{MSE}[\hat{O}] = E\big[(\hat{O} - \mathrm{Tr}[O\rho_0])^2\big]. \tag{1}$$

The ultimate goal of error mitigation is to reduce $\mathrm{MSE}[\hat{O}]$ as much as possible, but this needs to be achieved using only finite resources. To quantify this, it is useful to decompose the mean square error of an estimator into two components, the bias and the variance of the estimator,

$$\mathrm{MSE}[\hat{O}] = \mathrm{bias}[\hat{O}]^2 + \mathrm{var}[\hat{O}],$$

with the bias and the variance defined as

$$\mathrm{bias}[\hat{O}] = E[\hat{O}] - \mathrm{Tr}[O\rho_0],$$
$$\mathrm{var}[\hat{O}] = E[\hat{O}^2] - E[\hat{O}]^2.$$

In this review, when we say *bias* we are sometimes referring to the magnitude of the bias $|\mathrm{bias}[\hat{O}]|$; the exact meaning should be obvious from the context.

The simplest way to construct the estimator $\hat{O}$ is by directly measuring O on the noisy output state of the primary circuit ρ, and the measurement output is denoted simply using the random variable $\hat{O}_\rho$. After running the noisy primary circuit N_{cir} times, we can take the average of these noisy outputs (obtaining the noisy sample mean) to estimate the ideal expectation value $\mathrm{Tr}[O\rho_0]$. This noisy sample mean estimator is denoted as $\bar{O}_\rho$, and its mean square error is given by

$$\mathrm{MSE}[\bar{O}_\rho] = \underbrace{(\mathrm{Tr}[O\rho] - \mathrm{Tr}[O\rho_0])^2}_{\mathrm{bias}[\bar{O}_\rho]=\mathrm{bias}[\hat{O}_\rho]} + \underbrace{\frac{\mathrm{Tr}[O^2\rho] - \mathrm{Tr}[O\rho]^2}{N_{\mathrm{cir}}}}_{\mathrm{var}[\bar{O}_\rho]=\mathrm{var}[\hat{O}_\rho]/N_{\mathrm{cir}}}. \quad (2)$$

We see that the error contribution due to the variance, which is often called shot noise, will reduce as we increase the number of circuit runs N_{cir}. In the limit of a large number of circuit executions, the mean square error $\mathrm{MSE}[\bar{O}_\rho]$ will be mainly limited by the bias of the estimator $|\mathrm{bias}[\bar{O}_\rho]|$, which is a systematic error that cannot be reduced by increasing the number of circuit runs.

To reduce the bias, we can apply QEM using data obtained from the noisy primary circuit and its variants, as discussed in Sec. III. We construct an error-mitigated estimator $\bar{O}_{\mathrm{em}}$ using the same number of circuit runs N_{cir}. We want to construct the error-mitigated estimator $\bar{O}_{\mathrm{em}}$ in such a way that it can achieve a smaller bias than the naive noisy estimator $\bar{O}_\rho$,

$$|\mathrm{bias}[\bar{O}_{\mathrm{em}}]| \leq |\mathrm{bias}[\bar{O}_\rho]|.$$

This reduction in the bias is usually achieved by constructing a more complex estimator that extracts and amplifies the useful information buried within the noise. As a result, the error-mitigated estimator is also more sensitive to the variation in the sampled data, and thus its variance will usually increase,

$$\mathrm{var}[\bar{O}_{\mathrm{em}}] \geq \mathrm{var}[\bar{O}_\rho].$$

Such a bias-variance trade-off is illustrated in Fig. 1 and can be found in almost all areas of parameter estimation. As we later see, different ways of performing error mitigation often lead to different trade-offs between bias and variance, giving the user a choice between a quickly converging QEM method with large residual error and one that is more costly but more accurate.

We can define a "one-shot" error-mitigated estimator $\hat{O}_{\mathrm{em}}$ that will satisfy $E[\hat{O}_{\mathrm{em}}] = E[\bar{O}_{\mathrm{em}}]$ and $\mathrm{var}[\hat{O}_{\mathrm{em}}] = N_{\mathrm{cir}}\mathrm{var}[\bar{O}_{\mathrm{em}}]$. The number of circuit runs needed for a given

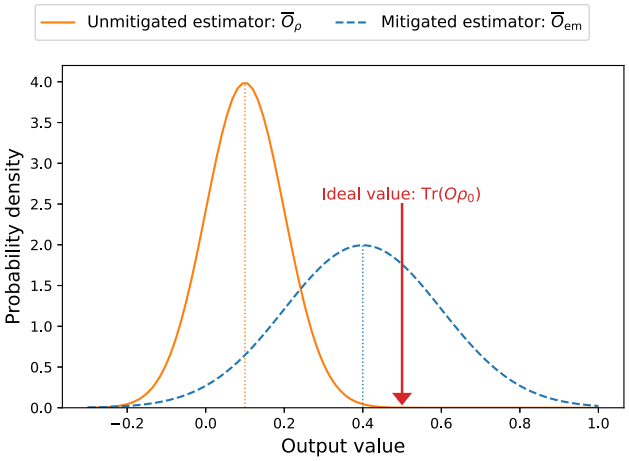

FIG. 1. Probability density distributions of the unmitigated estimator and the error-mitigated estimator. We see a decrease of bias and an increase of variance after performing error mitigation.

estimator $\hat{X}$ to achieve the shot noise level ϵ is given by $N^{\epsilon}_{\mathrm{shot}}(\hat{X}) = \mathrm{var}[\hat{X}]/\epsilon^2$. To reach the same shot noise level as the original noisy estimator, the error-mitigated estimator will require more circuit runs. This factor of increase in the number of circuit runs is called the *sampling overhead*, which is given by

$$C_{\mathrm{em}} = \frac{N^{\epsilon}_{\mathrm{shot}}(\hat{O}_{\mathrm{em}})}{N^{\epsilon}_{\mathrm{shot}}(\hat{O}_\rho)} = \frac{\mathrm{var}[\hat{O}_{\mathrm{em}}]}{\mathrm{var}[\hat{O}_\rho]}. \quad (3)$$

The sampling overhead can also be estimated using the range of the estimator through Hoeffding's inequality. The range of a random variable $\hat{X}$, denoted as $R[\hat{X}]$, is the difference between the maximum and minimum possible values taken by $\hat{X}$. Using Hoeffding's inequality, the number of samples that is sufficient to guarantee an estimation of $E[\hat{X}]$ to ϵ precision with $1-\delta$ probability is given by

$$N^{\epsilon,\delta}_{\mathrm{Hff}}(\hat{X}) = \frac{\ln(2/\delta)}{2\epsilon^2} R[\hat{X}]^2, \quad (4)$$

which can be used to estimate the sampling overhead required for the error-mitigated estimator $\hat{O}_{\mathrm{em}}$ to achieve the same ϵ and δ as the unmitigated estimator $\hat{O}_\rho$,

$$C_{\mathrm{em}} = \frac{N^{\epsilon}_{\mathrm{shot}}(\hat{O}_{\mathrm{em}})}{N^{\epsilon}_{\mathrm{shot}}(\hat{O}_\rho)} \sim \frac{N^{\epsilon,\delta}_{\mathrm{Hff}}(\hat{O}_{\mathrm{em}})}{N^{\epsilon,\delta}_{\mathrm{Hff}}(\hat{O}_\rho)} = \frac{R[\hat{O}_{\mathrm{em}}]^2}{R[\hat{O}_\rho]^2}. \quad (5)$$

C. Faults in the circuit

To gain intuition about the performance and costs of QEM, we need a way to quantify the damages due to noise in the circuit. Doing this by modeling noise in complete generality on a quantum system can be challenging (Lidar and Brun, 2013). One useful approximation that is widely employed in the field of quantum error correction is to model noise as discrete, probabilistic events named *faults* that can occur at the

various locations in the circuit including gates, idling steps, and measurements (Gottesman, 2009; Terhal, 2015). For simplicity, we assume that all locations are afflicted with Pauli noise with the error probability of location f given by p_f.[1] If we further assume that the error events in all locations are independent, then the total probability that there is no fault in the circuit is given by

$$P_0 = \prod_f (1 - p_f), \tag{6}$$

which we simply call the *fault-free probability* of the circuit. It decays exponentially with the increase in the number of fault locations in the circuit (and thus the number of qubits and the circuit depth). Note that the fault-free probability is not the fidelity of the output state, but it is a lower bound for the fidelity under our noise assumption.

We can also quantify the amount of noise in the circuit using the average number of faults in each circuit run, which is given by

$$\lambda = \sum_f p_f \tag{7}$$

and is called the *circuit fault rate*. In the simple case that all M fault locations in the circuit have the same error rate p, we then simply have $\lambda = Mp$. If the circuit contains a large number (more than dozens) of fault locations and the circuit fault rate is of the order of unity $\lambda \sim 1$, then the number of faults occurring in a given circuit run can be modeled using a Poisson distribution with mean λ using Le Cam's theorem (Le Cam, 1960; Endo, Benjamin, and Li, 2018; Cai, 2021a); i.e., the probability that ℓ faults occur in the circuit is given by $P_\ell = e^{-\lambda}\lambda^\ell/\ell!$. In this way, the fault-free probability is given by

$$P_0 = e^{-\lambda}, \tag{8}$$

which decay exponentially with the circuit fault rate. Note that for the intuitive arguments made in Sec. II.D, and indeed for general estimation about the feasibility of error mitigation, approximate estimates of P_0 and λ are often good enough to be useful.

D. Exponential scaling of the sampling overhead

We can perform bias-free QEM if we are able to postselect for the fault-free circuit runs without needing any additional circuit components. The fraction of circuit runs that are selected is simply given by the fault-free probability $e^{-\lambda}$ in Eq. (8), which means that we still require e^{λ} times more circuit runs to obtain the same number of "effective" circuit runs as a noise-free machine and achieve the same level of shot noise. Hence, even allowing for the "magical" postselection of fault-free circuit runs, the sampling overhead $C_{\mathrm{em}} = e^{\lambda}$ will still increase exponentially with the circuit fault rate (and thus the circuit size). This implies a sampling overhead of $C_{\mathrm{em}} \sim 150$ when $\lambda = 5$ and $C_{\mathrm{em}} \sim 10^4$ when $\lambda = 9$, which provides intuition on why QEM is unlikely to be efficient when the circuit fault rate is beyond $\mathcal{O}(1)$. This does not constitute a rigorous bound on the overhead of QEM.

We now move beyond trying to extract the error-free state and instead focus on obtaining the right expectation value for the observable of interest. Cai (2021a) and Wang, Fontana *et al.* (2021) showed that the expectation value of Pauli observables under Pauli gate noise is bounded by an exponential decay curve against the increase of the circuit fault rate λ. To resolve such an exponentially small quantity at large λ, we need an exponential number of samples [$C_{\mathrm{em}} = \mathcal{O}(e^{\beta\lambda})$ for some positive β]. This exponential scaling of sample overhead still applies when we consider error-mitigated estimators that are linear combinations of the output of such noisy circuits; see Appendix A.1.a.

For a given noisy circuit with the circuit fault rate λ, rather than performing active correction to reduce λ in each circuit run as in quantum error correction, QEM relies on postprocessing the outputs from an ensemble of circuit runs with the same circuit fault rate λ or above. Hence, through the aforementioned simple examples, we see that QEM cannot efficiently tackle noisy circuits with large λ on its own due to the exponential sampling overhead. However, as we later see, owing to the much lower implementation cost for QEM in terms of additional circuit components and qubits, it has become an effective means of stretching the application potential of near-term noisy devices and will be a useful tool to help alongside quantum error correction in the longer term.

III. METHODS

After introducing the overall concept of QEM, we now look at how the various error-mitigated estimators are actually constructed by performing different QEM methods.

A. Zero-noise extrapolation

In this section, we make use of noisy states obtained at different circuit fault rates. The state obtained at the circuit fault rate λ is denoted as ρ_λ. The noisy expectation value $\mathrm{Tr}[O\rho_\lambda]$ can be viewed as a function of λ. In this way, the ideal expectation value that we want is simply the value of the function at $\lambda = 0$. Trying to obtain this zero-noise value using data points at different circuit fault rates brings us to the concept of *zero-noise extrapolation* (also called error extrapolation), which was introduced by Li and Benjamin (2017) and Temme, Bravyi, and Gambetta (2017).

Using λ_1 to denote the smallest circuit fault rate that we can achieve, we can probe $\mathrm{Tr}[O\rho_\lambda]$ at a range of a boosted error rate $\{\lambda_m\}$ with $\lambda_m < \lambda_{m+1}$ to obtain a set of data points $\{(\lambda_m, \mathrm{Tr}[O\rho_{\lambda_m}])\}$. We can model $\mathrm{Tr}[O\rho_\lambda]$ using a parametrized

[1] All of the arguments will still apply if we define $1 - p_f$ as the coefficient of the error-free part of the Kraus representation of some general noise (we select the Kraus representation that gives the largest $1 - p_f$). In this case, $1 - p_f$ is not the average gate fidelity. For example, for any nonidentity unitary channel $1 - p_f$ will always be 0, which is not the case for the average gate fidelity. More generally we can always apply Pauli twirling to transform all noise to Pauli noise such that our arguments become valid.

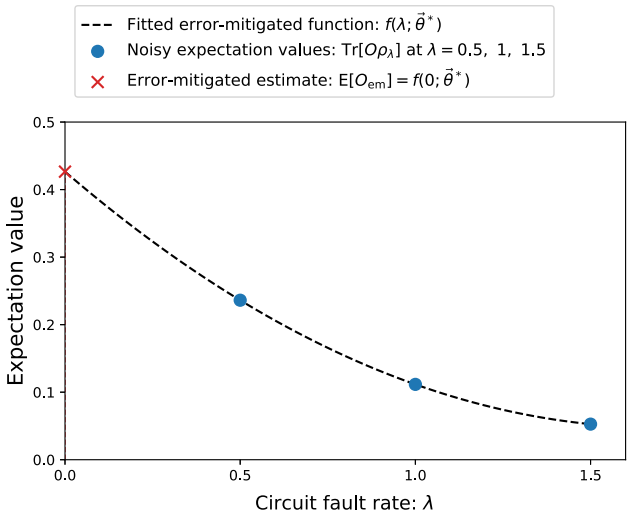

FIG. 2. Error-mitigated estimate obtained using zero-noise extrapolation. We perform extrapolation using three noisy expectation values at the circuit fault rate of $\lambda = 0.5$, 1, and 1.5, where 0.5 is the lowest circuit fault rate that we can achieve and the other two values are obtained by boosting the noise in the device.

function $f(\lambda;\vec{\theta})$ and fit it to the data points to obtain a set of optimal parameters $\vec{\theta}^*$. The error-mitigated estimate of the zero-noise output $\mathrm{Tr}[O\rho_0]$ is then given by

$$\mathrm{E}[\hat{O}_{\mathrm{em}}] = f(0;\vec{\theta}^*).$$

A simple illustration of zero-noise extrapolation is shown in Fig. 2.

If λ is small, Temme, Bravyi, and Gambetta (2017) showed that $\mathrm{Tr}[O\rho_\lambda]$ can be approximated using a polynomial function in the same spirit as a truncated Taylor expansion,

$$\mathrm{Tr}[O\rho_\lambda] \approx f(\lambda;\vec{\theta}) = \sum_{\ell=0}^{M-1} \theta_\ell \frac{\lambda^\ell}{\ell!}. \tag{9}$$

In Eq. (9) we have a polynomial of degree $M-1$ that has M different free parameters, and the zero-noise estimate that we want is $\theta_0^* = f(0;\vec{\theta}^*)$. The simplest case is linear extrapolation with $M = 2$ (Li and Benjamin, 2017). If we try to fit Eq. (9) with M data points, which is the minimal number of data points needed, we can perform Richardson extrapolation as discussed by Temme, Bravyi, and Gambetta (2017). The corresponding error-mitigated estimate obtained using the set of data points $\{(\lambda_m, \mathrm{Tr}[O\rho_{\lambda_m}])\}$ is (Giurgica-Tiron *et al.*, 2020)

$$\mathrm{E}[\hat{O}_{\mathrm{em}}] = \theta_0^* = \sum_{m=1}^{M} \mathrm{Tr}[O\rho_{\lambda_m}] \prod_{k\neq m} \frac{\lambda_k}{\lambda_k - \lambda_m}. \tag{10}$$

The bias in our estimate is due mostly to our omission of the higher degree terms in the polynomial approximation, and thus we should expect that $\mathrm{bias}[\hat{O}_{\mathrm{em}}] = \mathcal{O}(\lambda^M)$.

The error-mitigated expectation value in Eq. (10) is a linear combination of the set of noisy expectation values $\{\mathrm{Tr}[O\rho_{\lambda_m}]\}$; thus, it can be estimated using the Monte Carlo sampling method (see Appendix A.1), and the corresponding sampling overhead for Richardson extrapolation is given by

$$C_{\mathrm{em}} \sim \left(\sum_{m=1}^{M} \left|\prod_{k\neq m} \frac{\lambda_k}{\lambda_k - \lambda_m}\right|\right)^2. \tag{11}$$

We see that if any of the probed circuit fault rates λ_m are too large or the gap between any two data points $|\lambda_m - \lambda_k|$ is too small, then C_{em} will blow up and the extrapolation will become infeasible. For the simple case of equal-gap Richardson extrapolation ($\lambda_m = m\lambda_1$), the sampling overhead in Eq. (11) becomes $C_{\mathrm{em}} \sim (2^M - 1)^2$, which grows exponentially with the number of data points M. The example here is mainly for illustrating the scaling behavior of the sampling overhead. In practice, Richardson extrapolation usually takes a more general data gap Δ with data points at $\lambda_m = \lambda_1 + (m-1)\Delta$. Even more sophisticated Richardson extrapolation beyond the equal-gap variant is possible, which can reduce the sampling overhead (Krebsbach, Trauzettel, and Calzona, 2022).

As mentioned, theoretically Richardson extrapolation will be valid only for small λ due to the approximation that we made in Eq. (9). However, a recent experiment by Kim, Wood *et al.* (2023) showed that Richardson extrapolation can be effective at large λ in practice. They performed a 26-qubit simulation of the 2D transverse field Ising model in the limit of strong entanglement and showed that the error-mitigated evolution of the magnetization is competitive in comparison to standard tensor network methods. To look for an extrapolation method that is naturally compatible with large λ, we can consider a large circuit with Pauli noise. Such a circuit in the limit of $\lambda \to \infty$ will become a random circuit with zero expectation value for bounded traceless observables, as mentioned in Sec. II.D. A polynomial extrapolation function diverges at $\lambda \to \infty$ and thus does not fit the aforementioned intuition, which leads to extrapolation schemes that use an exponential decay curve or even a multiexponential decay curve instead. Exponential extrapolation has proven to be able to achieve smaller biases than Richardson and more generally polynomial extrapolation in some numerical simulations and cloud experiments (Endo, Benjamin, and Li, 2018; Giurgica-Tiron *et al.*, 2020), especially for Pauli noise (Cai, 2021a). Note that, going beyond Richardson extrapolation, the error-mitigated estimator obtained through least-squares fitting often does not have a closed-form representation, and thus there are not yet analytical expressions for their biases and variance in most of the cases. It is possible to combine exponential and Richardson extrapolation by performing Richardson extrapolation on the function $e^{\lambda}\mathrm{Tr}[O\rho_\lambda]$ instead of $\mathrm{Tr}[O\rho_\lambda]$, which can give explicit bounds on the sampling overheads and biases (Cai, 2021b).

When we try to boost the noise in the circuit, we must know the exact factor of increase in the circuit fault rate. Ideally, we want to increase the error strength in the various fault locations without changing their error models, which can be challenging to implement in experiments. Temme, Bravyi, and Gambetta (2017) showed that under the assumption that the noise was time invariant, the noise could be effectively amplified by stretching and recalibrating the gate pulses, which was demonstrated on a superconducting platform (Kandala *et al.*, 2019). When targeting only single-qubit gate errors, these

experiments can successfully perform Richardson extrapolation up to the fourth order, highlighting the accuracy of the noise amplification. Rescaling noise for two-qubit cross-resonance gates in these architectures can be more challenging due to reasons like more complicated drive Hamiltonians and drive-dependent coherence times (Chow *et al.*, 2011). Nevertheless, pulse-stretching experiments with slowed-down two-qubit cross-resonance gates have been shown to be effective with linear extrapolation (Kandala *et al.*, 2019), most recently demonstrated in experiments with up to 26 qubits and a depth of 120 (Kim, Wood *et al.*, 2023).

Alternatively, Dumitrescu *et al.* (2018), Giurgica-Tiron *et al.* (2020), and He *et al.* (2020) tried to boost the circuit fault rate by inserting a sequence of abundant gates that are equivalent to the identity if operated noiselessly. This is easier to implement and calibrate than the pulse-stretching method. However, while this will work well with depolarizing gate noise, it may change the gate error model for more general gate noise such that the factor of increase in the error strength is no longer well defined (Kim, Wood *et al.*, 2023). To address this, Henao, Santos, and Uzdin (2023)replaced the gate-based inverse with a pulse-based inverse that is applicable beyond depolarizing noise. They also adapted the weights of the noise-amplified circuits to the strength of the noise in order to extend beyond the weak noise assumption in Richardson extrapolation. Note that for both pulse stretching and gate insertion, an M-time increase in the noise strength might lead to an up to M-time increase in the circuit run-time. If the error models at the various fault locations are completely known, Li and Benjamin (2017) suggested that it is possible to controllably amplify the noise by probabilistically inserting gates to simulate the faults. In fact, as noted first by Cai (2021a) and later by Mari, Shammah, and Zeng (2021), it would be more efficient to use these probabilistically inserted gates to perform probabilistic error cancellation instead (Sec. III.B), which yields new data points at reduced error strength. Learning a representative noise model can be challenging for large-scale devices, especially for correlated noise.

Owing to the simplicity of zero-noise extrapolation, it is one of the most widely implemented QEM methods. Recently Kim *et al.* (2023) simulated the time dynamics of the Ising model up to a circuit size of 127 qubits and 60 layers of two-qubit gates, and with the help of zero-noise extrapolation they are able to produce results in agreement with state-of-the-art classical simulations for that circuit size (Begušić and Chan, 2023; Kechedzhi *et al.*, 2023; Tindall *et al.*, 2023). In addition to the aforementioned experiments, it was also successfully demonstrated in a wide range of other experiments, especially through cloud platforms (Klco *et al.*, 2018; Garmon, Pooser, and Dumitrescu, 2020; Keen *et al.*, 2020; Tacchino *et al.*, 2020; Yeter-Aydeniz, Pooser, and Siopsis, 2020).

B. Probabilistic error cancellation

An alternative quantum error-mitigation method referred to as probabilistic error cancellation was introduced by Temme, Bravyi, and Gambetta (2017). A particular feature of this method is that it can fully remove the bias of expectation values of generic quantum circuits ($\mathrm{bias}[\hat{O}_{\mathrm{em}}] = 0$). This comes at the expense of a sampling overhead C_{em} that grows exponentially with the circuit fault rate λ. The key idea is noting that the noise-free expectation value can be written as a linear combination of expectation values from a set of noisy quantum circuits. This real-valued, linear combination can be interpreted as a *quasiprobability decomposition* that can be sampled (cf. Appendix A.1), according to a Monte Carlo procedure. In this review we formalize the method using the superoperator representation (Gilchrist, Terno, and Wood, 2011), in which density matrices ρ are vectorized into $|\rho\rangle\!\rangle$ and quantum channels are written as matrices acting on $|\rho\rangle\!\rangle$, denoted using a script font like $\mathcal{U}$. Taking the trace with the observable O is then written as the inner product with the vectorized observable $\mathrm{Tr}[O\rho] = \langle\!\langle O|\rho\rangle\!\rangle$. In this section, we employ this representation over the standard density matrix formalism for clearer representations of the linear combination and decomposition of a set of noisy quantum circuits.

To construct an estimator for an ideal channel $\mathcal{U}$ from noisy operations, we need to choose a set of noisy basis operations $\{\mathcal{B}_n\}$ that we can implement in the physical hardware. These operations are, for example, noisy gates, state preparation operations, and measurements. These operations are assumed to be learned from the noisy hardware in experiments through some form of tomography. An example of a complete set of basis operations was discussed by Endo, Benjamin, and Li (2018). With a sufficiently large basis, we can decompose the ideal operation into

$$\mathcal{U} = \sum_n \alpha_n \mathcal{B}_n, \tag{12}$$

with real coefficients α_n. Some of the coefficients α_n in Eq. (12) can be negative, which means that this decomposition does not necessarily correspond to a probabilistic mixture of physical maps. The expansion therefore is often not a physical map that can be implemented directly. However, if we are applying $\mathcal{U}$ on some input state ρ_{in} in order to measure some observable O and obtain the ideal expectation value $\mathrm{Tr}[O\rho_0] = \langle\!\langle O|\rho_0\rangle\!\rangle$, then the ideal expectation value can be decomposed into

$$\langle\!\langle O|\rho_0\rangle\!\rangle = \langle\!\langle O|\mathcal{U}|\rho_{\mathrm{in}}\rangle\!\rangle = \sum_n \alpha_n \langle\!\langle O|\mathcal{B}_n|\rho_{\mathrm{in}}\rangle\!\rangle, \tag{13}$$

i.e., the ideal expectation value $\langle\!\langle O|\mathcal{U}|\rho_{\mathrm{in}}\rangle\!\rangle$ can be decomposed into a linear combination of noisy expectation values $\{\langle\!\langle O|\mathcal{B}_n|\rho_{\mathrm{in}}\rangle\!\rangle\}$ that we can obtain individually.

When the expansion in Eq. (13) has many terms, the linear combination of noisy expectation values in Eq. (13) can be estimated using the Monte Carlo sampling method (Appendix A.1). The method samples different basis operations $\mathcal{B}_n$ according to their weight in the expansion (i.e., the noisy circuit corresponding to $\mathcal{B}_n$ is chosen with the probability $|\alpha_n|Q^{-1}$, where $Q = \sum_n |\alpha_n|$), and the circuit output is multiplied by $\mathrm{sgn}(\alpha_n)Q$ before being used to estimate the error-mitigated expectation value $E[\hat{O}_{\mathrm{em}}]$. This multiplicative factor leads to an increase in the variance and, correspondingly, to a sampling overhead [given by Eq. (A4)]

$$C_{\mathrm{em}} \sim Q^2 = \left(\sum_n |\alpha_n|\right)^2. \tag{14}$$

In practice, we can often implement the noisy version of the target operation, denoted as $\mathcal{U}_p$, which can be written as $\mathcal{U}_p = (1-p)\mathcal{U} + p\mathcal{N}$, where p is the operation error rate and $\mathcal{N}$ is the noise element. It can be rewritten as

$$\mathcal{U} = \frac{1}{1-p}\mathcal{U}_p - \frac{p}{1-p}\mathcal{N}. \tag{15}$$

Comparing Eq. (15) to Eq. (12), we see that $\mathcal{U}_p$ is one of the basis operations $\mathcal{U}_p = \mathcal{B}_1$. In the simple case that $\mathcal{N}$ is also one of the basis operations that we can implement $\mathcal{N} = \mathcal{B}_2$, the sampling overhead for removing the noise in $\mathcal{U}_p$ is given by

$$C_{\mathrm{em}} \sim Q^2 = \left(\frac{1}{1-p} + \frac{p}{1-p}\right)^2 = \left(\frac{1+p}{1-p}\right)^2. \tag{16}$$

More generally $\mathcal{N}$ must be decomposed into the rest of the basis $\mathcal{N} = \sum_{n\neq 1} w_n \mathcal{B}_n$. For the common scenario in which $\mathcal{U}_p$ suffers from Pauli noise and we can perform high-fidelity Pauli gates to be used as our basis $\{\mathcal{B}_n\}$, the sampling overhead is still given by Eq. (16).

Thus far we have used the basis set to cancel out the noisy component $\mathcal{N}$ in $\mathcal{U}_p = (1-p)\mathcal{U} + p\mathcal{N}$. Alternatively, there is another means of decomposition to effectively "invert" the noise channel $\mathcal{E}_p$ in $\mathcal{U}_p = \mathcal{E}_p\mathcal{U}$ as discussed by Temme, Bravyi, and Gambetta (2017) and Endo, Benjamin, and Li (2018), and its resultant sampling overhead is similar. However, this noise-inversion implementation requires inserting additional operations after every noisy operation, and thus we might need up to twice the original circuit run-time.

To perform the previously described error cancellation, we need to have the full description of the noisy operation $\mathcal{U}_p$. This can be efficiently characterized only for individual local gates. Any circuit we want to implement can be decomposed into a sequence of M ideal gates $\{\mathcal{U}_m\}$, and the ideal expectation value to obtain will be

$$\langle\langle O|\rho_0\rangle\rangle = \langle\langle O|\prod_{m=1}^{M}\mathcal{U}_m|\rho_{\mathrm{in}}\rangle\rangle.$$

As in Eq. (13), we decompose the individual gates into the noisy basis using Eq. (12),

$$\begin{aligned}\langle\langle O|\rho_0\rangle\rangle = E[\hat{O}_{\mathrm{em}}] &= \langle\langle O|\prod_{m=1}^{M}\left(\sum_{n_m}\alpha_{mn_m}\mathcal{B}_{n_m}\right)|\rho_{\mathrm{in}}\rangle\rangle \\ &= \sum_{\vec{n}}\alpha_{\vec{n}}\langle\langle O|\mathcal{B}_{\vec{n}}|\rho_{\mathrm{in}}\rangle\rangle,\end{aligned} \tag{17}$$

where $\vec{n} = \{n_1, n_2, \ldots, n_M\}$ is the set of labels for a sequence of basis elements and we have defined $\mathcal{B}_{\vec{n}} = \prod_{m=1}^{M}\mathcal{B}_{n_m}$, $\alpha_{\vec{n}} = \prod_{m=1}^{M}\alpha_{mn_m}$, and $Q = \sum_{\vec{n}}|\alpha_{\vec{n}}|$. Note that the set of noisy basis elements $\{\mathcal{B}_n\}$ here includes the basis for all the gates in the circuit; it is thus overcomplete and contains the noisy version of all of the target gates. Again we can obtain samples of $\hat{O}_{\mathrm{em}}$ using Monte Carlo sampling as previously discussed with the label of a single basis n replaced by the label of a sequence of basis $\vec{n}$.

The overall sampling overhead of mitigating the errors in the entire circuit is simply the product of the sampling overhead of each gate. As explored by Temme, Bravyi, and Gambetta (2017) and subsequently by Endo, Benjamin, and Li (2018), if we assume all of the gates suffer from Pauli noise with the same error rate p, the circuit size M is large and the circuit fault rate $\lambda = Mp$ is finite, then taking the product of the gate-level sampling overhead in Eq. (16) gives

$$C_{\mathrm{em}} = \prod_{m=1}^{M}\left(\frac{1+p}{1-p}\right)^2 \approx e^{4Mp} = e^{4\lambda}. \tag{18}$$

In fact, the overhead here is still valid even if the gates in the circuit have different error rates, as long as the circuit faults follow a Poisson distribution; see Sec. II.C. More generally, as discussed first by Cai (2021a) and later by Mari, Shammah, and Zeng (2021), it is possible to apply probabilistic error cancellation only partially. When the resultant circuit fault rate is denoted as λ_{em}, the corresponding sampling cost is simply $C_{\mathrm{em}} = e^{4(\lambda-\lambda_{\mathrm{em}})}$, which grows exponentially with the reduction in the circuit fault rate.

As mentioned, to be able to perform the decomposition in Eq. (12) we need to have a set of basis elements that span the ideal operations, and full characterization of this basis. Endo, Benjamin, and Li (2018) showed that such a basis can be constructed using single-qubit Clifford gates and Z measurements with reasonable fidelity, while the characterization can be carried out efficiently using gate set tomography. In practice we usually supplement this set of basis elements with the noisy version of the target operation, such that it is overcomplete as discussed. Only noisy operations with local noise can be efficiently characterized using the aforementioned protocol.

A recent experimental implementation of the probabilistic error cancellation method used a correlated Pauli-noise model that is supported over the full gate layer on the device (Van den Berg *et al.*, 2023). The noise model is given in terms of a sparse Pauli Lindbladian $\mathcal{L}(\rho) = \sum_k \lambda_k(P_k\rho P_k - \rho)$ for a set of Pauli matrices P_k, which can be efficiently learned for a polynomial number of Pauli terms (usually low-weight Pauli terms). Although the noise model is correlated across the full circuit, it can be inverted efficiently and its quasiprobability distribution can be sampled exactly. A similar approach was also adopted by Ferracin *et al.* (2022), who used cycle benchmarking to characterize the low-weight correlated Pauli-noise components for performing probabilistic error cancellation. It is also possible to mitigate these correlated noise components by performing noise characterization using matrix product operators (Guo and Yang, 2022) or with the help of learning-based methods (Strikis *et al.*, 2021), as discussed in Sec. III.H.

From Eqs. (12) and (14), we see that the sampling overhead of probabilistic error cancellation is highly dependent on the basis that we choose. For the standard basis proposed by Endo, Benjamin, and Li (2018), it will perform well when the gates in the circuit suffer from Pauli noise (which can be achieved via Pauli twirling). Going beyond that,

Takagi (2021) derived a lower bound for the sampling overhead under general noise models. Such a lower bound on the sampling overhead has proven to be a good measure for many properties of the noise channel that we try to mitigate (Jiang, Wang, and Wang, 2021; Regula, Takagi, and Gu, 2021; Guo and Yang, 2023), but the basis required to reach this lower bound may not be implementable using the given hardware. To find a better practical basis beyond the standard basis, Piveteau, Sutter, and Woerner (2022) proposed using variational circuits to construct the basis and numerically tested this using real hardware noise models. For actual experiments, probabilistic error cancellation was successfully demonstrated by Song *et al.* (2019) and Zhang *et al.* (2020) on superconducting and trapped-ion platforms. Sun *et al.* (2021) showed that probabilistic error cancellation can also be applied to continuous noise processes, for which partial mitigation is possible by expanding the noise process into a perturbation series (Hama and Nishi, 2022). It is also possible to use the non-Markovianity in noise to reduce the sampling cost (Hakoshima, Matsuzaki, and Endo, 2021).

C. Measurement error mitigation

Depending on the error-mitigation procedure, it is necessary to make a distinction between the types of errors that occur during a calculation. Errors that occur during the state preparation and measurement (SPAM) stage are referred to as SPAM errors (Merkel *et al.*, 2013; Lin *et al.*, 2021). The error that occurs at the final measurement stage introduces an additional bias in the expectation value of interest. To put this into a more concrete form, we continue to use the superoperator representation introduced in Sec. III.B. When we perform measurements and obtain the binary string $x \in \{0,1\}^N$ as the output, ideally we want to perform projective measurements in the computational basis $\{\langle\langle x|\}$. However, some measurement noise $\mathcal{A}$ might occur and transform the projective measurements into some positive operator-value measures (POVM) $\{\langle\langle E_x|\} = \{\langle\langle x|\mathcal{A}\}$, leading to a different output statistic.

The origins of the measurement errors are as diverse as the hardware that is used to implement quantum processors. For example, a dominant error in the measurement of superconducting qubits is due to thermal excitations and T_1 decay (Blais *et al.*, 2004; Wallraff *et al.*, 2004), while for ion traps a major source of uncertainty arises from the difficulty of detecting an ion's dark state and collisions in the trap (Bergquist *et al.*, 1986; Nagourney, Sandberg, and Dehmelt, 1986; Sauter *et al.*, 1986). Other architectures experience noise in the measurement stage from different sources (Haroche and Raimond, 2006). From the perspective of the measurement error protocols most frequently considered in the literature, it is sufficient to consider a simplified model that is agnostic regarding the actual origin of the noise. This model makes the assumption that the noise channel $\mathcal{A}$ has the computational subspace $\{\langle\langle x||x \in \{0,1\}^N\}$ as its invariant subspace; i.e., the resultant POVM basis $\{\langle\langle E_x|\}$ lives within the computational subspace. This is not the most general measurement error model, but it is nonetheless the most frequently considered model, as other coherent errors are usually assumed to be part of the computational stage instead of the measurement stage. Under this assumption, the POVM $\langle\langle E_x|$ can be decomposed into the computational basis,

$$\langle\langle E_x| = \sum_y \langle\langle E_x|y\rangle\rangle\langle\langle y| = \sum_y \langle\langle x|\mathcal{A}|y\rangle\rangle\langle\langle y| = \sum_y A_{xy}\langle\langle y|. \quad (19)$$

In Eq. (19) the assignment matrix A is a transition matrix (stochastic matrix) whose entries $\langle\langle x|\mathcal{A}|y\rangle\rangle$ represent the transition probability from the measurement result y to x due to the noise channel $\mathcal{A}$ (Chow *et al.*, 2010). The entry $A_{xy} = \langle\langle x|\mathcal{A}|y\rangle\rangle = \langle\langle E_x|y\rangle\rangle$ can be obtained by estimating the probability of the x outcome when we prepare the computational state $|y\rangle\rangle$ (assumed to be almost perfect) and perform the set of noisy measurements $\{\langle\langle E_x|\}$. If A is full rank, then we can invert Eq. (19) and obtain

$$\langle\langle y| = \sum_x A^{-1}_{yx}\langle\langle E_x|. \quad (20)$$

That is, we can simulate the behavior of the ideal measurement $\{\langle\langle y|\}$ using a linear combination of the noisy measurements $\{\langle\langle E_x|\}$, just as we did in probabilistic error cancellation in Sec. III.B. Hence, the associated sampling overhead will increase exponentially with the measurement fault rate, as with Eq. (18) for probabilistic error cancellation.

For an incoming state $|\rho\rangle\rangle$, the output distribution using the ideal measurements is given by the vector $\vec{p}_0 = \{\langle\langle y|\rho\rangle\rangle\}$, while the output distribution using the noisy measurements is $\vec{p}_{\text{noi}} = \{\langle\langle E_x|\rho\rangle\rangle\}$. Applying Eqs. (19) and (20) on $|\rho\rangle\rangle$, we then have

$$\vec{p}_{\text{noi}} = A\vec{p}_0 \Rightarrow \vec{p}_0 = A^{-1}\vec{p}_{\text{noi}}. \quad (21)$$

For a given observable O with the spectrum $\vec{O} = \{O_x\}$, i.e., $\langle\langle O| = \sum_x O_x\langle\langle x|$, its ideal expectation value is

$$\langle\langle O|\rho\rangle\rangle = \vec{O}^T\vec{p}_0 = \vec{O}^T A^{-1}\vec{p}_{\text{noi}}. \quad (22)$$

Hence, the ideal expectation value can be obtained once we know the assignment matrix A and the noisy output distribution $\vec{p}_{\text{noi}}$.

In early experiments with only a few qubits, Kandala *et al.* (2017) performed a full readout tomography of the noisy output distribution $\vec{p}_{\text{noi}}$ in the computational basis. All entries of the assignment matrix A were then estimated, which can be used to calculate the ideal expectation value using Eq. (22). This approach is not efficiently scalable as the size of A scales exponentially with the number of qubits N.

The simplest way to tackle the problem is to assume that the measurement errors of different qubits are not correlated, which implies that the assignment matrix is simply the tensor product of the assignment matrices of the individual qubits $A = \bigotimes_{n=1}^{N} A_n$. However, realistic noise encountered in experiments may not be captured accurately by this simplified model, and it can be observed that correlations between individual bit flips are in fact present (Heinsoo *et al.*, 2018).

Bravyi *et al.* (2021) tried to construct the assignment matrix using continuous-time Markov processes with the generators

(transition rate matrices) $\{G_i\}$ being single- and two-qubit operators,

$$A = e^G, \quad \text{with} \quad G = \sum_{i=1}^{2N^2} r_i G_i. \tag{23}$$

The assignment matrix A is now determined by $2N^2$ positive coefficients $\{r_i\}$ that can be learned by only considering a polynomial number of input bit strings. Once the coefficients are learned, we can easily construct the inverse matrix $A^{-1} = e^{-G}$ for error mitigation.

As mentioned in Appendix A.2, we can perform Pauli twirling on the noise channel $\mathcal{A}$ by conjugating it with random Pauli operators, which will remove all the off-diagonal elements of $\mathcal{A}$ in the Pauli basis and produce a Pauli channel $\mathcal{D}$. This can be used to simplify measurement error mitigation, as discussed by Chen *et al.* (2021) and Van den Berg, Minev, and Temme (2022). Since we are interested only in the action of $\mathcal{A}$ on the computational subspace, we need to consider only Pauli basis elements that are the tensor products of Z denoted as $\{\langle\langle Z^x|\}$, with x as the bit string that marks the qubits that are acted on by Z. Ideally we want to perform the Pauli measurement $\langle\langle Z^x|$; however, we can perform only the noisy measurement $\langle\langle Z^x|\mathcal{A}$. Using Pauli twirling, we can transform the noisy measurement into

$$\langle\langle Z^x|\mathcal{D} = D_x \langle\langle Z^x| \tag{24}$$

using the fact that the twirled channel $\mathcal{D}$ is diagonal in the Pauli basis, with the entries being $D_x = \langle\langle Z^x|\mathcal{D}|Z^x\rangle\rangle = \langle\langle Z^x|\mathcal{A}|Z^x\rangle\rangle$. Thus, the noisy measurement $\langle\langle Z^x|\mathcal{D}$ is simply the ideal measurement $\langle\langle Z^x|$ rescaled by a factor D_x. For a given input state $|\rho\rangle\rangle$, we have

$$\underbrace{\langle\langle Z^x|\rho\rangle\rangle}_{\text{ideal}} = D_x^{-1} \underbrace{\langle\langle Z^x|\mathcal{D}|\rho\rangle\rangle}_{\text{twirled noisy}}.$$

Hence, by transforming the observable into Z^x and performing Pauli twirling, we need to rescale the noisy expectation value only by a factor D_x^{-1} to obtain the ideal expectation value. Note that, since conjugation with Z will have trivial effects within the computational subspace, we need only conjugate the noise channel with a random operator in $\{I, X\}^{\otimes N}$ (i.e., random bit flips) to achieve the aforementioned effect of Pauli twirling.

In practice, the ideal output distribution $\vec{p}_0$ is often sparse. With weak measurement noise, we would expect the corresponding noisy output distribution $\vec{p}_{\text{noi}}$ to also be sparse and to have nonzero probability at all positions that are nonzero in $\vec{p}_0$. Using this fact, Nation *et al.* (2021) proposed focusing on the action of measurement noise $\mathcal{A}$ within the subspace spanned by the basis of $\vec{p}_{\text{noi}}$ with nonzero probability, which gives an assignment matrix with a much smaller dimension. The ideal output distribution $\vec{p}_0$ can then be obtained by inverting the assignment matrix within this subspace. The inversion can be sped up by considering a matrix-free preconditioned iteration algorithm.

Owing to sampling noise, the estimation of the ideal distribution $\vec{p}_0$ obtained using matrix inversion in Eq. (21) may contain negative values and thus is not a valid probability distribution. However, one is still able to provide an unbiased expectation value estimate using Eq. (22) (Bravyi *et al.*, 2021). However, instead of using matrix inversion, one might try to solve a constrained optimization problem with the cost function $\|A\vec{p}_0 - \vec{p}_{\text{noi}}\|_2^2$ such that $\vec{p}_0$ is a valid probability distribution. This can be solved using maximum likelihood (Chen *et al.*, 2019; Geller, 2020; Maciejewski, Zimborás, and Oszmaniec, 2020) or iterative Bayesian unfolding (Nachman *et al.*, 2020).

There is a wide range of other techniques for combating measurement errors. Hamilton *et al.* (2020) proposed a representation on cumulant expansion to capture correlations between observables. Kwon and Bae (2021) looked into the use of Clifford twirling in the context of measurement error mitigation. Measurement error protocols that are directly tailored to calculations of the variational quantum eigensolver (VQE) type are possible (Barron and Wood, 2020). Other approaches have proposed the use of final premeasurement entangling circuits to combat noise (Hicks *et al.*, 2022). Tannu and Qureshi (2019) proposed exploiting a potential asymmetry in the noise strength of the assignment matrix A by flipping bit values into a configuration less likely to be affected by noise. The experimental observations have been used to train classical neural networks to infer predictions of the correct expectation values (Palmieri *et al.*, 2020). As measurement noise is a major obstacle in almost all experimental setups, implementation of measurement error mitigation is almost ubiquitous in all near-term experiments.

D. Symmetry constraints

A simple but effective scheme for suppressing errors is to identify errors that break the symmetries of the ideal quantum state and remove them via postselection. This notion originates from quantum error correction (Gottesman, 1997; Terhal, 2015), in which we explicitly define a set of measurements to detect and correct all local errors at the cost of additional qubit overhead. Though explicitly correcting errors is required for scalability (Shor, 1996), *quantum error detection* of artificially added symmetries has been widely recognized as an important milestone toward this end goal (Nigg *et al.*, 2014; Córcoles *et al.*, 2015; Kelly *et al.*, 2015; Gottesman, 2016; Linke *et al.*, 2017). In practical applications, quantum circuits often possess inherent symmetries that can be used for error mitigation without the need to execute a quantum circuit on an error-detection code. Measuring these inherent symmetries and discarding circuit runs that produce the wrong results produces a postselected state ρ_{sym}. The broad class of schemes that directly or indirectly measure $\text{Tr}[O\rho_{\text{sym}}]$ are known collectively as symmetry verification (Bonet-Monroig *et al.*, 2018; McArdle, Yuan, and Benjamin, 2019). If the symmetry measurements are perfect, ρ_{sym} must have nondecreasing overlap on the ideal state ρ_0 compared to the noisy state ρ, as we have thrown away states with zero overlap. In practice, symmetry measurements may themselves introduce errors into the state; thus, choosing the right

symmetries and optimizing their measurements are at the core of symmetry verification.

Bonet-Monroig *et al.* (2018) and McArdle, Yuan, and Benjamin (2019) proposed various easily accessible symmetry operators S for symmetry verification, drawing from those that naturally emerge in physical systems. In this context, given a physical system its symmetry operators S are operators that commute with the system Hamiltonian H: $[H,S]=HS-SH=0$. When this is the case, H and S may be simultaneously diagonalized, i.e., energy eigenstates $|\Psi_j\rangle$ can be chosen in such a way that $S|\Psi_j\rangle=s|\Psi_j\rangle$, where s is an eigenvalue of S. Thus, measuring S on a prepared quantum state, and postselecting on the correct symmetry eigenvalue s for the target energy eigenstate should project one closer to said energy eigenstate. Furthermore, time evolution by e^{iHt} leaves the eigenspaces of S invariant, which means that dynamic properties of the physical system can be studied entirely within these eigenspaces as well. Common examples of symmetries are the parity $\prod_i Z_i$ and the Z component of the total spin $\sum_i Z_i$ of a spin system, or the particle number $\sum_i n_i$ of a fermionic system. Note that the Jordan-Wigner transformation maps $\sum_i n_i$ to $\sum_i Z_i$ modulo a constant shift. Given an N-qubit simulation, the eigensubspaces of these symmetries have dimension 2^{N-1}, $\binom{N}{(N-s)/2}$, and $\binom{N}{s}$, respectively, with s as the eigenvalue of the given eigensubspace. With the right s, these subspaces are large enough to execute classically intractable quantum algorithms. Note, though, that it is not necessary to enforce symmetries during an entire circuit in order to verify them at the end (Dallaire-Demers *et al.*, 2019). Even when one employs a circuit that does not conserve the symmetry of the target physical problem, symmetry verification can still be used for projecting back into the appropriate spin or number sector (Yen, Lang, and Izmaylov, 2019; Tsuchimochi, Mori, and Ten-no, 2020; Khamoshi, Evangelista, and Scuseria, 2021), which can be viewed as mitigating algorithmic errors (Sec. V.C).

As mentioned, the process of direct symmetry verification is carried out by measuring both the symmetry operators S and the target observable O in every circuit run and discarding runs that produce the wrong output for S (i.e., that fail the symmetry check). Since symmetries S are typically global observables, measuring them alongside the target observable O is nontrivial. The additional circuit components required for their measurements can introduce additional errors, reducing or even nullifying the effect of error mitigation. Various ways to measure multiple operators (such as the symmetry operator and the target observable) in the same circuit run are discussed in Appendix A.3. One way is to use a Hadamard test to measure a Pauli symmetry like the number parity (Bonet-Monroig *et al.*, 2018; McArdle, Yuan, and Benjamin, 2019). Other practical schemes involve measuring qubitwise commuting operators (Izmaylov, Yen, and Ryabinkin, 2019), i.e., when both S and O can be obtained through postprocessing the same set of single-qubit Pauli measurements across all qubits; see Appendix A.3. This implies that if we are using only single-qubit rotations and readout to perform the measurements, symmetries such as the Z component of the total spin $\sum_i Z_i$ or parity $\prod_i Z_i$ cannot be measured simultaneously with any operator that is not diagonal in the computational basis.

The issue of qubitwise commutativity presents a specific problem in chemistry, where the fermion hopping operator $c_i^\dagger c_j + c_j^\dagger c_i$ (our target observable) commutes with the particle number $\sum_i n_i$ (the symmetry operator), but not qubitwise. An identical problem presents in spin physics between the operators $X_iX_j + Y_iY_j$ and $\sum_i Z_i$. This can be solved by noting that the rotation

$$\exp\left[i\frac{\pi}{8}(X_iY_j - Y_jX_i)\right] \tag{25}$$

maps $X_iX_j + Y_iY_j$ to $Z_i - Z_j$ while leaving $Z_i + Z_j$ invariant (Huggins, McClean *et al.*, 2021). In terms of fermionic systems, this is the operator $e^{i(\pi/2)\mathrm{FSWAP}_{i,j}}$, where FSWAP $= c_i^\dagger c_j + c_j^\dagger c_i - n_i - n_j$ (Google Quantum AI *et al.*, 2020a). In a spin system, this allows for the joint measurement of hopping terms and the Z component of the total spin using only a constant depth circuit (assuming all-to-all coupling). However, this measurement does not parallelize efficiently: one can measure hopping terms simultaneously only between disjoint pairs of qubits, requiring $O(N)$ distinct measurements to estimate all spin-spin hopping terms simultaneously with the Z component of the total spin. In contrast, without the requirement to simultaneously measure the total spin-Z component, all spin-spin hopping terms can be estimated by only two distinct choices of single-qubit rotation and readout. In a fermionic system, we need $O(N)$ distinct measurements in the first place for estimating all fermionic hopping terms due to the lack of mutually commuting terms (Bonet-Monroig, Babbush, and O'Brien, 2020), so simultaneous measurements of the particle number do not represent significant overhead.

Instead of constructing circuits to simultaneously diagonalize and measure the symmetry operators S and the target observable O, Bonet-Monroig *et al.* (2018) showed that it is possible to perform effective postselection via postprocessing in symmetry verification, which turns out to be closely related to subspace expansion (McClean *et al.*, 2017); see Sec. III.F. We now consider the simple example in which there is only a single Pauli symmetry operator S and where the ideal state lives within the $+1$ eigenspace of S defined by the projector $\Pi=(1/2)(1+S)$. In this way, if the state prepared prior to the postselection is ρ, the postselected state is then

$$\rho_{\mathrm{sym}} = \frac{\Pi\rho\Pi}{\mathrm{Tr}[\Pi\rho\Pi]}. \tag{26}$$

The symmetry-verified expectation value for the target observable O is given by

$$\mathrm{Tr}[O\rho_{\mathrm{sym}}] = \frac{\mathrm{Tr}[O\Pi\rho\Pi]}{\mathrm{Tr}[\Pi\rho\Pi]} = \frac{\mathrm{Tr}[O_{\mathrm{sym}}\rho]}{\mathrm{Tr}[\Pi\rho]}, \tag{27}$$

where $O_{\mathrm{sym}}=\Pi O\Pi$ is the symmetrized observable; i.e., the verified expectation value $\mathrm{Tr}[O\rho_{\mathrm{sym}}]$ is simply the quotient between the noisy expectation value of O_{sym} and Π. The symmetrized observable O_{sym} and the symmetry projector Π can be further decomposed into the Pauli basis. For instance,

if O is Pauli, the Pauli decomposition of O_{sym} is simply $O_{\mathrm{sym}} = \Pi O \Pi = (O + SO + OS + SOS)/4$. In this way, the expectation values of O_{sym} and Π can be obtained by measuring the expectation values of these Pauli basis operators (or via the Monte Carlo sampling in Appendix A.1). In the previously described method, we need only measure one Pauli observable in a given circuit run, which can be carried out using only single-qubit Pauli measurements without needing additional circuit components. The aforementioned simple examples can be further generalized to the cases of multiple symmetries and non-Pauli symmetries with a change in the definition of the projector Π, reflecting a change in the symmetry subspace. We can also decompose the observables into bases beyond the Pauli basis as long as the components are easy to measure. Instead of combining the measurement results of the basis to reconstruct the projected observable O_{sym} and the projector Π, Cai (2021c) showed that it is possible to achieve smaller biases by combining these measurement results with different weights, at the cost of larger sampling overhead.

Note that as we try to describe different ways of performing the aforementioned symmetry verification, we clearly separate postselection from postprocessing (for a clear distinction between different methods) even though postselection is technically a specific type of postprocessing in the most general context of QEM.

For direct in-circuit symmetry verification, the fraction of circuit runs that are "useful" is simply the "pass rate" of the symmetry checks given by $\mathrm{Tr}[\Pi\rho]$. The corresponding sampling overhead is simply the inverse of this pass rate: $C_{\mathrm{em}} \sim \mathrm{Tr}[\Pi\rho]^{-1}$. The fault-free postselection discussed in Sec. II.D can be viewed as the ideal symmetry verification that can detect all faults. It can achieve zero bias, but correspondingly its sampling overhead also sets the upper bound for all possible direct symmetry verification. As discussed, we can greatly simplify the measurement circuit through postprocessing, but this will come at a higher sampling overhead, which scales as $C_{\mathrm{em}} \sim \mathrm{Tr}[\Pi\rho]^{-2}$ (Cai, 2021c; Huggins, McClean *et al.*, 2021).

Though it is simplest to rely on the native symmetries of a system for the purpose of error mitigation, one can consider adding more symmetries artificially or unitarily transforming a system to improve the error-mitigation power of a set of symmetries. This is important, as symmetry-based methods cannot mitigate against errors that commute with all symmetries of the system. Bonet-Monroig *et al.* (2018) showed that one can unitarily transform chemistry Hamiltonians such that no single-qubit operator commutes with all symmetries, and that this can be preferable even to removing qubits from the system. They also described a basic scheme to add artificial symmetries to a system. However, this scheme makes local system operators highly nonlocal and thus is relatively unscalable.

A solution to the aforementioned problem was found in the Bravyi-Kitaev superfast transformation, where artificial symmetries are used to transform local fermionic Hamiltonians to local qubit Hamiltonians (Bravyi and Kitaev, 2002; Setia and Whitfield, 2018). These symmetries are necessary for implementing a geometrically local fermion-to-qubit transformation in more than one dimension, and at the same time they provide a natural boon for symmetry-based QEM. The list of fermion-to-qubit transformations has seen significant development and optimization in recent years (Steudtner and Wehner, 2018, 2019; Setia *et al.*, 2019), with attempts to optimize the number of local errors that can be mitigated (Jiang *et al.*, 2019; Derby and Klassen, 2021). Jiang *et al.* (2019) pointed out that this need not be constrained to fermionic lattice models: Since time evolution on a fermionic system can be mapped to a series of operations that are local on a lattice (without any asymptotic growth in the circuit depth), one can implement this using a mapping intended for a local fermionic lattice without having some system observables become extensively large. Jiang *et al.* (2019) further found an encoding (the "Majorana loop stabilizer code") that can mitigate or even correct all single-qubit errors, making it in effect an error correcting code of distance 3. However, the Eastin-Knill theorem suggests that these methods cannot be extended to construct codes of arbitrarily large distances (Eastin and Knill, 2009).

Encoding the qubits into QEC codes is also a way to add artificial symmetry. Performing direct in-circuit verification in this case is simply quantum error detection. However, if the symmetries (stabilizers) of the code are hard to directly measure due to high weight or connectivity constraints, then as discussed one can instead perform postprocessing verification for QEC codes (McClean *et al.*, 2020). This was later extended to include midcircuit stabilizer checks (Tsubouchi *et al.*, 2023) and bosonic codes (Endo *et al.*, 2022).

A large number of experiments have demonstrated stabilizer measurements in error correction codes throughout the 2010s (Nigg *et al.*, 2014; Córcoles *et al.*, 2015; Kelly *et al.*, 2015; Linke *et al.*, 2017; Vuillot, 2018). However, to our knowledge the first experimental demonstration of symmetry verification using natural symmetries ($\prod_i Z_i$) in a quantum algorithm was by Sagastizabal *et al.* (2019) through postprocessing, as low-cost techniques to measure natural symmetries were not previously known. Later Google Quantum AI *et al.* (2020a) made direct simultaneous measurement of the number operator and the fermionic one-particle reduced density matrix (1RDM) using the FSWAP rotation [Eq. (25)] and combined this with McWeeny purification (Sec. III.E). Stanisic *et al.* (2022) successfully demonstrated the verification of multiple symmetries (number, particle-hole symmetry, and total spin) and also combined it with learning-based methods (Sec. III.H). Owing to the simplicity and effectiveness of symmetry verification, it has been employed in a wide range of other experiments (Google Quantum AI *et al.*, 2020b, 2022; Neill *et al.*, 2021; Dborin *et al.*, 2022; Stanisic *et al.*, 2022).

E. Purity constraints

Many quantum algorithms target the preparation of an ideal state ρ_0 that is pure: $\rho_0 = |\psi_0\rangle\langle\psi_0|$. Many common noise channels are stochastic, which will turn our ideal pure state ρ_0 into some noisy mixed state ρ. At a high level, error-mitigation techniques based on purity constraints attempt to reduce the bias in the expectation value by trying to approximate the pure state closest to ρ. If we look at the spectral decomposition of ρ,

$$\rho = \sum_{i=1}^{2^N} p_i |\phi_i\rangle\langle\phi_i|, \tag{28}$$

where the eigenvalues are ordered $p_i \geq p_j$ for $i < j$ and we assume that $p_1 > p_2$ for simplicity, then the closest pure state to ρ in the trace distance is simply the dominant eigenvector $|\phi_1\rangle\langle\phi_1|$. In principle, one could use the quantum principle component analysis to sample from the eigenbasis of ρ (Lloyd, Mohseni, and Rebentrost, 2014). Combined with postselection this would allow for the efficient preparation of the dominant eigenvector. However, the additional circuit required is too deep for mitigating errors in near-term devices, and thus simpler strategies are needed.

One such strategy, referred to as virtual distillation (VD) (Huggins, McArdle *et al.*, 2021) or error suppression by derangement (ESD) (Koczor, 2021b), uses collective measurements of M copies of ρ in order to access expectation values with respect to the Mth degree purified state

$$\rho_{\mathrm{pur}}^{(M)} = \frac{\rho^M}{\mathrm{Tr}[\rho^M]} = \frac{1}{\sum_{i=1}^{2^N} p_i^M} \sum_{i=1}^{2^N} p_i^M |\phi_i\rangle\langle\phi_i|. \tag{29}$$

We see that, under the assumption that p_1 is strictly greater than p_2, we have $\lim_{M\to\infty} \rho_{\mathrm{pur}}^{(M)} = |\phi_1\rangle\langle\phi_1|$. The rate at which this is achieved is exponential in the number of copies used for purification M. The remaining bias in the VD or ESD estimator at $M \to \infty$ comes from the deviation between $|\phi_1\rangle\langle\phi_1|$ and the target state ρ_0, which is sometimes known as the *coherent mismatch* or *noise floor* (Huggins, McArdle *et al.*, 2021; Koczor, 2021b). As the largest sources of noise in state-of-the-art quantum devices are typically incoherent, this coherent mismatch can be expected to be significantly smaller than the error in the unmitigated state. Indeed, numerical and analytic studies have confirmed that the error suppression from VD or ESD can be of multiple orders of magnitude for large systems, even using as little as $M = 2$ copies of the state (Huggins, McArdle *et al.*, 2021; Koczor, 2021a, 2021b).

It was shown by Huggins, McArdle *et al.* (2021) and Koczor (2021b) that one can estimate expectation values of the purified states in Eq. (29) without ever having to prepare them on a quantum device. The expectation value of this purified state with respect to the observable of interest O is $\mathrm{Tr}[O\rho_{\mathrm{pur}}^{(M)}] = \mathrm{Tr}[O\rho^M]/\mathrm{Tr}[\rho^M]$. In VD or ESD, this is obtained by estimating $\mathrm{Tr}[O\rho^M]$ and $\mathrm{Tr}[\rho^M]$ in separate measurements. If we let S_M be the cyclic permutation operator between M copies of ρ, the quantity $\mathrm{Tr}[O\rho^M]$ can be estimated using

$$\mathrm{Tr}[O\rho^M] = \mathrm{Tr}[S_M O_m \rho^{\otimes M}] = \mathrm{Tr}[S_M \bar{O} \rho^{\otimes M}], \tag{30}$$

where O_m is the operator O acting on the mth copy and $\bar{O} = (1/M)\sum_m O_m$ is the observable symmetrized under copy permutation. Thus, $\mathrm{Tr}[O\rho^M]$ can be obtained by measuring $S_M O_m$ or $S_M \bar{O}$ on M noisy copies of ρ. A diagrammatic proof of Eq. (30) is shown in Fig. 3. One can extend this to estimate $\mathrm{Tr}[\rho^M]$ by putting I in the place of O.

Measuring a global operator like S_M can be challenging, but it can be decomposed into transversal operations among different copies $S_M = \otimes_{n=1}^N \tilde{S}_M^{(n)}$, where $\tilde{S}_M^{(n)}$ cyclically permute the nth qubits of different copies as shown in Fig. 4. If the observable O acts on a single qubit, then the symmetrized observable $\bar{O}$ will commute with all $\tilde{S}_M^{(n)}$ (and thus $S_M \bar{O}$ can be obtained by measuring low-weight operators $\tilde{S}_M^{(n)}$ and O_m for all n and m) and then postprocess. This requires only transversal operations among the identically labeled qubits of each copy of ρ, thereby avoiding global measurement. Explicit circuits for $O = Z$ and $M = 2$ and 3 without ancilla qubits were given by Huggins, McArdle *et al.* (2021); more general measurements can be achieved with Hadamard tests using ancilla qubits (Huggins, McArdle *et al.*, 2021; Koczor, 2021b). If the observable O is not single qubit but rather is a tensor product of single-qubit operators, $O = \otimes_{n=1}^N G^{(n)}$ (for instance, O is Pauli), then the observable $S_M O_m$ in Eq. (30) can be decomposed into a tensor product of low-weight operators $\otimes_{n=1}^N \tilde{S}_M^{(n)} G_m^{(n)}$ that can be measured in a transversal manner (Cai, Siegel, and Benjamin, 2023). We can use Hadamard tests to measure each $\tilde{S}_M^{(n)} G_m^{(n)}$, which requires N ancilla qubits in total. To efficiently carry out any of the aforementioned low-weight measurement schemes, we need transversal operations among different copies, which can be challenging to implement in practice and may involve long-range interactions. A hardware architecture with native transversal operations among different copies was proposed (Cai, Siegel, and Benjamin, 2023) in which an implementation of VD or ESD is shown with almost no space-time overhead.

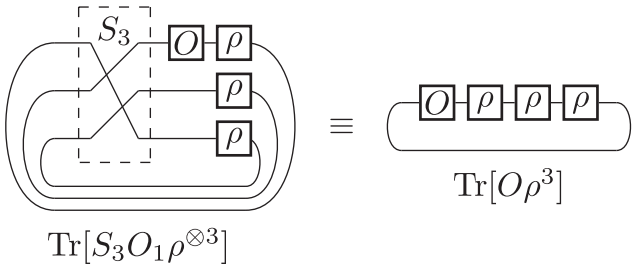

FIG. 3. Diagrammatic proof of Eq. (30) for three copies ($M = 3$) with the observable acting on the first copy ($m = 1$). The proof uses tensor network notations (Bridgeman and Chubb, 2017) and can be easily extended to more copies and/or with O acting on the mth copy rather than the first.

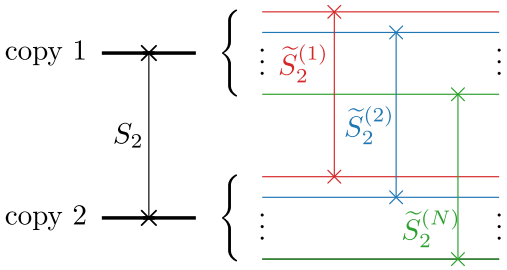

FIG. 4. Decomposition of the copy-swap operator S_2 into transversal qubit-swap operators $\tilde{S}_2$. A similar decomposition also applies to other cyclic copy-permutation operators S_M with $M > 2$.

In the more general case, one can imagine either unitarily transforming O to a single-qubit observable (if possible) or linearly decomposing it into a sum of simpler terms (which is always possible). One could additionally imagine decomposing $S_M O_m$ or $S_M \bar{O}$ into a linear combination of Pauli operators and measuring the entire set of operators using shadow tomography (Huang, Kueng, and Preskill, 2020; Hu *et al.*, 2022; Seif *et al.*, 2023). This has the advantage that one can reuse a single copy of ρ rather than using multiple physical copies. However, the sampling cost in this case scale exponentially with the number of qubits and thus may be difficult to scale beyond small system sizes (Seif *et al.*, 2023).

Instead of estimating $\mathrm{Tr}[O\rho^2]$ using two copies of ρ separated in space, it is possible to do the estimation using two copies separated in time. This technique has been given multiple names since its initial incarnation (Cai, 2021d; O'Brien *et al.*, 2021; Huo and Li, 2022), but here we refer to it as echo verification (EV). To perform EV, we ideally want to measure the ideal state projector $\rho_0 = |\psi_0\rangle\langle\psi_0|$ at the end of the circuit for postselection, which returns 1 if the output state is the ideal state and 0 otherwise, just like a symmetry check. The measurement of the projector ρ_0 is carried out by applying the inverse of the primary circuit (O'Brien *et al.*, 2021), and the noise in the inverse circuit means that we are effectively measuring the noisy operator $\bar{\rho}$ instead of ρ_0. If we measure the observable of interest O along with the noisy projector $\bar{\rho}$ in the same circuit run and postselect according to the outcome of $\bar{\rho}$, we are effectively performing a measurement of O on the state $(\bar{\rho}\rho + \rho\bar{\rho})/2\mathrm{Tr}[\bar{\rho}\rho]$ (Huo and Li, 2022). Compared to Eq. (29), we see that this is simply the second-degree purified state with $\bar{\rho}$ in the place of ρ. The measurement of O alongside $\bar{\rho}$ is usually carried out using a Hadamard test (O'Brien *et al.*, 2021). Alternatively, it is possible to achieve a similar degree of error suppression without the ancilla by preparing a superposition of ρ_0 with a known eigenstate of O and applying the gate O on the quantum state (assuming that O is unitary) (Lu, Bañuls, and Cirac, 2021; O'Brien *et al.*, 2021), though this is not formally equivalent to second-degree purification. The error-mitigation power of EV has been demonstrated in numerical simulation (O'Brien *et al.*, 2021) and a four-qubit experiment (Huo and Li, 2022). The results differ depending on the exact circuit implementation of the methods, even in simplified noise models (O'Brien *et al.*, 2021), indicating that further optimizing the circuit implementation could be an interesting direction to investigate. Gu *et al.* (2023) recently proved that applying EV to control-free phase estimation (Lu, Bañuls, and Cirac, 2021; O'Brien *et al.*, 2021; Russo *et al.*, 2021) corrects any noise source with Hermitian Kraus operators to first order.

Owing to their similarity, EV and VD or ESD are directly comparable in terms of performance and resource requirements. If one wants to suppress the incoherent contributions of ρ beyond second-degree purification, then one can only use VD or ESD instead of EV. However, EV has a smaller qubit footprint and requires less circuit overhead for measurement (especially as it removes the need to perform operations across multiple copies of the same state). In fact, EV can be combined with VD or ESD to achieve a high degree of purification with a lower qubit and circuit overhead (Cai, 2021d). There is also a difference between these two approaches in terms of their sampling overhead, similar to the difference between the sampling overhead of direct and postprocessing symmetry verification (Sec. III.D). For EV, the error mitigation is done through direct postselection, and thus the sampling overhead is simply the inverse of the success probability $C_{\mathrm{em}} \sim \mathrm{Tr}[\bar{\rho}\rho]^{-1} \sim \mathrm{Tr}[\rho^2]^{-1}$. For the two-copy version of VD or ESD, the error mitigation is done through effective postselection (postprocessing), and thus the sampling overhead increases more steeply, scaling as $C_{\mathrm{em}} \sim \mathrm{Tr}[\rho^2]^{-2}$. More generally, the sampling cost of VD or ESD scales exponentially in the number of copies M (as $\mathrm{Tr}[\rho^M]$ is exponentially small in M unless ρ is pure). If the faults in the circuit follow a Poisson distribution (Sec. II.C), we then have $\mathrm{Tr}[\rho^2] \lesssim P_0^2 = e^{-2\lambda}$, where P_0 is the fault-free probability of the circuit and λ is the circuit fault rate. This suggests that both EV and VD or ESD can incur a sampling cost growing exponentially with the circuit fault rate, as we discussed for general bias-free QEM in Sec. II.D. Furthermore, both methods (in their standard form) lack the parallelizability of unmitigated expectation value estimation. As mentioned, VD or ESD typically estimates expectation values of only N operators (rather than their products). EV is restricted further; Polla, Anselmetti, and O'Brien (2023) showed that EV cannot be efficiently parallelized for even commuting observables, reducing the information extracted by the method to a single bit per state preparation. Though this can be significantly optimized over a simple Pauli decomposition of a complex O (Polla, Anselmetti, and O'Brien, 2023), it presents roughly an $O(N)$ overhead in sampling compared to VD or ESD.

The original formulation of EV in terms of projection on the initial state gives a clear connection between purification- and (symmetry) projection-based techniques (Sec. III.D). Though they are not equivalent, this raises the question of whether a similar connection can be made to VD or ESD. A connection was originally made by Huggins, McArdle *et al.* (2021), who pointed out that in the eigenbasis $\{|\phi_j\rangle\}$ of ρ the swap operator measurement has zero expectation value on terms in $\rho \otimes \rho$ except for those of the form $|\phi_j\rangle\langle\phi_j| \otimes |\phi_j\rangle\langle\phi_j|$. Cai (2021c) showed that VD and ESD can naturally arise from the verification of copy-permutation symmetry by replacing the symmetry projector with a general linear combination of permutation operators, which is also connected to subspace expansion using symmetry operators (as discussed in Sec. III.F). Yoshioka *et al.* (2022) further exploited the connection between VD or ESD and subspace expansion and looked at the effect of performing error mitigation using a linear combination of states with different degrees of purification. Both EV and VD or ESD were successfully implemented experimentally by O'Brien *et al.* (2023), where they performed variational ground state energy estimations up to 20 qubits (two copies of ten-qubit states for VD or ESD) and achieved 1 to 2 order of magnitude error reductions by applying these QEM methods.

F. Subspace expansions

In some quantum tasks, one has knowledge of not only the ideal circuit and potential noise sources but also the structure of the task at hand. One such class of methods that can take

advantage of the knowledge on the problem are *quantum subspace expansion* techniques (McClean *et al.*, 2017, 2020; Takeshita *et al.*, 2020). Many tasks in quantum computing, such as optimization or state preparation, can be phrased as the desire to minimize the objective function $\langle\psi|H|\psi\rangle$ with respect to the state $|\psi\rangle$ for a known Hermitian operator H. Often we are unable to directly prepare the optimal state, due to either a coherent error in our implementation of the ideal circuit or a simple lack of knowledge of the ideal circuit. However, there is usually a set of M different states $\{|\phi_i\rangle\}$ (linearly independent but not necessarily orthogonal to each other) that we can easily prepare, which can be used to construct our target state through linear combination $|\psi_{\vec{w}}\rangle = \sum_{i=0}^{M-1} w_i|\phi_i\rangle$. In this way, the problem of finding the optimal state becomes finding the optimal set of coefficients $\vec{w}^*$,

$$\vec{w}^* = \underset{\vec{w}}{\operatorname{argmin}} \langle\psi_{\vec{w}}|H|\psi_{\vec{w}}\rangle \quad \text{such that} \quad \langle\psi_{\vec{w}}|\psi_{\vec{w}}\rangle = 1. \tag{31}$$

This has the well-known exact solution in the form of a generalized linear eigenvalue problem

$$\bar{H}W = \bar{S}WE, \tag{32}$$

where

$$\bar{H}_{ij} = \langle\phi_i|H|\phi_j\rangle, \qquad \bar{S}_{ij} = \langle\phi_i|\phi_j\rangle. \tag{33}$$

That is, $\bar{H}$ is the $M \times M$ matrix representation of H in the chosen basis set $\{|\phi_i\rangle\}$ and $\bar{S}$ is the overlap matrix for the basis set. In Eq. (32), W and E are the matrices of eigenvectors and eigenvalues, respectively, of the solved problem. The eigenvector in W with the lowest eigenvalue is precisely the optimal combination of coefficients $\vec{w}^*$ for the state $|\psi_{\vec{w}^*}\rangle$.

If we want to obtain the improved expectation value of some observable O with respect to our new found state $|\psi_{\vec{w}^*}\rangle$, it is given simply as

$$\langle\psi_{\vec{w}^*}|O|\psi_{\vec{w}^*}\rangle = \sum_{i,j=0}^{M-1} w_i^* w_j^* \langle\phi_i|O|\phi_j\rangle.$$

That is, we can construct it using the optimal weight $\vec{w}^*$ and the measurement results of $\langle\phi_i|O|\phi_j\rangle$, without needing to explicitly prepare $|\psi_{\vec{w}^*}\rangle$. For the special case of $O = H$, the improved expectation value is given simply by the smallest eigenvalue in E when we solve Eq. (32). If desired, however, one could prepare the state $|\psi_{\vec{w}^*}\rangle$ via linear combination of unitaries methods (Childs and Wiebe, 2012) in order to use it as the input state for a subsequent quantum routine.

If one takes the limit of choosing $|\phi_i\rangle$ to be a complete basis for the entire Hilbert space, then solving the optimization problem will return the ideal state; however, choosing an exponentially large space to perform classical optimization defeats the purpose of using a quantum computer to begin with. Hence, how to choose the right set of basis states $\{|\phi_i\rangle\}$ is the key to the success of quantum subspace expansion. The first basis state $|\phi_0\rangle$ that we select is usually the best state that we can prepare before performing quantum subspace expansion. In this way, in the worst case we simply obtain $|\phi_0\rangle$ through subspace expansion.

For the other basis states, the original work of McClean *et al.* (2017) suggested that we can draw inspiration from the configuration interaction expansions (Helgaker, Jorgensen, and Olsen, 2000) in quantum chemistry, which is commonly used for improving energy and properties of mean-field states as well as determining excited states for response properties. There each of the other basis states is generated by applying an expansion basis operator G_i on the original state such that $G_i|\phi_0\rangle = |\phi_i\rangle$. Knowledge of a good set of expansion basis operators $\{G_i\}$ can come from symmetry considerations, from excitation operators, or simply from knowing that correcting (as opposed to replacing) the state $|\phi_0\rangle$ requires that the additional states are connected directly or indirectly through H. In this way, the expanded state is now $|\psi_{\vec{w}}\rangle = \Gamma_{\vec{w}}|\phi_0\rangle$, where $\Gamma_{\vec{w}} = \sum_{i=0}^{M-1} w_i G_i$ is the expansion operator, which is a weighted sum of expansion basis operators. Once the expansion basis operators are determined, the matrix elements required for solving the optimization equation in Eq. (32) can be measured on a quantum computer through

$$\begin{aligned} \bar{H}_{ij} &= \langle\phi_0|G_i^\dagger H G_j|\phi_0\rangle = \mathrm{Tr}[G_i^\dagger H G_j \rho], \\ \bar{S}_{ij} &= \langle\phi_0|G_i^\dagger G_j|\phi_0\rangle = \mathrm{Tr}[G_i^\dagger G_j \rho] \end{aligned} \tag{34}$$

without needing detailed knowledge of the original state $\rho = |\phi_0\rangle\langle\phi_0|$.

Thus far both the starting state $|\phi_0\rangle$ and the expanded state $|\psi_{\vec{w}^*}\rangle$ are pure states; as a result, any errors that we have removed are coherent errors. This is in stark contrast to our focus on incoherent errors in Sec. III.E. The right-hand side of Eq. (34) is suggestive of the fact that we can apply expansion around a mixed state ρ to remove incoherent errors. This was indeed first conjectured in the original work of McClean *et al.* (2017) and was later confirmed by several experimental implementations of the method (Colless *et al.*, 2018; Sagastizabal *et al.*, 2019; Urbanek *et al.*, 2020). These observations were put on more solid theoretical footing by Bonet-Monroig *et al.* (2018) and McClean *et al.* (2020). The effective state after performing a subspace expansion on a noisy state ρ was shown to be (McClean *et al.*, 2020)

$$\rho_{\mathrm{sub}} = \frac{\Gamma_{\vec{w}} \rho \Gamma_{\vec{w}}^\dagger}{\mathrm{Tr}[\Gamma_{\vec{w}} \rho \Gamma_{\vec{w}}^\dagger]}. \tag{35}$$

We see that Eq. (35) is similar to the symmetry-verified state in Eq. (26), with the symmetry projector Π replaced by the expansion operator $\Gamma_{\vec{w}}$. This implies that, using the symmetry operators as our expansion basis operators, we can recover the symmetry subspace by performing subspace expansion. Cai (2021c) tried to further generalize this by searching for expanded "states" of the form $\Gamma_{\vec{w}}\rho/\mathrm{Tr}[\Gamma_{\vec{w}}\rho]$ instead, which allows us to also incorporate purification-based QEM in Sec. III.E under this formalism.

A number of recent works have looked into other possible sets of expansion basis operators $\{G_k\}$. For example, when the operators $\{G_k\}$ are chosen to be powers of the Hamiltonian $\{H^k\}$, we see that these methods coincide with

quantum Krylov subspace methods like the QLanczos algorithm (Motta *et al.*, 2020) or other methods based on filtering the eigenspectrum via functions of the Hamiltonian (Suchsland *et al.*, 2021). A previously mentioned recent work (Yoshioka *et al.*, 2022) included operators that are powers of the density matrix, making close ties to the purification-based QEM methods. By doing so, optimal combinations of states can now exploit problem-specific knowledge related to purity in addition to general knowledge.

G. *N* representability

Often, in the context of quantum simulation (and sometimes more broadly), the goal is ultimately to measure a set of observables corresponding to a marginal of the total density matrix known as the reduced density matrix (RDM). Given a general quantum state ρ on N qubits, the set of p-qubit RDMs is obtained by integrating out q qubits (such that $N-q=p$) of the joint distribution,

$$ {}^{p}\rho_{m_1,\ldots,m_p} = \mathrm{Tr}_{n_1,n_2,\ldots,n_q}[\rho], \tag{36}$$

resulting in $\binom{n}{p}$ different RDMs, each of dimension $2^p \times 2^p$. The coefficients $n_1, \ldots, n_q$ on the trace operator indicate which qubits are integrated out of ρ, and coefficients $m_1, \ldots, m_p$ label the subsystem marginal. The result of this marginalization is a distribution over p qubits.

The connection to error mitigation is that these RDMs are known to have special geometric structures; not all marginals that one can write are consistent with having come from a valid (or, in the parlance of this field, representable) wave function. In principle there is a set of conditions that constrain the space of representable RDMs. Articulating and evaluating these equality and inequality constraints is known as the N-representability problem (Mazziotti, 2016), and the problem is formally quantum Merlin-Arthur complete (Liu, Christandl, and Verstraete, 2007). However, for most RDMs of interest one can write and evaluate at least some of the N-representability conditions. That knowledge can often be used to mitigate errors, in a spirit similar to the application of symmetry constraints, but generally using different methods.

The focus on measuring RDMs is especially common when simulating many-body systems of identical particles. For example, because real fermions interact pairwise, most properties of interest can be obtained using only the one-particle and two-particle reduced density matrices [and because the particles are identical, there is only a single 1RDM and a single two-particle RDM (2RDM)]. Using the second quantized fermionic creation and annihilation operators acting on sites p, $a_p^\dagger$, and a_p, the fermionic 1RDM and 2RDM can be expressed as

$$ {}^{1}D_j^i = \mathrm{Tr}[a_i^\dagger a_j {}^{n}D] = \langle\psi|a_i^\dagger a_j|\psi\rangle, \tag{37}$$

$$ {}^{2}D_{rs}^{pq} = \mathrm{Tr}[a_p^\dagger a_q^\dagger a_s a_r {}^{n}D] = \langle\psi|a_p^\dagger a_q^\dagger a_s a_r|\psi\rangle, \tag{38}$$

where ${}^{k}D$ is the k-particle RDM and the equalities on the right-hand side correspond to the case of pure states. For a system of N sites this 2RDM is only of dimensions $N^2 \times N^2$ (in contrast to $2^N \times 2^N$ for the full density matrix) and yet completely determines the energy of a fermionic system with pairwise interactions.

Some of the simpler constraints we can express on the 1RDM and the 2RDM are as follows:

(1) Hermiticity of the density matrices

$$ {}^{1}D_i^j = ({}^{1}D_j^i)^*, \tag{39}$$

$$ {}^{2}D_{rs}^{pq} = ({}^{2}D_{pq}^{rs})^*. \tag{40}$$

(2) Antisymmetry of the two-particle marginal

$$ {}^{2}D_{rs}^{pq} = -{}^{2}D_{sr}^{pq} = -{}^{2}D_{rs}^{qp} = {}^{2}D_{sr}^{qp}. \tag{41}$$

(3) The $(p-1)$-marginal is related to the p-marginal by contraction; for instance, the 2-marginal can be contracted to the 1-marginal

$$ {}^{1}D_j^i = \frac{1}{n-1}\sum_k {}^{2}D_{jk}^{ik}. \tag{42}$$

(4) The trace of each marginal is fixed by the number of particles in the system

$$ \mathrm{Tr}[{}^{1}D] = n, \tag{43}$$

$$ \mathrm{Tr}[{}^{2}D] = n(n-1). \tag{44}$$

(5) The marginals are proportional to density matrices and are thus positive semidefinite,

$$ \{{}^{1}D, {}^{2}D\} \succeq 0. \tag{45}$$

Note that this is not an exhaustive list of all N-representability constraints.

Rubin, Babbush, and McClean (2018) first suggested that knowledge of these constraints could be used for mitigating errors when measuring RDMs. The essential idea is that in the course of a NISQ simulation one might measure an RDM that violates N-representability conditions as a consequence of errors corrupting the estimation of the tomography elements composing the RDM. However, one can use even a partial list of RDM conditions in order to project the noisy and unrepresentable RDM estimate back to the nearest RDM consistent with a list of RDM constraints. Because the RDM constraints all take the form of equality and inequality constraints, this can be performed using semidefinite programming. Such an approach was demonstrated numerically by Rubin, Babbush, and McClean (2018) and experimentally by Smart and Mazziotti (2020).

Of special interest for error mitigation is a subdiscipline of the N-representability field known as pure-state N representability that is concerned with describing the geometry of pure density matrices (i.e., N representability that must hold for pure states). In principle using pure-state representability would potentially allow one to measure the 2RDM of a partially decohered state and then project that estimate back to the nearest RDM consistent with having come from a pure state. While this idea was first discussed by Rubin, Babbush,

and McClean (2018), it has been difficult to realize in practice since pure N-representability conditions are extremely difficult to compute. The state of the art in the field is that specialized computer algebra systems are needed to generate the pure-state conditions (Klyachko, 2006; DePrince, 2016). Nevertheless, some work has succeeded in using some of these conditions in quantum simulations (Smart and Mazziotti, 2019).

In the case when the 1RDM of a fermionic system is expected to be idempotent ($D^2 = D$), a special type of purification known as McWeeny purification (McWeeny, 1960) is possible. This purification scheme is achieved by iterating on a nonidempotent 1RDM estimate as

$$^1D_{i+1} = 3(^1D_i)^2 - 2(^1D_i)^3 \tag{46}$$

until idempotency is restored. The only fermionic states with idempotent 1RDMs are Slater determinants, a factor that limits the applicability of this scheme in strongly correlated systems. Despite the lack of theoretical justification, McCaskey *et al.* (2019) demonstrated moderate mitigation success from McWeeny purification in a correlated four-qubit chemistry simulation. The most notable success, however, came from the application made by Google Quantum AI *et al.* (2020a) to a Hartree-Fock state, for which the idempotency assumption is justified. They demonstrated McWeeny purification between 1 and 2 orders of magnitude of error suppression, on top of the other mitigation methods used. Hope remains that similar results can be demonstrated in other experiments by enforcing the just-discussed pure-state representability constraints, or by applying purification-based QEM methods discussed in Sec. III.E.

H. Learning-based methods

Given the primary circuit $\boldsymbol{P}$, its noisy expectation value is denoted as $E(\boldsymbol{P})$, and we are trying to estimate its noiseless expectation value $E_0(\boldsymbol{P})$. To achieve this, we obtain the error-mitigated expectation value $E_{\vec{\theta}}(\boldsymbol{P})$ as a function of the primary circuit and a set of parameters $\vec{\theta}$. Thus far we have seen that the function parameters $\vec{\theta}$ in different QEM methods are obtained through different noise calibration processes. However, we can also obtain $\vec{\theta}$ through *learning-based methods* (Czarnik *et al.*, 2021; Strikis *et al.*, 2021) using training circuits. We construct a training circuit $\boldsymbol{T}$ here that satisfies the folllowing conditions:

(1) It is similar to $\boldsymbol{P}$, usually in terms of circuit structures, such that it contains circuit faults similar to $\boldsymbol{P}$.
(2) It is classically simulable; i.e., its ideal expectation value $E_0(\boldsymbol{T})$ can be obtained via classical simulation.

Knowing the exact value of $E_0(\boldsymbol{T})$ enables us to find a good set of parameters $\vec{\theta}$ for the error-mitigation function $E_{\vec{\theta}}$ by minimizing the difference between $E_0(\boldsymbol{T})$ and $E_{\vec{\theta}}(\boldsymbol{T})$. We assume that the error-mitigation protocol $E_{\vec{\theta}}$ obtained from the training circuit $\boldsymbol{T}$ also works well for the primary circuit $\boldsymbol{P}$ due to their similarity, which give us the error-mitigated result $E_{\vec{\theta}}(\boldsymbol{P})$ as an estimate of the ideal result $E_0(\boldsymbol{P})$. More generally we can have more than one training circuit, which is denoted using the training set $\mathbb{T}$. The simplest loss function we can construct to obtain the optimal $\vec{\theta}$ is

$$L_{\mathbb{T}}(\vec{\theta}) = \frac{1}{|\mathbb{T}|}\sum_{\boldsymbol{T}\in\mathbb{T}}[E_0(\boldsymbol{T}) - E_{\vec{\theta}}(\boldsymbol{T})]^2. \tag{47}$$

The entire process of learning-based QEM is summarized in Fig. 5. If the error-mitigated estimate $E_{\vec{\theta}}(\boldsymbol{T})$ is linear in $\vec{\theta}$, which is the case for many QEM schemes, then the optimal $\vec{\theta}$ can be obtained using linear least squares.

The simplest error-mitigation function simply rescales and shifts the noisy expectation value to approximate the ideal expectation value,

$$E_0(\boldsymbol{A}) \approx E_{\vec{\theta}}(\boldsymbol{A}) = \theta_0 + \theta_1 E(\boldsymbol{A}). \tag{48}$$

In Eq. (48) the input circuit $\boldsymbol{A}$ can be the primary circuit $\boldsymbol{P}$ or the training circuits $\boldsymbol{T}\in\mathbb{T}$. Czarnik *et al.* (2021) first proposed using such a linear function to mitigate errors, whose coefficients can be obtained by training using the Clifford variants of the primary circuit. This has been shown to be

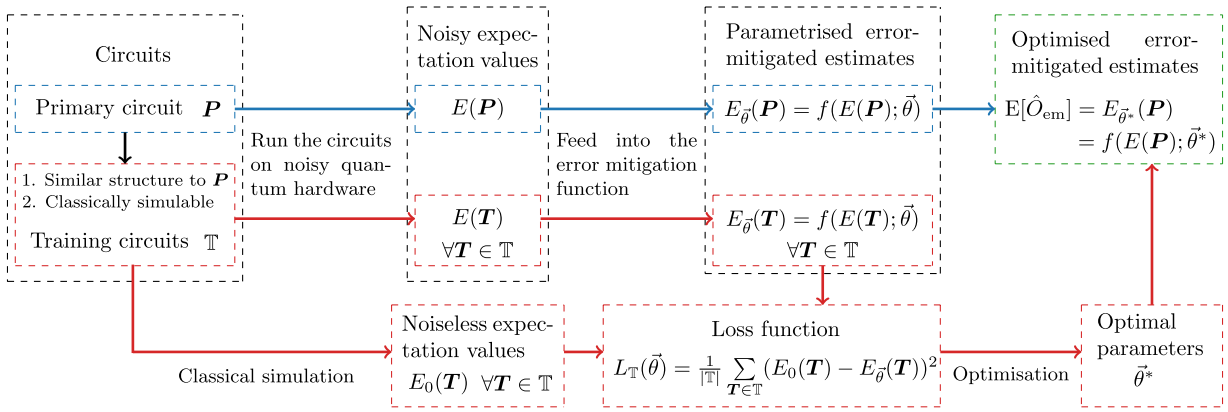

FIG. 5. The process of learning-based quantum error mitigation. For cases in which we need to construct variants of the input circuit for the error-mitigation function as in Eq. (49), the entire process is similar, but instead of running only the input circuits on the quantum hardware, we need to run all the different variants of the input circuits.

effective in experiments (Mi *et al.*, 2021; Urbanek *et al.*, 2021). When performing fermion simulations, it is also possible to train using the closest free-fermion model as proposed and demonstrated by Google Quantum AI *et al.* (2020b), or using fermionic linear optics (Montanaro and Stanisic, 2021) as demonstrated experimentally by Stanisic *et al.* (2022). In some particular use cases like measuring the out-of-time-order correlator (Mi *et al.*, 2021), by simply removing or replacing a small number of gates of the primary circuit, one can analytically derive the output of the resultant circuits, which can also be used as the training circuits. Rosenberg, Ginsparg, and McMahon (2022) experimentally compared the different ways to perform such rescaling methods.

The linear error-mitigation function in Eq. (48) can naturally arise when we assume that all noise sources in the circuit are globally depolarizing. In such a case, the resultant noisy state is simply a mixture of the ideal state and the completely mixed state $\rho = P_0\rho_0 + (1-P_0)I/2^N$ (Mi *et al.*, 2021; Vovrosh *et al.*, 2021), where P_0 is the fault-free probability of the circuit (Sec. II.C). Hence, the ideal expectation value takes the simple form of

$$\underbrace{\mathrm{Tr}[O\rho_0]}_{E_0(\boldsymbol{P})} = \frac{1}{P_0}\underbrace{\mathrm{Tr}[O\rho]}_{E(\boldsymbol{P})} - \frac{1-P_0}{P_0 2^N}\mathrm{Tr}[O],$$

which is in the same form as Eq. (48). The assumption of global depolarizing noise motivates the linear error-mitigation function but is not a necessary assumption for applying this error-mitigation function using training circuits. There is evidence that global depolarizing noise is an effective phenomenological error model emerging from gatewise error models when the gate number is large (Qin, Chen, and Li, 2023). However, if indeed there is only global depolarizing noise in the circuit, we can actually estimate the rescaling factor P_0^{-1} using Eq. (8) if we know the circuit fault rate, or obtain P_0^{-1} by measuring

$$\mathrm{Tr}[\rho^2] = P_0^2 + P_0(1-P_0)/2^{N-1} + (1-P_0)^2/2^{2N}$$

and solving the quadratic equation (Vovrosh *et al.*, 2021). Multiple ways to measure $\mathrm{Tr}[\rho^2]$ have been discussed in purification-based QEM (Sec. III.E). These circuits for obtaining P_0^{-1} do not have the same circuit structure as the primary circuit $\boldsymbol{P}$, so instead of viewing them as the training circuits for applying learning-based rescaling and shifting it may be more appropriate to view them as the noise calibration circuits for performing probabilistic error cancellation (Sec. III.B) against global depolarizing channels or for a special case of linear error extrapolation (Sec. III.A) (Cai, 2021a).

In Eq. (48), $E_{\vec{\theta}}(\boldsymbol{A})$ is simply a function of the noisy expectation value $E(\boldsymbol{A})$ of the input circuit $\boldsymbol{A}$; thus, only the input circuit $\boldsymbol{A}$ needs to be run on the quantum hardware. This is not the case for many of the previously discussed QEM methods. In general for a given input circuit $\boldsymbol{A}$, we need to construct a set of response measurement circuits $\{\boldsymbol{A}_{\mathrm{rsp},i}\}$ that are variants of the input circuit $\boldsymbol{A}$ by adding or replacing gates and/or adding measurements. These circuits can be, for instance, circuits of different noise levels for zero-noise extrapolation, circuits with different added gates for probabilistic error cancellation, and circuits with different added measurements for symmetry verification. The error-mitigated expectation value will be a function of the outputs of all of these response measurement circuits instead of merely the input circuit,

$$E_{\vec{\theta}}(\boldsymbol{A}) = f\big(\{E(\boldsymbol{A}_{\mathrm{rsp},i})\};\vec{\theta}\big). \tag{49}$$

One such example would be the error-mitigation estimate for probabilistic error cancellation in Eq. (17),

$$E[\hat{O}_{\mathrm{em}}] = \Sigma_{\vec{n}}\alpha_{\vec{n}}\langle\!\langle O|\mathcal{B}_{\vec{n}}|\rho_{\mathrm{in}}\rangle\!\rangle \mapsto E_{\vec{\theta}}(\boldsymbol{A}) = \Sigma_{\vec{n}}\theta_{\vec{n}}E(\boldsymbol{A}_{\mathrm{rsp},\vec{n}}).$$

Here the response measurement circuit $\boldsymbol{A}_{\mathrm{rsp},\vec{n}}$ corresponds to preparing ρ_{in}, applying the sequence of operation $\mathcal{B}_{\vec{n}}$, and measuring O, which as mentioned differs from $\boldsymbol{A}$ by additions or replacements of some subset of gates. The parameters $\theta_{\vec{n}}$ can be obtained as $\alpha_{\vec{n}}$ through device calibration, as discussed in Sec. III.B. Note that we can also obtain $\theta_{\vec{n}}$ via learning-based methods (Strikis *et al.*, 2021). Applying learning-based methods to probabilistic error cancellation implies that we do not have enough information about the gate errors in the primary circuit (otherwise, we will apply probabilistic error cancellation directly). Hence, we would need to assume Pauli gate errors in the primary circuit or apply Pauli twirling such that the set of response measurement circuits can be constructed by simply adding Pauli gates. Other than optimizing over $\theta_{\vec{n}}$ directly, we see that the response circuit coefficient $\alpha_{\vec{n}}$ is actually the product of the coefficients for individual gates in Sec. III.B. In a similar way, we can write $\theta_{\vec{n}}$ as the product of the coefficients for individual gates and optimize over the gate coefficients instead. This would greatly simplify the optimization problem if the number of gate types in the circuit is small. By incorporating the appropriate response measurement circuit $\boldsymbol{A}_{\mathrm{rsp},\vec{n}}$, it is also possible to mitigate spatially and temporally correlated noise using learning-based methods.

We continue to use probabilistic error cancellation with Clifford training (Strikis *et al.*, 2021) as an example to illustrate how the training circuits can be constructed. In probabilistic error cancellation, we want to remove all faults in the circuit. To find a way to mitigate these faults using the training circuits, the same faults must exist in these training circuits. Such training circuits can be constructed by replacing gates in the primary circuit with gates that have the same error channels. If we compile the primary circuit such that all the multiqubit gates in the primary circuit are Clifford ones, we then need only replace the single-qubit gates to construct Clifford training circuits that are classically simulable. If we further assume that all the single-qubit gates have the same error channel or they have negligible error rates compared to the multiqubit gates, then the fault distribution of these Clifford training circuits will be the same as the primary circuit. We use $\mathbb{C}$ and $\mathbb{U}$ to denote the set of all possible circuits generated by replacing the single-qubit gates in the primary circuit with random single-qubit Clifford and random single-qubit unitary circuits, respectively (note that $\mathbb{C} \subset \mathbb{U}$ and $\boldsymbol{P} \in \mathbb{U}$). By constructing the training circuits in the previously

outlined manner, the training loss function $L_{\mathbb{T}}(\vec{\theta})$ in Eq. (47) is a homogeneous polynomial of degree 2 in matrix elements of the single-qubit gates in the circuits. Since the Clifford group is a unitary 2-design, the loss function satisfies $L_{\mathbb{C}}(\vec{\theta}) = L_{\mathbb{U}}(\vec{\theta})$ (Wang, Chen *et al.*, 2021). Therefore, by training over the Clifford circuits and minimizing $L_{\mathbb{C}}(\vec{\theta})$, we are minimizing the errors in the more general unitary circuits in $\mathbb{U}$. When $L_{\mathbb{C}}(\vec{\theta})$ goes to zero, we have $L_{\mathbb{U}}(\vec{\theta}) = 0$, which implies $E_{\vec{\theta}}(\boldsymbol{A}) = E_0(\boldsymbol{A})$ (i.e., all errors are perfectly mitigated) for all $\boldsymbol{A} \in \mathbb{U}$, including the primary circuit $\boldsymbol{P}$.

Since the size of the Clifford set $\mathbb{C}$ grows exponentially with the number of qubits, it is impractical to evaluate $E_0(\boldsymbol{T})$ and the corresponding noisy response measurements $\{E(\boldsymbol{T}_{\mathrm{rsp},i})\}$ for all Clifford training circuits $\boldsymbol{T} \in \mathbb{C}$. Furthermore, the majority of noisy Clifford training circuits have near-zero expectation values that are costly to evaluate in terms of sampling overhead. One way to circumvent this is to truncate the Clifford training set using only circuits with large noiseless expectation values $|E_0(\boldsymbol{T})|$ due to their more significant contribution to the loss function, which was shown to be effective in numerical simulations (Strikis *et al.*, 2021; Czarnik *et al.*, 2022). It is also possible to use Monte Carlo sampling and a variational update of the parameters $\vec{\theta}$ to overcome the large size of $\mathbb{C}$ (Strikis *et al.*, 2021). Since it is possible to classically simulate circuits with a small number of non-Clifford gates, we can keep a few single-qubit gates in the primary circuit untouched when we define the training set. Alternative ways to truncate the Clifford set can be explored, for instance, based on their similarity to the primary circuits.

If the training set is large enough that the training can target a wide range of application circuits (computational tasks), then the training can be carried out at the device calibration stage and the training sampling overhead can be omitted when a particular computational task is considered. However, when the training set is small such that the training is targeting a small set of application circuits (computational tasks), then the training is best viewed as a part of a task itself and the training overhead needs to be included in any resource audit for that computational task. Note that a smaller training set usually also means a smaller sampling overhead for training. The relation between the size of the training set $|\mathbb{T}|$ and the performance of applying the trained result onto the target circuit $\boldsymbol{P}$ is rigorously developed only when the training uses the full Clifford set $\mathbb{C}$ for probabilistic error cancellation, as previously discussed. More general studies into the trade-off between them would be essential for different learning-based methods.

Thus far we have only explicitly talked about learning-based methods applied to rescaling and shifting the noisy result [Eq. (48)] and to probabilistic error cancellation. However, learning-based methods are in general compatible with almost all QEM methods that we have discussed, for instance, error extrapolation (Lowe *et al.*, 2021) and purification-based methods (Bultrini *et al.*, 2023). There are also suggestions that it can be applied to symmetry-based methods (Cai, 2021c). Some possible roles for learning-based methods in other QEM methods will be further discussed in the next section.

IV. COMPARISONS AND COMBINATIONS

A. Comparison among QEM methods

In Sec. III, we provided a detailed account of the individual QEM methods. To see the connections and differences among them more clearly and provide a discussion of their respective costs in a more coherent framework, Cai (2021b) divided the process of QEM into the following two stages:

(1) *Noise calibration* measures the strength of some given noise components.
(2) *Response measurement* measures how the observable of interest responds to changes in the noise components that were calibrated in the last step.

Combining the two components will inform us about how the observable of interest changes due to the presence of the calibrated noise components, and thus will enable us to construct an error-mitigated estimator protected from the calibrated noise components. We already discussed such a structure for QEM methods when introducing the learning-based methods (Sec. III.H); there the noise calibration process is simply the training process. The division between noise calibration and response measurement is not always clear cut. For most of the QEM methods that we have discussed, the error-mitigated estimator will be a linear combination of the results obtained from the response measurement, and the way to combine them (the weightings of each term) will be determined by the noise calibration. Treating QEM as a two-stage process enables us to discuss the costs of noise calibration and response measurement separately.

1. Noise calibration overhead

When previously talking about sampling overhead, we were referring mostly to the response measurement sampling overhead. We often assumed certain knowledge about the noise without discussing the cost of obtaining it. We now look more closely into the cost of the noise calibration for various types of QEM methods.

a. Gate error mitigation

These QEM methods target the *gate errors* in the primary circuit. Their noise calibration can simply be carried out using standard gate noise benchmarking and characterization techniques (Eisert *et al.*, 2020; Kliesch and Roth, 2021) such as gate infidelity estimation for zero-noise extrapolation (Sec. III.A), full gate error characterization for probabilistic error cancellation (Sec. III.B), and detector error characterization for measurement error mitigation (Sec. III.C). These can all be done in the device calibration stage. If this is the case, then ideally noise calibration is effectively free at the stage of QEM application. However, fully correlated noise models are exponentially expensive (in terms of qubit number) to characterize. Various ways to circumvent this were mentioned in Secs. III.B and III.C. Moreover, device parameters can drift in time, and thus routine recalibration might be needed. In such a case, the noise calibration cost is no longer negligible at the QEM application stage and low-cost device calibration techniques would be essential (Google Quantum AI *et al.*, 2020b).

TABLE I. Summary of the assumptions, costs, and performances associated with some of the QEM methods mentioned in this review. Some of the expressions of the sampling overhead and the fidelity boost are derived under the assumption that the occurrence of faults in the circuit follows a Poisson distribution with a circuit fault rate of λ, as discussed in Sec. II.C. We use ρ and ρ_0 to denote the unmitigated noisy state and the ideal noiseless state, respectively. Only a specific instance of error extrapolation is included here. Measurement error mitigation is not included here, since it can be viewed as a special case of probabilistic error cancellation focusing on measurement noise. Quantum subspace expansion and N representability are also not included, since their implementation costs and performance are highly problem specific.

Methods	Probabilistic error cancellation	Richardson extrapolation (equal gap[a])	Symmetry verification	Virtual distillation	Echo verification
Main assumptions	Full knowledge of the noise	Ability to scale the noise; small λ[b]	The ideal state contains inherent symmetry	The ideal state ρ_0 is pure; the noise is stochastic such that ρ_0 is the dominant eigenvector of the noisy state ρ[c]	
Hyperparameters	Circuit fault rate after mitigation: λ_{em}	Number of data points: M	Projector of the symmetry subspace: Π	Degree of purification: M	0
Qubit overhead	1	1	1[d]	M	1
Circuit run-time overhead	Up to ~ 2	Up to $\sim M$	1	~ 1[e]	2
Sampling overhead (C_{em})	$e^{4(\lambda-\lambda_{\mathrm{em}})}$	$(2^M-1)^2$	Postselection $\mathrm{Tr}[\Pi\rho]^{-1}$ Postprocessing: $\mathrm{Tr}[\Pi\rho]^{-2}$	$\mathrm{Tr}[\rho^M]^{-2} \gtrsim e^{2M\lambda}/[1+(e^{\lambda}-1)^M]^2$	$\mathrm{Tr}[\rho^2]^{-1} \gtrsim e^{2\lambda}/[1+(e^{\lambda}-1)^2]$
Fidelity boost[f]	$e^{\lambda-\lambda_{\mathrm{em}}}$	$e^{\lambda}+\mathcal{O}(\lambda^M)$	$\mathrm{Tr}[\Pi\rho]^{-1}$	$\mathrm{Tr}[\rho_0\rho]^{M-1}/\mathrm{Tr}[\rho^M] \gtrsim e^{\lambda}/[1+(e^{\lambda}-1)^M]$	Same as VD with $M=2$
Bias	Can reach 0 when $\lambda_{\mathrm{em}}=0$.	$\mathcal{O}(\lambda^M)$	Can be upper bounded using the error-mitigated fidelity using Eq. (53), which in turn is related to the achieved fidelity boost.		

[a]There are other variants of Richardson extrapolation that may provide better scaling for the overhead (Sidi, 2003).
[b]There are successful experiments operating beyond the small-λ assumption. The small-λ assumption is not needed for some other extrapolation methods (Sec. III.A).
[c]If ρ_0 is not the dominant eigenvector of the noisy state, the achieved ultimate fidelity will be limited by coherent mismatch (Sec. III.E).
[d]Assuming that only the inherent symmetries of the physical problems are used. Additional qubit overhead might be needed if we want to introduce additional symmetries.
[e]Additional circuit components are needed to swap in between the noisy copies and the corresponding additional run-time required will depend on the connectivity of the hardware.
[f]Fidelity boost is the factor of increase in the fidelity against the ideal state ρ_0 after applying error mitigation (Sec. IV.B.3).

b. State error mitigation

These QEM methods target the errors on the output state of the primary circuit. Since we do not know the exact form of the ideal state, we now try to probe errors that violate known constraints on the output state like symmetry constraints (Sec. III.D) and purity constraints (Sec. III.E). Their noise calibration will measure the strength of the noise that violates these constraints, for instance, the fraction of the circuit runs that fail the symmetry verification. In these cases, both noise calibration and response measurement involve measuring additional operators on the unmitigated noisy state. Hence, in some settings the noise calibration can be performed alongside response measurement by simply measuring additional operators without additional circuit runs (Bonet-Monroig *et al.*, 2018; Cai, 2021b,2021c), which means that the noise calibration is essentially free.

c. Observable error mitigation

Not all errors in the primary circuit will affect our observable of interest, and this class of QEM methods target only the error components that are damaging to our observable. Methods like subspace expansion (Sec. III.F) and N-representability (Sec. III.G) target error components that violate some given constraints on the noiseless observables (which can be the observable of interest or its components). For observable error mitigation, the noise calibration will be dependent on the observable of interest. Hence, the noise calibration accuracy required is highly dependent on the problems that we try to solve, and thus the associated cost has not been analytically derived.

As mentioned, learning-based methods (Sec. III.H) are a means to perform the noise calibration process using training circuits. Compared to the circuits used in the original noise calibration process for different QEM methods, the training circuits are usually more closely related to the primary circuit, so we can target faults that are more specific to the primary circuit and that optimistically can reduce the calibration cost. For example, when trying to construct the training circuit, we can make use of the structure of the primary circuit for gate error mitigation, or we can make use of the known observable of interest for state error mitigation. The training cost for learning-based methods is thus highly problem specific and usually hard to analytically quantify, just like the cost for noise calibration in observable error mitigation.

The aforementioned categorization of QEM methods is mainly for providing more intuition and does not represent a definitive guide. For example, the RDM measured in the N-representability method can be viewed as an observable, in which case the method is seen as a type of observable error mitigation. Alternatively, it can be viewed as a state in a reduced subspace such that the method is a form of state error mitigation. After this discussion of noise calibration requirements, we now move on to more general comparisons between QEM methods beyond the noise calibration stage.

2. Mean square errors

As discussed in Sec. II.B, the most straightforward metric for comparing two different error-mitigated estimators is simply to compare their mean square errors [Eq. (1)]. However, general comparisons between two QEM methods cannot be made in a similar manner, because for each QEM method there is a wide range of error-mitigated estimators that we can construct using different hyperparameters (the degree of purification for purification-based QEM, the number of symmetries in symmetry-based QEM, the number of data points in zero-noise extrapolation, etc.). Each of these estimators has a different trade-off between bias and variance and thus a different mean square error. Comparisons between two QEM methods are further complicated by their different sets of assumptions (whether there are known constraints to states, knowledge about the noise, etc.) and different hardware requirements (such as qubit number and connectivity). Hence, to compare two QEM methods we often need to know the exact primary circuit whose noise we are trying to mitigate and the set of experimental constraints that we need to adhere to (on things like qubit numbers, run-time, and hardware error rate), such that we can specify exact implementations of the two given QEM methods and deduce the corresponding mean square errors. Even in such cases the bias and variance often can be calculated analytically for only some canonical implementations. Hence, in most of the QEM literature mentioned in this review, when trying to assess or compare the performance of QEM methods, their bias and variance are often obtained using numerical simulations or physical experiments by applying them to a specific use case. This is not a good indicator of the general performance of a given QEM method, but it can usually demonstrate key characteristics of the QEM methods such as their target noise types and their scaling behavior.

While recognizing that it is difficult to make general comparisons between different QEM methods, nevertheless we summarize the costs and performance of some *canonical implementations* of typical QEM methods in Table I. The intention is to capture the essential distinguishing features of these QEM approaches.

B. Benchmarking QEM from other perspectives

Besides computing the mean square error for a specific implementation of a given QEM method, we can also look into other general characteristics of the QEM method by viewing the process of QEM from another perspective.

1. State discrimination

QEM allows us to better estimate the expectation values of the various noisy states output from a noisy system. However, the data-processing inequality never allows the distinguishability between these noisy states to increase (Nielsen and Chuang, 2010). This fact was used by Takagi *et al.* (2022) to obtain explicit bounds on the range that the error-mitigated estimator may take, which in turn determined the number of samples needed to achieve a given level of performance.

They considered an error-mitigation protocol in which the first step was obtaining M noisy copies of the error-free state ρ_0 through a process denoted as $\mathcal{L}$,

$$\mathcal{L}(\rho_0) = \bigotimes_{m=1}^{M} \mathcal{E}_m(\rho_0),$$

where $\{\mathcal{E}_m\}$ are the different effective noise channels acting on different copies. The second step is constructing an error-mitigated observable O_{em} based on the QEM technique and the observable of interest O such that measuring O_{em} on these noisy copies will output the error-mitigated estimator $\hat{O}_{\mathrm{em}}$,

$$E[\hat{O}_{\mathrm{em}}] = \mathrm{Tr}[O_{\mathrm{em}}\mathcal{L}(\rho_0)].$$

The maximum bias that can be achieved by the QEM method for all possible observables of interest O and target states ρ_0 is given as

$$B_{\max} = \max_{O,\rho_0} \mathrm{bias}[\hat{O}_{\mathrm{em}}] = \max_{O,\rho_0} \big[\mathrm{Tr}[O_{\mathrm{em}}\mathcal{L}(\rho_0)] - \mathrm{Tr}[O\rho_0]\big].$$

Using this, Takagi *et al.* (2022) proved the following lower bound for the range of the error-mitigated estimator:

$$R[\hat{O}_{\mathrm{em}}] \geq \max_{\rho_0,\sigma_0} \frac{D_{\mathrm{tr}}(\rho_0,\sigma_0) - 2B_{\max}}{D_{\mathrm{tr}}(\mathcal{L}(\rho_0),\mathcal{L}(\sigma_0))}, \tag{50}$$

where D_{tr} denotes the trace distance. The range $\mathrm{R}[\hat{O}_{\mathrm{em}}]$ can be used to obtain the sufficient number of samples required via Eq. (4). A scaling relationship similar to $R[\hat{O}_{\mathrm{em}}]$ for the case of depolarising noise was proven by Wang, Czarnik *et al.* (2021).

Building on the aforementioned framework, Takagi, Tajima, and Gu (2023) further obtained the explicit lower bound for the number of samples M necessary for achieving the target accuracy δ (this deviation includes both the bias and the shot noise) with success probability $1-\epsilon$,

$$M \geq \max_{\substack{\rho_0,\sigma_0 \\ D_{\mathrm{tr}}(\rho_0,\sigma_0)\geq 2\delta}} \min_{\mathcal{E}\in\{\mathcal{E}_m\}} \frac{2(1-2\epsilon)^2}{\ln 2 S(\mathcal{E}(\rho_0)||\mathcal{E}(\sigma_0))}. \tag{51}$$

In Eq. (51) $S(\rho\|\sigma) = \mathrm{Tr}(\rho\log\rho) - \mathrm{Tr}(\rho\log\sigma)$ is the relative entropy, which is related to the trace distance through the quantum Pinsker's inequality $D_{\mathrm{tr}}(\rho,\sigma) \leq \sqrt{\ln 2 S(\rho\|\sigma)/2}$ (Hiai, Ohya, and Tsukada, 2008). It has been shown that the relative entropy between two output states from noisy circuits under local depolarizing noise decreases exponentially with the circuit depth (Kastoryano and Temme, 2013;

Müller-Hermes, França, and Wolf, 2016), which implies that the sampling cost in Eq. (51) will grow exponentially with the circuit depth in this case. Note that by using fidelity instead of relative entropy a bound tighter than Eq. (51) can also be obtained, as discussed by Takagi, Tajima, and Gu (2023).

Quek *et al.* (2022) managed to construct a circuit structure for which the relative entropy of two different output states decreases exponentially in both circuit depth and the number of qubits, which allowed them to prove that the worst-case sampling lower bound for QEM scaled exponentially with the circuit size (instead of just depth), confirming our intuition obtained in Sec. II.D. They further showed that, for a geometrically local circuit, this exponential scaling is determined by the number of gates in the light cone of observables rather than the circuit size. Note that the bounds obtained in this section are more for the purpose of demonstrating the fundamental limits on the performance of a given QEM setup than for use as a metric for comparing the practical performance of different QEM methods.

2. Quantum estimation theory

In estimation theory, the variance of any unbiased estimator can be lower bounded by the inverse of the Fisher information through the Cramér-Rao inequality. Without loss of generality, we assume that our goal is to estimate $\mathrm{Tr}(O\rho_0)$ for some traceless observable O, but the state is suffering from the noise channel $\mathcal{E}$. We can still obtain the ideal expectation value if we have a way to measure the operator $Y = \mathcal{E}^\dagger(O)$ on the noisy state $\mathcal{E}(\rho_0)$. Watanabe, Sagawa, and Ueda (2010) showed that the quantum Fisher information for estimating the observable O using the noisy state $\mathcal{E}(\rho_0)$ is given by $J_O = \{\mathrm{Tr}[\mathcal{E}(\rho_0)Y^2] - \mathrm{Tr}[\mathcal{E}(\rho_0)Y]^2\}^{-1}$. When one uses the quantum Cramér-Rao inequality, the number of samples M needed for evaluating the noiseless estimator with additive error ϵ can then be bounded as

$$M \geq \frac{1}{\epsilon^2 J_O}. \tag{52}$$

Thus, Watanabe, Sagawa, and Ueda (2010) concluded that the cost-optimal strategy for obtaining the unbiased estimator of $\mathrm{Tr}(O\rho_0)$ from the noisy state $\mathcal{E}(\rho_0)$ is simply to measure $\mathcal{E}(\rho_0)$ with the observable Y.

Tsubouchi, Sagawa, and Yoshioka (2023) applied these arguments to the NISQ scenario with the estimator as an error-mitigated estimator and obtained two lower bounds on the sampling cost of QEM. Their first lower bound is for generic layered quantum circuits under a wide class of Markovian noise. By showing that the quantum Fisher information decays exponentially with the depth of the layered quantum circuit, they demonstrated the exponential growth of the sampling overhead as the depth of the circuit increases. In the special case of global depolarizing errors, the required number of samples M for achieving an additive error ϵ in the standard deviation can be lower bounded as $M \gtrsim [\mathrm{Tr}(O^2)/2^N\epsilon^2](1-p)^{-2L}$, which can be saturated by simply rescaling the measurement result by a factor of $(1-p)^{-L}$; see Sec. III.H. For random quantum circuits under local noise, they further showed that the sampling cost grows exponentially with both the circuit depth and the number of qubits through analytical arguments and numerical simulations. The scaling behavior found here is consistent with Sec. IV.B.1.

3. State extraction

In symmetry-based or purification-based QEM, it is natural to think of QEM as a process of extracting the symmetry-verified or purified state out of the noisy state (Bonet-Monroig *et al.*, 2018; McArdle, Yuan, and Benjamin, 2019; Huggins, McArdle *et al.*, 2021; Koczor, 2021b). More generally, in most of the QEM methods the error-mitigated expectation value $E[\hat{O}_{\mathrm{em}}]$ is a linear function of the observable of interest O, which can be written as

$$E[\hat{O}_{\mathrm{em}}] = \mathrm{Tr}[O\rho_{\mathrm{em}}].$$

Here ρ_{em} can be viewed as the error-mitigated component that we try to extract out of the noisy state using the given QEM method (Cai, 2021b).

A higher fidelity of the error-mitigated state ρ_{em} against the ideal state ρ_0, denoted as $F(\rho_0, \rho_{\mathrm{em}})$, has been shown to correspond to lower biases in various numerical simulations (Cai, 2021c; Huggins, McArdle *et al.*, 2021; Koczor, 2021b). More exactly, using the results of Koczor (2021a) and the Fuchs–Van de Graaf inequality (Fuchs and Van de Graaf, 1999), the bias after error mitigation can be bounded by

$$\begin{aligned}\mathrm{bias}[\hat{O}_{\mathrm{em}}] &\leq 2\|O\|_\infty D_{\mathrm{tr}}(\rho_0, \rho_{\mathrm{em}}) \\ &\leq 2\|O\|_\infty \sqrt{1 - F(\rho_0, \rho_{\mathrm{em}})},\end{aligned} \tag{53}$$

where $\|O\|_\infty$ is the largest absolute eigenvalue of O.

As discussed by Cai (2021b), by assuming the ideal state ρ_0 to be a pure state we can decompose the noisy state ρ into the error-mitigated component ρ_{em} and an erroneous component ρ_{err} (not necessarily a valid density matrix) that is orthogonal to the ideal state ρ_0 ($\mathrm{Tr}[\rho_0\rho_{\mathrm{err}}] = 0$),

$$\rho = p_{\mathrm{em}}\rho_{\mathrm{em}} + (1 - p_{\mathrm{em}})\rho_{\mathrm{err}}. \tag{54}$$

In Eq. (54) p_{em} is the amount of the error-mitigated component contained in the noisy state. Only a partial amount of this error-mitigated component can be successfully extracted using the given QEM method; this amount is denoted as q_{em}, with $q_{\mathrm{em}} \leq p_{\mathrm{em}}$.

In this picture of extracting error-mitigated states, the factor of improvement in the fidelity, which we call the *fidelity boost* here, is given by

$$B_{\mathrm{em}} = \frac{\mathrm{Tr}[\rho_0\rho_{\mathrm{em}}]}{\mathrm{Tr}[\rho_0\rho]} = p_{\mathrm{em}}^{-1}$$

(where ρ_0 is assumed to be pure), and the corresponding sampling overhead is given as

$$C_{\mathrm{em}} \sim q_{\mathrm{em}}^{-2}. \tag{55}$$

Cai (2021b) gave a list of examples showing how p_{em} and q_{em} can be calculated to obtain B_{em} and C_{em}, with some of the results shown in Table I.

The fraction of the error-mitigated state that is successfully extracted using the QEM method is called the extraction rate,

$$r_{\mathrm{em}} = \frac{q_{\mathrm{em}}}{p_{\mathrm{em}}} = \frac{B_{\mathrm{em}}}{\sqrt{C_{\mathrm{em}}}}.$$

This is simply the ratio between the fidelity boost and the square root of the sampling overhead, and thus is a natural indicator for the cost effectiveness of a given QEM method. Successful extraction of all of the error-mitigated components contained in the noisy state corresponds to the maximum extraction rate $r_{\mathrm{em}} = 1$, which can be achieved by symmetry verification (Cai, 2021b), indicating its cost effectiveness. Note that if we are able to perform direct postselection instead postprocessing to extract the error-mitigated state (as in direct symmetry verification and echo verification), then the sampling overhead will be $C_{\mathrm{em}} \sim q_{\mathrm{em}}^{-1}$ instead and the extraction rate will be $r_{\mathrm{em}} = B_{\mathrm{em}}/C_{\mathrm{em}}$. As in Sec. IV.B.1, the arguments in this section also do not apply to QEM techniques that nonlinearly combine the results from response measurements.

C. Combinations of QEM methods

Rather than trying to pick a better QEM method, one might instead try to combine different QEM methods in the hope of being able to target more noise components and/or achieving a better trade-off between bias and variance. We mentioned some possible connections and combinations of QEM methods while discussing individual techniques in Sec. III, and we provide additional insights in this section.

The simplest way of combining different QEM methods is applying them in parallel such that each of them targets a different noise source in the circuit. For example, using zero-noise extrapolation or probabilistic error cancellation to target noise in the main computation while using measurement error mitigation to tackle noise at the measurement stage. For a circuit consisting of Pauli-symmetry-preserving components affected by Pauli noise, we can detect and partially mitigate the circuit faults that anticommute with the symmetry using symmetry verification, while the rest of the circuit faults that commute with the symmetry and thus are immune to symmetry verification can be mitigated using zero-noise extrapolation (through only scaling this "commuting" noise) or probabilistic error cancellation (Cai, 2021a). As discussed in Sec. V.C, QEM can be used to mitigate errors in the circuit compilation. Hence, it is possible to use one QEM method to mitigate these compilation errors while using another QEM method to mitigate noise in the circuit. Note that we can also apply a given QEM method multiple times, each time targeting a different noise source, which has been demonstrated for zero-noise extrapolations (Otten and Gray, 2019).

When the different QEM methods applied interfere with each other rather than working entirely independently, their order of application becomes important. The QEM method that is first applied will be called the *base QEM method*. In a sense, applying one QEM method after another can be viewed as "concatenating" the QEM methods, and the overall sampling overhead is simply the product of the sampling overheads of both stages of QEM. If we want to apply symmetry- or purification-based QEM along with another QEM method, we can view the base QEM method simply as a process of extracting error-mitigated states out of the noisy state, as discussed in Sec. IV.B.3, and then apply symmetry constraints or purity constraints to this error-mitigated state (Cai, 2021b). In this way, we are able to remove the remaining symmetry- or purity-violating noise that was not targeted by the base QEM method. In a similar way, for a given error-mitigated state we can also perform subspace expansion around it or we can measure its RDM, which means that we can directly perform a subspace expansion or N representability in addition to other QEM methods. Google Quantum AI *et al.* (2020a) experimentally demonstrated the application of N representability along with symmetry verification.

One can apply zero-noise extrapolation with another QEM method by performing extrapolation using error-mitigated expectation values (McArdle, Yuan, and Benjamin, 2019). However, the error models of the circuit can be altered by the base QEM method such that the noise scaling factor might no longer be well defined. Furthermore, the shape of the extrapolation curves may be altered by the base QEM method. It is possible to circumvent this by using probabilistic error cancellation as the base QEM method since it can be applied partially without changing the error models, yielding data points of reduced noise strength for extrapolation, as mentioned in Sec. III.A (Cai, 2021a). For other base QEM methods, as long as we know the effective noise scaling factor and the circuit fault rate is small after the base QEM, Richardson extrapolation will still be applicable; for instance, it has been applied along with purification-based methods in numerical simulations (Koczor, 2021b) and together with instances of probabilistic error cancellation that changes the error models (Sun *et al.*, 2021). In some specific use cases, the new shape of the extrapolation curve can be analytically deduced. One such example was given by Cai (2021a), with symmetry verification being the base QEM method. More generally, we can try to use learning-based methods to predict the shape of the extrapolation curves after performing the base QEM, especially in the special case of rescaling and shifting the noisy expectation value; see Sec. III.H. Applying rescaling in addition to symmetry verification has been demonstrated in numerical simulations (Stanisic *et al.*, 2022) and experiments (Google Quantum AI *et al.*, 2020b). As mentioned in Sec. III.H, rescaling and shifting can be viewed as performing probabilistic error cancellation for mitigating global depolarizing noise. However, probabilistic error cancellation in a more general context is difficult to perform alongside other QEM methods because the effective error model in the circuit will change due to the base QEM method.

Learning-based methods can be viewed as a way to perform the noise calibration process rather than a stand-alone QEM method, as mentioned in Sec. IV.A.1. Hence, it can be easily combined with the other methods by replacing their noise calibration process (Cai, 2021b; Lowe *et al.*, 2021), which we have mentioned in several examples. We can also use learning-based methods to obtain the optimal hyperparameters for various QEM methods (Bultrini *et al.*, 2023).

There were also attempts to construct a unified framework containing hyperparameters that can lead to different existing QEM methods for different values taken. Such frameworks give us access to a wide range of other possible QEM implementations by taking hyperparameters values sitting between existing QEM methods. We can think of this as interpolating between different QEM methods instead of concatenating different QEM methods. These new "interpolated" QEM implementations can achieve different bias-variance trade-offs beyond the existing QEM methods and often also target a different set of noise components. This allows us to choose the implementation that best fits our target problem and experimental constraints rather than being restricted to the existing QEM methods. We mentioned some of these examples in Sec. III, for instance, the combination of purification-based methods with subspace expansion (Yoshioka *et al.*, 2022), the generalization of symmetry verification and subspace expansion (Cai, 2021c), and the combination of zero-noise extrapolation and probabilistic error cancellation (Mari, Shammah, and Zeng, 2021). Some QEM combinations mentioned earlier in this section can also be viewed as attempts to construct a unified framework (Lowe *et al.*, 2021; Bultrini *et al.*, 2023). Note that many QEM methods can also be natively combined with shadow tomography (Jnane *et al.*, 2023), which may provide performance advantages when we want to estimate a large number of observables.

D. Comparison to the other error-suppression methods

Quantum error mitigation is one family of strategies for suppressing errors. It is natural to consider how it relates to other well-known error-suppression methods, such as decoherence-free subspace or subsystem (DFS), dynamical decoupling (DD), and QEC. Symmetry is at the heart of all the methods mentioned here; thus, we first try to compare QEC against symmetry verification to gain some intuition.

In QEC, we perform regular symmetry (stabilizer) measurements to detect and correct errors accumulated in the quantum system. If we discard the quantum states that violate these symmetry checks instead of correcting them, we have quantum error detection instead. Symmetry verification (Sec. III.D) can usually be seen as an error-detection scheme whose symmetries are given by the physical problem of interest, and the detection can be carried out via postprocessing. Such native symmetries and the possible use of post-processing detection typically mean that many fewer qubits and a much lower gate fidelity are required for symmetry verification than for QEC to achieve the "break-even" point of performing better than the uncorrected or unmitigated circuit. Furthermore, performing logical operations in QEC code usually comes with higher space-time overhead than the physical operations performed in QEM, leading to a longer circuit run-time and even more qubit overhead for QEC.

Even though we are talking about symmetry verification in this comparison, much of the intuition provided is applicable to general QEM methods. As mentioned in Sec. I, general QEM methods aim to correct the output distribution in an ensemble of circuit runs, while QEC aims to correct the output in every circuit run. If we are trying to remove almost all bias as in fault-tolerant QEC, the sampling overhead of general QEM will grow exponentially with the circuit size, as discussed in Sec. II.D. Using all of the aforementioned arguments, we can compile a rough guide for the differences between QEC and QEM, as outlined in Table II. Note that there are intersections between QEC and symmetry-based QEM like postprocessing quantum error detection and adding more symmetries to QEM (at the cost of adding more qubits), as discussed in Sec. III.D. Adding to what we have discussed, Cao *et al.* (2022) also looked at the difference between QEM and QEC in the setting of communication instead of computation.

Without knowing the specific application task, it is difficult to compare the amount of resources required for QEC and QEM. There have been attempts to estimate the resource required to obtain *classically intractable* Fermi-Hubbard ground state energy using VQE with the help of symmetry verification and error extrapolation (Cai, 2020). To overcome the sampling overhead, which is partly due to QEM, it was estimated that more than 100 independent quantum processors are needed for parallelization, each containing 50 qubits and having a gate error rate of 10^{-4}. The gate fidelity and the total number of qubits required here are actually similar to those required for solving the problem using a fault-tolerant algorithm (Kivlichan *et al.*, 2020). However, given the same total number of qubits, building hundreds of identical quantum processors (cores) using commercial fabrication techniques is usually easier than building the single large integrated processor required for the fault-tolerant algorithm. Note that what we have discussed here is not a sheer comparison between QEC and QEM, but rather more of a comparison between a NISQ algorithm with the help of QEM and a fault-tolerant algorithm for a specific application.

TABLE II. Rough guide to the differences between QEC and QEM. The difference will be dependent on the QEC codes, the QEM method, and the exact application scenario.

	QEC	QEM
Qubit overhead	High	Low
Circuit run-time	Often scales with code distance	Like the unmitigated circuit
Sampling overhead	Constant	Exponential in the circuit size[a]
Error rate	Must be below code threshold	Important to keep it low; no threshold[b]
Midcircuit measurement	Essential and frequent	Infrequent or not required

[a]Assuming that we try to construct a nearly unbiased estimator as in QEC and with fixed gate error rates.
[b]Lower error rate means smaller sampling overheads. No threshold for the gate noise in the unmitigated circuit, but there might be requirements on the additional gates needed for QEM.

In practice, QEC and QEM are more likely to be complementary rather than competing methods since in many practical applications we can apply QEM in addition to QEC, as further discussed in Sec. V.B. It is also possible to use QEM to mitigate compilation errors, which is not possible using QEC, as discussed in Sec. V.C.

Unlike QEC and QEM, both DD and DFS methods are *open-loop* quantum control techniques whose action does not depend on the status of the quantum state, and thus no measurements are needed for their implementation (Lidar, 2014; Suter and Álvarez, 2016). In fact, DFS is a passive technique that requires no action at all: we need only choose the correct symmetry subspace that is immune to the target set of noise. QEC and symmetry-based QEM all have some element of DFS in them, in the sense that they are all immune to noise generated by the symmetry operators. Conversely, DD is achieved by applying a sequence of decoupling pulses to "average out" the harmful interaction between the quantum system and the environment. In the limit of many rounds of instantaneous decoupling pulses, group-based DD is shown to be equivalent to making continuous symmetry measurements to achieve the quantum Zeno effect (Facchi, Lidar, and Pascazio, 2004; Burgarth *et al.*, 2019). Such a DD sequence can be applied using the native symmetry of the physical problem as in symmetry-based QEM, which has been shown to be effective in quantum simulations (Tran *et al.*, 2021).

V. APPLICATIONS

Thus far we have discussed QEM mostly in the general context of expectation value estimation without overspecifying the application scenarios. There is a range of software packages for carrying out the QEM techniques that we have mentioned. For example, QERMIT, MITIQ, and QISKIT each provide its own implementation of zero-noise extrapolation, probabilistic error cancellation, and measurement error mitigation (LaRose *et al.*, 2022; Cirstoiu *et al.*, 2023; Qiskit Contributors, 2023c). In addition, QERMIT and MITIQ contain the Clifford variants of the learning-based method. There is also a wide range of other packages providing measurement error mitigation like `QCompute`, `PennyLane`, and QREM (Bergholm *et al.*, 2018; Maciejewski *et al.*, 2020; Baidu, 2023a).

Through numerical experiments, QEM has proven effective in a wide range of applications including linear equation solvers (Vazquez, Hiptmair, and Woerner, 2022), quantum metrology (Yamamoto *et al.*, 2022; Conlon *et al.*, 2023), Monte Carlo simulations (Yang, Lu, and Li, 2021), and numerous other examples mentioned in Sec. III. Moreover, many physical experiments have successfully employed QEM for noise suppression, with some pioneering and state-of-the-art examples summarized in Table III, spanning across different application scenarios such as Fermi-Hubbard models (Stanisic *et al.*, 2022) and multiparticle bound states (Google Quantum AI *et al.*, 2022), with a circuit size of up to 127 qubits and 60 layers of two-qubit gates (Kim *et al.*, 2023). Different application scenarios and hardware platforms will give rise to different types of noise. In this section, we look into ways to adapt the implementation of QEM to different noise types and discuss the expected results.

A. Coherent errors

The degree of coherence in an error channel can be quantified by how well it preserves the purity of an average incoming state (Wallman *et al.*, 2015). In this sense the "most coherent" errors are simply unitary errors, while incoherent errors usually refer to stochastic Pauli channels. More generally the term *coherent errors* refers to a broad spectrum of noise that sits closer to the unitary errors than the Pauli errors. There are cases in which coherent errors can be mitigated using coherent control solutions such as dynamical decoupling (Lidar, 2014; Suter and Álvarez, 2016). The remnant coherent errors can be mitigated using most of the aforementioned QEM techniques (with the exception of purity-based methods). In fact, subspace expansion (Sec. III.F) was originally constructed to target coherent

TABLE III. Examples of pioneering and state-of-the-art experimental applications of QEM. Note that measurement error mitigation is present in almost all experimental work and thus is not explicitly included here. SC, superconducting qubits; ion, trapped-ion qubits; ZNE, zero-noise extrapolation; PEC, probabilistic error cancellation; SYM, symmetry constraints; PUR, purity constraints; SUB, subspace expansion; NRP, *N* representability; LEA, learning-based methods.

QEM methods	Platform	Applications	Reference(s)
ZNE	SC	Variational eigensolver	Dumitrescu *et al.* (2018) and Kandala *et al.* (2019)
ZNE	SC	Real-time dynamic simulation	Kim *et al.* (2023) and Kim, Wood *et al.* (2023)
ZNE	Ion	Entanglement entropy measurement	Foss-Feig *et al.* (2022)
PEC	SC	Deterministic quantum computation with pure states	Song *et al.* (2019)
PEC	SC	Real-time dynamic simulation	Van den Berg *et al.* (2023)
PEC	Ion	Gate fidelity estimation	Zhang *et al.* (2020)
PEC	Ion	Real-time dynamic simulation	Chen *et al.* (2023)
SYM	SC	Variational eigensolver	Sagastizabal *et al.* (2019)
SYM	Ion	Variational state preparation	Zhu *et al.* (2020)
PUR	SC	Variational eigensolver	O'Brien *et al.* (2023)
SUB	SC	Variational eigensolver	Colless *et al.* (2018)
LEA	SC	Variational eigensolver	Dborin *et al.* (2022)
LEA	SC	Quantum information scrambling	Mi *et al.* (2021)
SYM and NRP	SC	Variational state preparation	Google Quantum AI *et al.* (2020a)
SYM and LEA	SC	Variational eigensolver	Stanisic *et al.* (2022)
SYM and LEA	SC	Real-time dynamic simulation	Google Quantum AI *et al.* (2020b)

errors (McClean *et al.*, 2017). However, for many other QEM methods Pauli errors can be far easier to mitigate than coherent errors, as previously discussed. The advantages of mitigating Pauli noise are summarized as follows:

- *Zero-noise extrapolation*. Arguments can be made that the expectation value of a Pauli observable should decay following a multiexponential curve under Pauli noise. Knowing the form of the extrapolation curve can lead to lower sampling costs and smaller biases.
- *Probabilistic error cancellation*. Pauli noise requires less overhead to be characterized and removed using the standard basis, in a manner similar to measurement error mitigation.
- *Symmetry constraints*. When one deals with symmetry-preserving circuit components and Pauli symmetry, analytical arguments can be made about the proportion of noise removed when considering Pauli noise. This also enables us to select a good set of symmetries to use.
- *Purity constraints*. Errors that are less coherent can improve the ultimate accuracy that can be reached using purity constraints.
- *Learning-based methods*. Standardizing error channels by transforming them into Pauli noises will result in a more similar error model between the training circuits and target circuits. The effects of Pauli noise on the observable are usually easier to characterize and thus require less training.

Even for subspace expansion, though Pauli errors might not be easier to mitigate, they are certainly easier to analyze and thus enable us to select a better set of expansion operators. Furthermore, coherent errors usually accumulate at a faster rate than their Pauli counterpart (Gutiérrez and Brown, 2015; Sanders, Wallman, and Sanders, 2015; Kueng *et al.*, 2016). Hence, one of the most effective ways to deal with coherent errors in the context of QEM is simply to transform them into Pauli errors.

Any arbitrary quantum channel can be transformed into a Pauli channel using *Pauli twirling*, which originated from entanglement purification (Bennett, Brassard *et al.*, 1996; Bennett, DiVincenzo *et al.*, 1996; Knill, 2004) and is widely used in areas like quantum benchmarking (Kliesch and Roth, 2021) for transforming quantum states or noise channels into a standardized form. When applying Pauli twirling to environmental noise or noise from Clifford gates, we simply insert random Pauli gates during the circuit compilation (as discussed in Appendix A.2). Most of these random Pauli gates can in fact be absorbed into the existing Pauli gates (Wallman and Emerson, 2016), and thus it can often be implemented with a relatively small gate cost. It is also possible to twirl over the Clifford group, which will further homogenize the probability of different Pauli errors, resulting in a depolarizing channel, which is even easier to analyze and mitigate. However, Clifford gates, especially multiqubit ones, can be much more difficult to implement than Pauli gates.

B. Logical errors in fault-tolerant quantum computation

The ultimate goal of quantum computation is to have a fully scalable fault-tolerant quantum system in which we can suppress the logical errors to an arbitrary small level by increasing the size of the system. However, before we achieve that it is likely that there will be an extended period of time in which quantum error correction is successfully implemented, but the minimum logical error rate achievable is still substantial due to reasons such as hardware limitations in the system size, challenges in large-scale decoding, and insufficient resources for magic state distillation. In this "early fault-tolerant" era, many tasks of interest requiring an effectively zero error rate will be impossible to perform. As long as the results of the fault-tolerant algorithm are obtained through expectation value estimation, the residual logical errors can be mitigated using all of the aforementioned QEM methods. We can follow the same arguments outlined previously and simply replace the qubit registers, the operations, and the error channels with their logical equivalents.

In particular, the role of probabilistic error cancellation in the Clifford $+ T$ paradigm of universal fault-tolerant quantum computation was studied extensively in several works (Lostaglio and Ciani, 2021; Piveteau *et al.*, 2021; Suzuki *et al.*, 2022). The logical noise associated with Clifford gates can be transformed into logical Pauli channels using Pauli twirling by inserting random Pauli gates, as mentioned in Sec. V.A. To mitigate the resultant logical Pauli noise using probabilistic error cancellation, we need only insert Pauli gates into the unmitigated circuit. These additional logical Pauli gates required for error mitigation can be applied virtually and noiselessly by updating the Pauli frame (Suzuki *et al.*, 2022), thereby boosting the effectiveness of probabilistic error cancellation in the fault-tolerant setting. The errors in the noisy logical T gate can be mitigated in a similar way, but to twirl them we need to apply additional logical Clifford gates instead of just Pauli gates; see Appendix A.2. Suzuki *et al.* (2022) showed that to achieve a logical circuit fault rate of 10^{-3} with a sampling overhead of 100, probabilistic error cancellation can reduce the physical qubit overhead by 80% for some classically intractable problems and more than 45% in many practical fault-tolerant applications.

The magic state distillation process needed for implementing fault-tolerant T gates accounts for significant costs in fault-tolerant computation even with recent advances (O'Gorman and Campbell, 2017; Litinski, 2019). Therefore, it makes sense to study the details of error mitigation for imperfect logical T gates. Piveteau *et al.* (2021) considered logical T gates implemented through either gate teleportation or code switching and found that it is sufficient to consider the effective logical error as a dephasing channel (with the help of logical twirling), which is much easier to characterize and mitigate using probabilistic error cancellation. Assuming perfect Clifford gates, they showed that one can use probabilistic error cancellation to remove errors in a circuit that has 2000 T gates and a physical error rate of 10^{-3} at a sampling overhead of 1000. This is well beyond the classically tractable regime of $\sim$50 T gates (Bravyi and Gosset, 2016). Lostaglio and Ciani (2021) discussed the application of QEM to noisy logical T gates but with more emphasis on the resource theoretical aspect. They introduced a resource measure called quantum-assisted robustness of magic, which indicates the speedup of quantum circuit simulation with noisy non-Clifford gates using probabilistic error cancellation.

Besides gate errors, there will also be compilation errors when we try to approximate an arbitrary unitary gate using the Clifford + T gate set. The form of the resultant compilation errors can be analytically computed for local gates, which can then be mitigated using probabilistic error cancellation (Suzuki *et al.*, 2022). The resultant compilation error will decrease exponentially with the increase of the number of T gates used (Kitaev, 1997; Dawson and Nielsen, 2006; Kliuchnikov, Maslov, and Mosca, 2013; Ross and Selinger, 2016). Hence, there is a trade-off between the number of T gates used and the cost of the QEM we need to apply. We discuss more instances of compilation errors beyond the context of gate synthesis in Sec. V.C.

C. Algorithmic (compilation) errors

To execute a unitary operation, we first need to compile it into a sequence of gates that the quantum computer can efficiently execute. The compiled circuit does not always exactly represent the target unitary (often due to a limited circuit depth); any resultant mismatches are called *compilation errors*. This will lead to errors in the output expectation value that we call *algorithmic errors*. We discussed one instance of compilation errors in the context of gate synthesis in Sec. V.B. In applications like variational eigensolvers, the target unitary is represented using a parametrized ansatz circuit, which can lead to algorithmic errors due to incorrect values of parameters or limited representation power of ansatz circuits. Algorithmic errors will be present even if all gates can be executed perfectly without any noise, and thus they cannot be removed through QEC. However, algorithmic errors can be removed through some of the QEM techniques that we have mentioned, and this is one of the areas where QEM is nontrivially different from QEC.

Endo *et al.* (2019) made the first attempt to mitigate such algorithmic errors via zero-noise extrapolation, which essentially operates in the manner described in Sec. III.A, but with algorithmic errors in the place of physical errors. For a given application, there will be different ways to compile the target unitary using different circuit structures (for instance, different circuit depths) or different circuit parameters, which will output data points with different algorithmic errors. If we know the relative scaling of the algorithmic errors between these data points, we can then apply zero-noise extrapolation to them to remove the algorithmic errors. Endo *et al.* (2019) looked at the specific problem of Hamiltonian simulation using Trotterization (Suzuki, 1990, 1991). Higher-order Trotterization will have fewer algorithmic errors but deeper circuits. Balancing these two aspects means there is an optimal order of Trotterization in a given practical setting. Data points of different algorithmic errors can be obtained using different orders of Trotterization up to the optimal order, using which we can apply Richardson extrapolation to estimate the result of infinite-order Trotterization with zero algorithmic errors. Zero-noise extrapolation can also be used to mitigate algorithmic errors in energy minimization algorithms (quantum optimization algorithms) like quantum annealing and variational eigensolver by extrapolating to the infinite-annealing-time limit or the zero-energy-variance limit (Cao *et al.*, 2023).

Mitigating algorithmic errors is even more natural for QEM methods that are based on the known constraints of the ideal state or the ideal observables since algorithmic errors can violate these constraints, just like gate errors. One such example was discussed by Huggins, McArdle *et al.* (2021). They looked at the randomized Trotterization (Campbell, 2019; Childs, Ostrander, and Su, 2019; Ouyang, White, and Campbell, 2020) in which the compiled circuit is inherently probabilistic while the ideal state is known to be pure. Hence, it is natural to apply purification-based QEM here to remove the stochastic errors due to the randomized compilation process. Another example is subspace expansion applied in the context of a variational eigensolver (McClean *et al.*, 2017) to overcome the limited representation power of the ansatz circuit.

VI. OPEN PROBLEMS

A. Overarching problems

Having thus surveyed the diverse concepts, implementations, and applications of QEM, it is now appropriate to reflect on the key unanswered questions in the field. We identify several such questions next.

(1) *What is the full landscape of the zoo of QEM?* In Sec. III, we discussed a set of QEM methods and their variants, which are grouped into three categories in Sec. IV.A.1. However, this is by no means the full landscape of QEM. Are there better classifications of the different QEM methods? Moreover, are there new methods waiting to be discovered?

(2) *What are some good performance metrics for QEM?* In Sec. II.B, we identified the bias and variance of the error-mitigated estimator as its key performance metrics, and in Sec. IV we further defined some performance metrics by viewing QEM from other perspectives. However, these metrics are greatly dependent on the exact problem we are trying to solve, and many of them can be obtained analytically for only some canonical cases. Are there other performance metrics for QEM that can better reflect their practical performance and/or can be analytically calculated for a wider range of methods? Can these metrics take into account the various additional information and resources required by different QEM methods? Instead of performance metrics, should we identify a set of well-defined use cases that can then be used to benchmark QEM methods through numerics or experiments, much like the MNIST and ImageNet databases in computer vision research?

(3) *What is the optimal QEM strategy for a practical use case?* This is a natural question that follows from the first two. After we have a full view of all possible QEM methods and good performance metrics for their practical performance, we can then identify the optimal QEM strategy for some practical use cases. This is most likely to be a hybrid of different QEM methods.

(4) *What is the connection between QEM and QEC?* We attempted to connect QEM to QEC in Sec. IV.D. There most of the connection was made specifically for

symmetry-based QEM. It would be interesting to see a more systemic approach to connect general QEM methods to QEC, or even to merge them under a larger common framework of error suppression. For this to happen, we first need a good description of the framework for QEM, which is the first question that we presented here.

(5) *Can we extend QEM beyond expectation value estimation?* While expectation value estimation has been at the heart of many useful algorithms to date, there are still many algorithms that fall outside this paradigm, for instance, a repeat-until-success algorithm like Shor's algorithm and algorithms that output a single-shot measurement, such as quantum phase estimations using the quantum Fourier transform. Can we identify an equivalent of QEM for these algorithms and usefully apply concepts such as those described in this review? For example, it should be possible to improve single-shot algorithms via postselection in the same spirit of quantum error detection.

B. Technical questions

There are also many other more technical questions about the implementations and applications of different QEM techniques, many of which we have mentioned in this review when talking about individual techniques. We list some sample questions here.

(1) Are there general protocols to scale the noise without changing the error model? If yes, can we deduce the shape of the extrapolation curve from the error model? See Sec. III.A.

(2) How can we efficiently handle drifts in the error model when applying probabilistic error cancellation in practice? See Sec. III.B.

(3) What is the training cost for learning-based QEM? See Sec. III.H.

(4) Various quantum computing hardware architectures have been realized, at least at the prototype level, with differing connectivities and native gate sets. What is the interplay between the hardware feature set and the suitable choices of QEM strategy?

VII. CONCLUSION

In this review, we have provided a comprehensive survey of QEM which has ranged from basic concepts and motivations to the implementation details of specific techniques. Owing to the low hardware requirements of QEM compared to QEC, as well as the broad range of mitigation techniques available to target diverse application scenarios, QEM has already become an integral part of many recent experimental demonstrations of quantum hardware. Even though we cannot rely on QEM alone to suppress all errors in all cases, especially when the circuit fault rate is large, we can always aim for a sweet spot between the amount of noise removed and the resource required. Consequently we can expect QEM techniques to continue to establish themselves as indispensable enablers, which is vital to maximizing the reach of each generation of hardware. We anticipate that this will remain true even into the fault-tolerant era since recent work has shown that it is possible to reduce the hardware requirements of fault tolerance by applying QEM alongside QEC. Indeed, there are applications of QEM concepts that are entirely beyond handling physical device imperfections and instead mitigate compilation (algorithmic) imperfections, i.e., by honing the performance of algorithms that of necessity produce only approximate answers.

Despite already playing a significant role in practical applications, the current landscape of QEM is still dynamic and complex, with many unexplored territories. While we have made every effort to present the rather tangled threads of this topic in a clear way in this review, it is evident that a more systematic and unified structure for QEM is desirable as the field matures. There are still many open problems left unanswered, as summarized in Sec. VI. Solving these problems may be the key to a clearer and more structured view of QEM and its role in the grand scheme of error suppression. We hope that the community of theoretical and experimental researchers who will drive the field forward may find that the survey of ideas and methods presented in this review provides useful guidance along the way.

LIST OF SYMBOLS AND ABBREVIATIONS

C_{em}	error-mitigation sampling overhead
N	number of qubits
N_{cir}	number of circuit runs
O	observable of interest
$\hat{O}_{\mathrm{em}}$	error-mitigated estimator
$\hat{O}_\rho$	random variable from measuring O on state ρ
p	physical gate error rate
P_0	circuit fault-free probability
λ	circuit fault rate: the average number of faults per circuit run
ρ	unmitigated noisy state
ρ_0	ideal noiseless state

ACKNOWLEDGMENTS

We thank Jay Gambetta, Abhinav Kandala, and Kristan Temme for their valuable insights and useful discussions. Z. C. and S. E. are grateful to Bálint Koczor, Ryuji Takagi, and Nobuyuki Yoshioka for the helpful discussions. S. E. is also grateful for the useful discussions with Yasunari Suzuki, Kento Tsubouchi, and Raam Uzdin. R. B. thanks Nicholas Rubin for the helpful discussions about pure-state N representability. Z. C. is supported by the Junior Research Fellowship from St John's College, Oxford. S. C. B. acknowledges financial support from EPSRC Hub grants under Grant Agreement No. EP/T001062/1, and from the IARPA funded LogiQ project. Z. C. and S. C. B. also acknowledge support from EPSRC's Robust and Reliable Quantum Computing (RoaRQ) project (Grant Agreement No. EP/W032635/1). S. E. is supported by JST Moonshot R&D Grant No. JPMJMS2061, MEXT Q-LEAP Grant No. JPMXS0120319794, and JST PRESTO Grant No. JPMJPR2114. Y. L. acknowledges the

support of the National Natural Science Foundation of China (Grants No. 11875050 and No. 12088101) and NSAF (Grant No. U1930403).

APPENDIX: PRACTICAL TECHNIQUES IN IMPLEMENTATIONS

1. Monte Carlo sampling

In many QEM techniques, the error-mitigated expectation value $E[\hat{O}_{\mathrm{em}}]$ is often a linear sum of the expectation values of a set of K random variables $\{\hat{O}_n\}$ that are the outputs of the set of response measurement circuits for the QEM technique,

$$E[\hat{O}_{\mathrm{em}}] = \sum_{n=1}^{K} \alpha_n E[\hat{O}_n], \tag{A1}$$

where $\{\alpha_n\}$ are real coefficients. A naive way to estimate $E[\hat{O}_{\mathrm{em}}]$ would be to perform an estimation of the individual terms $E[\hat{O}_n]$ up to a certain precision, then combine the results. In such a way, the variance of the error-mitigated estimator $\hat{O}_{\mathrm{em}}$ is given as

$$\mathrm{var}[\hat{O}_{\mathrm{em}}] = \sum_{n=1}^{K} |\alpha_n|^2 \mathrm{var}[\hat{O}_n].$$

The component random variables $\hat{O}_n$, which are generated from circuits that are variants of the primary circuit, can be expected to have a variance similar to the unmitigated estimator $\hat{O}_\rho$ generated from the noisy primary circuit: $\mathrm{var}[\hat{O}_n] \sim \mathrm{var}[\hat{O}_\rho]$. Hence, we have

$$\mathrm{var}[\hat{O}_{\mathrm{em}}] = \left(\sum_{n=1}^{K} |\alpha_n|^2\right) \mathrm{var}[\hat{O}_\rho].$$

Therefore, each component observable $\hat{O}_n$ is associated with a sampling overhead of $\sum_{n=1}^{K} |\alpha_n|^2$. Since there are K of them, the total sampling overhead is

$$C_{\mathrm{em}}^{\mathrm{naive}} = K \sum_{n=1}^{K} |\alpha_n|^2. \tag{A2}$$

This method is not scalable if the number of terms K is large, which is the case for many QEM methods. Instead, we can construct the estimator $\hat{O}_{\mathrm{em}}$ using Monte Carlo methods. We can rewrite Eq. (A1) as

$$E[\hat{O}_{\mathrm{em}}] = A \sum_{n=1}^{K} \frac{|\alpha_n|}{A} \mathrm{sgn}(\alpha_n) E[\hat{O}_n] = A E[\hat{O}_{\mathrm{mix}}], \tag{A3}$$

where $A = \sum_{n=1}^{K} |\alpha_n|$. In Eq. (A3) we have defined a new random variable $\hat{O}_{\mathrm{mix}}$ that is a probabilistic mixture of the set of random variables $\{\mathrm{sgn}(\alpha_n)\hat{O}_n\}$, with $\mathrm{sgn}(\alpha_n)\hat{O}_n$ chosen with the probability $|\alpha_n|/A$. Each sample of $\hat{O}_{\mathrm{em}}$ is simply a sample of $\hat{O}_{\mathrm{mix}}$ scaled by the factor A,

$$\hat{O}_{\mathrm{em}} = A \hat{O}_{\mathrm{mix}}.$$

Using the same $\mathrm{var}[\hat{O}_n] \sim \mathrm{var}[\hat{O}_\rho]$ assumption, we then have $\mathrm{var}[\hat{O}_{\mathrm{mix}}] \sim \mathrm{var}[\hat{O}_\rho]$, and thus

$$\mathrm{var}[\hat{O}_{\mathrm{em}}] = A^2 \mathrm{var}[\hat{O}_{\mathrm{mix}}] \sim A^2 \mathrm{var}[\hat{O}_\rho].$$

Hence, the sampling overhead [Eq. (3)] of an error-mitigated estimation performed using $\hat{O}_{\mathrm{em}}$ instead of $\hat{O}_\rho$ is then

$$C_{\mathrm{em}}^{\mathrm{MC}} = A^2 = \left(\sum_{n=1}^{K} |\alpha_n|\right)^2. \tag{A4}$$

Using the Cauchy-Schwarz inequality to compare the sampling costs in Eqs. (A2) and (A4), we have

$$\left(\sum_{n=1}^{K} |\alpha_n|\right)^2 \le K \sum_{n=1}^{K} |\alpha_n|^2$$
$$\Rightarrow C_{\mathrm{em}}^{\mathrm{MC}} \le C_{\mathrm{em}}^{\mathrm{naive}}.$$

In other words, the Monte Carlo method is always more sample efficient under our assumptions.

Instead of making assumptions about the variance of the component random variables $\hat{O}_n$, we can obtain a similar sampling overhead using Hoeffding's inequality, as shown in Eq. (5). This uses the fact that the component random variables $\hat{O}_n$ usually have the same range as $\hat{O}_\rho$ since they are typically obtained from the measurement of the same observable. Even if the observables are different, they are often all Pauli observables that are in the same range.

Here we have discussed using Monte Carlo sampling for estimating the error-mitigated expectation value. Similar arguments can be applied to the estimation of loss functions in learning-based methods and other comparable situations.

a. Exponential sampling overhead

As mentioned in Sec. II.D, if the component random variables $\{\hat{O}_n\}$ are obtained by measuring Pauli observables on circuits suffering from Pauli gate noise with a circuit fault rate of λ or more, then its expectation value will decay exponentially with λ as

$$E[\hat{O}_n] = \mathcal{O}(e^{-\beta_n \lambda})$$

for a positive β_n. We also have

$$\mathrm{var}[\hat{O}_n] = E[\hat{O}_n^2] - E[\hat{O}_n]^2 = 1 - \mathcal{O}(e^{-2\beta_n \lambda}),$$

where we have used $\hat{O}_n^2 = 1$ as the Pauli observable.

As mentioned, $\hat{O}_{\mathrm{mix}}$ is the probabilistic mixture of the set of random variable $\{\mathrm{sgn}(\alpha_n)\hat{O}_n\}$ with the probability distribution $\{p_n = |\alpha_n|/A\}$. Using the properties of random variables from mixture distribution, we have

$$E[\hat{O}_{\mathrm{mix}}] = \sum_n p_n \mathrm{sgn}(\alpha_n) E[\hat{O}_n] = \mathcal{O}(e^{-\beta\lambda}),$$

with $\beta = \min_n \beta_n$, and we also have

$$\begin{aligned}\mathrm{var}[\hat{O}_{\mathrm{mix}}] &= E[\hat{O}_{\mathrm{mix}}^2] - E[\hat{O}_{\mathrm{mix}}]^2 \\ &= \sum_n p_n E[\hat{O}_n^2] - E[\hat{O}_{\mathrm{mix}}]^2 \\ &= 1 - \mathcal{O}(e^{-2\beta\lambda}),\end{aligned}$$

where we have again used $\hat{O}_n^2 = 1$. This shows that when we are considering a Pauli observable under Pauli circuit noise, the assumption $\mathrm{var}[\hat{O}_{\mathrm{mix}}] \sim \mathrm{var}[\hat{O}_n]$ mentioned in Appendix A.1 is valid at large λ even without the need to consider exactly how the various $\hat{O}_n$ are constructed.

Since the error-mitigated estimator is $\hat{O}_{\mathrm{em}} = A\hat{O}_{\mathrm{mix}}$, we then have

$$E[\hat{O}_{\mathrm{em}}] = AE[\hat{O}_{\mathrm{mix}}] = \mathcal{O}(Ae^{-\beta\lambda}).$$

To ensure that the expectation value of the error-mitigated estimator does not decay with the noise level λ, we need to have $A = \mathcal{O}(e^{\beta\lambda})$, which implies that the sampling cost using Eq. (A4) is

$$C_{\mathrm{em}} = A^2 = \mathcal{O}(e^{2\beta\lambda}).$$

That is, it increases exponentially with λ.

2. Pauli twirling

Given a noise process $\mathcal{N}$, twirling it over a symmetry group $\mathbb{G}$ means conjugating $\mathcal{N}$ with random elements in $\mathbb{G}$,

$$\boldsymbol{T}_{\mathbb{G}}(\mathcal{N}) = \frac{1}{|\mathbb{G}|}\sum_{\mathcal{G}\in\mathbb{G}} \mathcal{G}\mathcal{N}\mathcal{G}.$$

Note that here we are using the superoperator formalism. Twirling over the Pauli group, which is called Pauli twirling, will remove all of the off-diagonal elements of $\mathcal{N}$ in the Pauli basis, thus transforming the error channel into Pauli noise.

Now suppose that we have a noisy Clifford gate $\mathcal{C}_\epsilon$ that is simply the ideal Clifford channel $\mathcal{C}$ followed by the noise channel $\mathcal{N}$: $\mathcal{C}_\epsilon = \mathcal{N}\mathcal{C}$. If we want to twirl the noise channel $\mathcal{N}$ of the noisy Clifford gate $\mathcal{C}_\epsilon$, we can apply $\mathcal{C}_\epsilon$ with a random Pauli $\mathcal{G}$ and its corresponding Pauli $\mathcal{C}\mathcal{G}\mathcal{C}^\dagger$ in each circuit run,

$$\begin{aligned}\boldsymbol{T}'_{\mathbb{G}}(\mathcal{C}_\epsilon) &= \frac{1}{|\mathbb{G}|}\sum_{\mathcal{G}\in\mathbb{G}} (\mathcal{C}\mathcal{G}\mathcal{C}^\dagger)\mathcal{C}_\epsilon\mathcal{G} = \frac{1}{\mathbb{G}}\sum_{\mathcal{G}'\in\mathbb{G}} \mathcal{G}'\mathcal{C}_\epsilon(\mathcal{C}^\dagger\mathcal{G}'\mathcal{C}) \\ &= \left(\frac{1}{|\mathbb{G}|}\sum_{\mathcal{G}\in\mathbb{G}} \mathcal{G}\mathcal{N}\mathcal{G}\right)\mathcal{C}.\end{aligned}$$

To twirl the error of a sequence of Clifford gates $\prod_{m=M}^{1}\mathcal{C}_m$, the random Pauli gates for twirling consecutive Clifford gates can be merged.

To twirl a noisy T gate $\mathcal{T}_\epsilon = \mathcal{N}\mathcal{T}$, we then need to apply the following circuit:

$$\boldsymbol{T}'_{\mathbb{G}}(\mathcal{T}_\epsilon) = \frac{1}{|\mathbb{G}|}\sum_{\mathcal{G}'\in\mathbb{G}} \mathcal{G}\mathcal{T}_\epsilon(\mathcal{T}^\dagger\mathcal{G}\mathcal{T}) = \left(\frac{1}{|\mathbb{G}|}\sum_{\mathcal{G}\in\mathbb{G}} \mathcal{G}\mathcal{N}\mathcal{G}\right)\mathcal{T},$$

where $\mathcal{T}^\dagger\mathcal{G}\mathcal{T}$ is a Clifford gate.

3. Measurement techniques

To measure an arbitrary operator, we can always measure its Pauli basis and then combine the results. Hence, without loss of generality we focus mostly on Pauli measurements in this section.

In practical experiments, it is often the case that we can perform only single-qubit Z measurements to high accuracy. Hence, one way to measure a Pauli operator is by transforming it into a single-qubit Z using Clifford circuits. Given linear qubit connectivity, the additional Clifford circuits needed will require long-range gates of depth $\mathcal{O}(\log(N))$ or local gates of depth $\mathcal{O}(N)$, a concept discussed in the context of symmetry verification by Bonet-Monroig *et al.* (2018). Alternatively, for a given Pauli operator O, if we can implement controlled-O gates, we can also indirectly measure O through the Hadamard test, as shown in Fig. 6. The controlled-O gate can be implemented using a Clifford circuit that transforms O to single-qubit Z along with a controlled-Z gate. The cost of this is similar to the previously discussed direct measurement.

Any given Pauli operator O can be written as the tensor product of single-qubit Pauli operators $O = \otimes_{n=1}^{N} G_n$, where G_n is the action of O on the nth qubit. Hence, the controlled O in the Hadamard test can also be decomposed into low-weight controlled-G_n gates instead. In fact, we can measure O directly by performing single-qubit Pauli measurements of its components $\{G_n\}$ and multiplying the results. In this way, the additional circuitry needed for measuring O is only one layer of single-qubit Clifford for changing the measurement basis.

Now suppose that we want to measure two commuting Pauli operators in the same circuit run, which can be useful in symmetry verification, we can use the aforementioned single-qubit measurements plus postprocessing scheme, but we must make sure the two Pauli operators qubitwise commute (Izmaylov, Yen, and Ryabinkin, 2019); i.e., for every qubit the single-qubit Pauli components acting on it from different observables commute. For example, XXI and IXZ are qubitwise commuting, but XX and YY are not. Note that for a set of operators to qubitwise commute, their actions on a given qubit must involve the same Pauli operator or the identity. More generally the set of operators that are linear combinations of a set of qubitwise commuting Pauli operators are also qubitwise

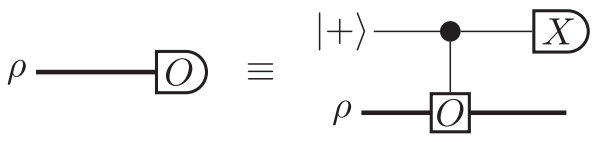

FIG. 6. Hadamard test circuit for performing Pauli O measurements. If O is a general unitary, then the Hadamard test circuit in which we measure $X \otimes I$ will output the expectation value $\mathrm{Re}(\mathrm{Tr}[O\rho])$, while measuring $Y \otimes I$ will output the expectation value $\mathrm{Im}(\mathrm{Tr}[O\rho])$.

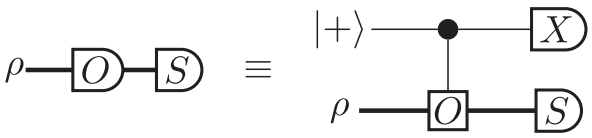

FIG. 7. Circuits for measuring $OS_+ = (OS + SO)/2$, in which O is Pauli and S is Hermitian. In the left circuit, we first perform a nondestructive measurement of O and then measure S and take the product of the measurement results. In the right circuit, the nondestructive measurement of O is carried out using Hadamard tests, and the expectation value of OS_+ is obtained by simply measuring $X \otimes S$ at the end.

commuting and can be measured simultaneously using single-qubit Pauli measurements in the same circuit run. Whenever two Pauli operators commute, we can make them qubitwise commuting using a suitable choice of Clifford circuit. Examples of this were discussed for direct symmetry verification in Sec. III.D.

In some other cases, we are more interested in the expectation values of a set of commuting operators than in their exact measurement results in any given circuit run. In these scenarios, instead of trying to measure multiple observables in every circuit run, it is possible to obtain these expectation values through *shadow tomography* (Huang, Kueng, and Preskill, 2020) using random single-qubit Pauli measurements and postprocessing.

We now move beyond measuring commuting Pauli operators and consider the case in which we want to measure the product of a Pauli operator O and a general Hermitian operator S that does not necessarily commute with O. For the circuits shown in Fig. 7, if we first perform a projective measurement of the Pauli operator O and then measure the operator S, then the latter measurement is equivalent to measuring the components of S that commute with O, which is simply $S_+ = (S + OSO)/2$ (Mitarai and Fujii, 2019; Cai, 2021d; Huo and Li, 2022). Taking the product of the two measurements, we can obtain the expectation value of the symmetrized product $OS_+ = (OS + SO)/2$, which is useful for the symmetry verification in Sec. III.D (in which S is the symmetry) and the echo verification in Sec. III.E (in which S is the dual state).

The component of S that anticommutes with O, denoted as $S_- = (S - OSO)/2$, can be obtained by first applying the Pauli rotation $\exp[-i(\pi/4)O]$ and then measuring S (Mitarai and Fujii, 2019; Cai, 2021d; Huo and Li, 2022). Combined with the previous measurement of S_+, we can obtain the expectation value of the product observable OS. Alternatively, if we can implement controlled-O and controlled-S gates, then they can be composed to give a controlled-OS gate, which can be used to measure OS using the Hadamard test.

REFERENCES

Abobeih, M. H., Y. Wang, J. Randall, S. J. H. Loenen, C. E. Bradley, M. Markham, D. J. Twitchen, B. M. Terhal, and T. H. Taminiau, 2022, "Fault-tolerant operation of a logical qubit in a diamond quantum processor," Nature (London) **606**, 884–889.

Acín, Antonio, *et al.*, 2018, "The quantum technologies roadmap: A European community view," New J. Phys. **20**, 080201.

Aharonov, D., and M. Ben-Or, 1997, "Fault-tolerant quantum computation with constant error," in *Proceedings of the 29th Annual ACM Symposium on Theory of Computing (STOC '97), El Paso, TX, 1997* (Association for Computing Machinery, New York), pp. 176–188.

Altman, Ehud, *et al.*, 2021, "Quantum simulators: Architectures and opportunities," PRX Quantum **2**, 017003.

Arute, Frank, *et al.*, 2019, "Quantum supremacy using a programmable superconducting processor," Nature (London) **574**, 505–510.

Asavanant, Warit, *et al.*, 2019, "Generation of time-domain-multiplexed two-dimensional cluster state," Science **366**, 373–376.

Baidu, 2023a, `QCompute` code, https://github.com/baidu/QCompute.

Barron, George S., and Christopher J. Wood, 2020, "Measurement error mitigation for variational quantum algorithms," arXiv:2010.08520.

Begušić, Tomislav, and Garnet Kin-Lic Chan, 2023, "Fast classical simulation of evidence for the utility of quantum computing before fault tolerance," arXiv:2306.16372.

Bennett, Charles H., Gilles Brassard, Sandu Popescu, Benjamin Schumacher, John A. Smolin, and William K. Wootters, 1996, "Purification of Noisy Entanglement and Faithful Teleportation via Noisy Channels," Phys. Rev. Lett. **76**, 722–725.

Bennett, Charles H., David P. DiVincenzo, John A. Smolin, and William K. Wootters, 1996, "Mixed-state entanglement and quantum error correction," Phys. Rev. A **54**, 3824–3851.

Bergholm, Ville, *et al.*, 2018, "`PennyLane`: Automatic differentiation of hybrid quantum-classical computations," arXiv:1811.04968.

Bergquist, J. C., Randall G. Hulet, Wayne M. Itano, and D. J. Wineland, 1986, "Observation of Quantum Jumps in a Single Atom," Phys. Rev. Lett. **57**, 1699–1702.

Biamonte, Jacob, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd, 2017, "Quantum machine learning," Nature (London) **549**, 195–202.

Blais, Alexandre, Ren-Shou Huang, Andreas Wallraff, S. M. Girvin, and R. J. Schoelkopf, 2004, "Cavity quantum electrodynamics for superconducting electrical circuits: An architecture for quantum computation," Phys. Rev. A **69**, 062320.

Bonet-Monroig, X., R. Sagastizabal, M. Singh, and T. E. O'Brien, 2018, "Low-cost error mitigation by symmetry verification," Phys. Rev. A **98**, 062339.

Bonet-Monroig, Xavier, Ryan Babbush, and Thomas E. O'Brien, 2020, "Nearly Optimal Measurement Scheduling for Partial Tomography of Quantum States," Phys. Rev. X **10**, 031064.

Bravyi, Sergey, and David Gosset, 2016, "Improved Classical Simulation of Quantum Circuits Dominated by Clifford Gates," Phys. Rev. Lett. **116**, 250501.

Bravyi, Sergey, and Alexei Kitaev, 2002, "Fermionic quantum computation," Ann. Phys. (Amsterdam) **298**, 210–226.

Bravyi, Sergey, Sarah Sheldon, Abhinav Kandala, David C. Mckay, and Jay M. Gambetta, 2021, "Mitigating measurement errors in multiqubit experiments," Phys. Rev. A **103**, 042605.

Breuckmann, Nikolas P., and Jens Niklas Eberhardt, 2021, "Quantum low-density parity-check codes," PRX Quantum **2**, 040101.

Bridgeman, Jacob C., and Christopher T. Chubb, 2017, "Hand-waving and interpretive dance: An introductory course on tensor networks," J. Phys. A **50**, 223001.

Bultrini, Daniel, Max Hunter Gordon, Piotr Czarnik, Andrew Arrasmith, M. Cerezo, Patrick J. Coles, and Lukasz Cincio, 2023, "Unifying and benchmarking state-of-the-art quantum error mitigation techniques," Quantum **7**, 1034.

Burgarth, Daniel, Paolo Facchi, Giovanni Gramegna, and Saverio Pascazio, 2019, "Generalized product formulas and quantum control," J. Phys. A **52**, 435301.

Cai, Zhenyu, 2020, "Resource Estimation for Quantum Variational Simulations of the Hubbard Model," Phys. Rev. Appl. **14**, 014059.

Cai, Zhenyu, 2021a, "Multi-exponential error extrapolation and combining error mitigation techniques for NISQ applications," npj Quantum Inf. **7**, 80.
Cai, Zhenyu, 2021b, "A practical framework for quantum error mitigation," arXiv:2110.05389.
Cai, Zhenyu, 2021c, "Quantum error mitigation using symmetry expansion," Quantum **5**, 548.
Cai, Zhenyu, 2021d, "Resource-efficient purification-based quantum error mitigation," arXiv:2107.07279.
Cai, Zhenyu, Adam Siegel, and Simon Benjamin, 2023, "Looped pipelines enabling effective 3D qubit lattices in a strictly 2D device," PRX Quantum **4**, 020345.
Calderbank, A. R., and Peter W. Shor, 1996, "Good quantum error-correcting codes exist," Phys. Rev. A **54**, 1098–1105.
Campbell, Earl, 2019, "Random Compiler for Fast Hamiltonian Simulation," Phys. Rev. Lett. **123**, 070503.
Cao, Chenfeng, Yunlong Yu, Zipeng Wu, Nic Shannon, Bei Zeng, and Robert Joynt, 2023, "Mitigating algorithmic errors in quantum optimization through energy extrapolation," Quantum Sci. Technol. **8**, 015004.
Cao, Ningping, Junan Lin, David Kribs, Yiu-Tung Poon, Bei Zeng, and Raymond Laflamme, 2022, "NISQ: Error correction, mitigation, and noise simulation," arXiv:2111.02345.
Chen, Senrui, Wenjun Yu, Pei Zeng, and Steven T. Flammia, 2021, "Robust shadow estimation," PRX Quantum **2**, 030348.
Chen, Wentao, Shuaining Zhang, Jialiang Zhang, Xiaolu Su, Yao Lu, Kuan Zhang, Mu Qiao, Ying Li, Jing-Ning Zhang, and Kihwan Kim, 2023, "Error-mitigated quantum simulation of interacting fermions with trapped ions," arXiv:2302.10436.
Chen, Yanzhu, Maziar Farahzad, Shinjae Yoo, and Tzu-Chieh Wei, 2019, "Detector tomography on IBM quantum computers and mitigation of an imperfect measurement," Phys. Rev. A **100**, 052315.
Childs, Andrew M., Aaron Ostrander, and Yuan Su, 2019, "Faster quantum simulation by randomization," Quantum **3**, 182.
Childs, Andrew M., and Nathan Wiebe, 2012, "Hamiltonian simulation using linear combinations of unitary operations," Quantum Inf. Comput. **12**, 901–924.
Chow, J. M., L. DiCarlo, J. M. Gambetta, A. Nunnenkamp, Lev S. Bishop, L. Frunzio, M. H. Devoret, S. M. Girvin, and R. J. Schoelkopf, 2010, "Detecting highly entangled states with a joint qubit readout," Phys. Rev. A **81**, 062325.
Chow, Jerry M., *et al.*, 2011, "Simple All-Microwave Entangling Gate for Fixed-Frequency Superconducting Qubits," Phys. Rev. Lett. **107**, 080502.
Cirstoiu, Cristina, Silas Dilkes, Daniel Mills, Seyon Sivarajah, and Ross Duncan, 2023, "Volumetric benchmarking of error mitigation with QERMIT," Quantum **7**, 1059.
Colless, J. I., V. V. Ramasesh, D. Dahlen, M. S. Blok, M. E. Kimchi-Schwartz, J. R. McClean, J. Carter, W. A. de Jong, and I. Siddiqi, 2018, "Computation of Molecular Spectra on a Quantum Processor with an Error-Resilient Algorithm," Phys. Rev. X **8**, 011021.
Conlon, Lorcán O., *et al.*, 2023, "Approaching optimal entangling collective measurements on quantum computing platforms," Nat. Phys. **19**, 351–357.
Córcoles, A. D., Easwar Magesan, Srikanth J. Srinivasan, Andrew W. Cross, M. Steffen, Jay M. Gambetta, and Jerry M. Chow, 2015, "Demonstration of a quantum error detection code using a square lattice of four superconducting qubits," Nat. Commun. **6**, 6979.
Czarnik, Piotr, Andrew Arrasmith, Patrick J. Coles, and Lukasz Cincio, 2021, "Error mitigation with Clifford quantum-circuit data," Quantum **5**, 592.
Czarnik, Piotr, Michael McKerns, Andrew T. Sornborger, and Lukasz Cincio, 2022, "Improving the efficiency of learning-based error mitigation," arXiv:2204.07109.
Dallaire-Demers, Pierre-Luc, Jonathan Romero, Libor Veis, Sukin Sim, and Alán Aspuru-Guzik, 2019, "Low-depth circuit ansatz for preparing correlated fermionic states on a quantum computer," Quantum Sci. Technol. **4**, 045005.
Dawson, C. M., and M. A. Nielsen, 2006, "The Solovay-Kitaev algorithm," Quantum Inf. Comput. **6**, 81–95.
Dborin, James, Vinul Wimalaweera, F. Barratt, Eric Ostby, Thomas E. O'Brien, and A. G. Green, 2022, "Simulating groundstate and dynamical quantum phase transitions on a superconducting quantum computer," Nat. Commun. **13**, 5977.
DePrince, A. Eugene, 2016, "Variational optimization of the two-electron reduced-density matrix under pure-state N-representability conditions," J. Chem. Phys. **145**, 164109.
Derby, Charles, and Joel Klassen, 2021, "A compact fermion to qubit mapping part 2: Alternative lattice geometries," arXiv:2101.10735.
Dinur, Irit, Shai Evra, Ron Livne, Alexander Lubotzky, and Shahar Mozes, 2022, "Locally testable codes with constant rate, distance, and locality," in *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC 2022), Rome, 2022* (Association for Computing Machinery, New York), pp. 357–374.
Dumitrescu, E. F., A. J. McCaskey, G. Hagen, G. R. Jansen, T. D. Morris, T. Papenbrock, R. C. Pooser, D. J. Dean, and P. Lougovski, 2018, "Cloud Quantum Computing of an Atomic Nucleus," Phys. Rev. Lett. **120**, 210501.
Eastin, Bryan, and Emanuel Knill, 2009, "Restrictions on Transversal Encoded Quantum Gate Sets," Phys. Rev. Lett. **102**, 110502.
Ebadi, Sepehr, *et al.*, 2021, "Quantum phases of matter on a 256-atom programmable quantum simulator," Nature (London) **595**, 227–232.
Egan, Laird, *et al.*, 2021, "Fault-tolerant control of an error-corrected qubit," Nature (London) **598**, 281–286.
Eisert, Jens, Dominik Hangleiter, Nathan Walk, Ingo Roth, Damian Markham, Rhea Parekh, Ulysse Chabaud, and Elham Kashefi, 2020, "Quantum certification and benchmarking," Nat. Rev. Phys. **2**, 382–390.
Endo, Suguru, Simon C. Benjamin, and Ying Li, 2018, "Practical Quantum Error Mitigation for Near-Future Applications," Phys. Rev. X **8**, 031027.
Endo, Suguru, Yasunari Suzuki, Kento Tsubouchi, Rui Asaoka, Kaoru Yamamoto, Yuichiro Matsuzaki, and Yuuki Tokunaga, 2022, "Quantum error mitigation for rotation symmetric bosonic codes with symmetry expansion," arXiv:2211.06164.
Endo, Suguru, Qi Zhao, Ying Li, Simon Benjamin, and Xiao Yuan, 2019, "Mitigating algorithmic errors in a Hamiltonian simulation," Phys. Rev. A **99**, 012334.
Facchi, P., D. A. Lidar, and S. Pascazio, 2004, "Unification of dynamical decoupling and the quantum Zeno effect," Phys. Rev. A **69**, 032314.
Farhi, Edward, Jeffrey Goldstone, and Sam Gutmann, 2014, "A quantum approximate optimization algorithm," arXiv:1411.4028.
Ferracin, Samuele, Akel Hashim, Jean-Loup Ville, Ravi Naik, Arnaud Carignan-Dugas, Hammam Qassim, Alexis Morvan, David I. Santiago, Irfan Siddiqi, and Joel J. Wallman, 2022, "Efficiently improving the performance of noisy quantum computers," arXiv:2201.10672.
Feynman, Richard P., 1982, "Simulating physics with computers," Int. J. Theor. Phys. **21**, 467–488.
Foss-Feig, Michael, *et al.*, 2022, "Entanglement from Tensor Networks on a Trapped-Ion Quantum Computer," Phys. Rev. Lett. **128**, 150504.

Fowler, Austin G., Matteo Mariantoni, John M. Martinis, and Andrew N. Cleland, 2012, "Surface codes: Towards practical large-scale quantum computation," Phys. Rev. A **86**, 032324.

Fuchs, C. A., and J. van de Graaf, 1999, "Cryptographic distinguishability measures for quantum-mechanical states," IEEE Trans. Inf. Theory **45**, 1216–1227.

Garmon, J. W. O., R. C. Pooser, and E. F. Dumitrescu, 2020, "Benchmarking noise extrapolation with the `OpenPulse` control framework," Phys. Rev. A **101**, 042308.

Geller, Michael R., 2020, "Rigorous measurement error correction," Quantum Sci. Technol. **5**, 03LT01.

Gilchrist, Alexei, Daniel R. Terno, and Christopher J. Wood, 2011, "Vectorization of quantum operations and its use," arXiv:0911.2539.

Giurgica-Tiron, Tudor, Yousef Hindy, Ryan LaRose, Andrea Mari, and William J. Zeng, 2020, "Digital zero noise extrapolation for quantum error mitigation," in *Proceedings of the IEEE International Conference on Quantum Computing and Engineering (QCE), Denver, 2020* (IEEE, New York), pp. 306–316.

Google Quantum AI, 2023, "Suppressing quantum errors by scaling a surface code logical qubit," Nature (London) **614**, 676–681.

Google Quantum AI *et al.*, 2020a, "Hartree-Fock on a superconducting qubit quantum computer," Science **369**, 1084–1089.

Google Quantum AI *et al.*, 2020b, "Observation of separated dynamics of charge and spin in the Fermi-Hubbard model," arXiv:2010.07965.

Google Quantum AI *et al.*, 2022, "Formation of robust bound states of interacting microwave photons," Nature (London) **612**, 240–245.

Gottesman, Daniel, 2009, "An introduction to quantum error correction and fault-tolerant quantum computation," arXiv:0904.2557.

Gottesman, Daniel, 2014, "Fault-tolerant quantum computation with constant overhead," Quantum Inf. Comput. **14**, 1339–1371.

Gottesman, Daniel, 2016, "Quantum fault tolerance in small experiments," arXiv:1610.03507.

Gottesman, Daniel Eric, 1997, "Stabilizer codes and quantum error correction," Ph.D. thesis (California Institute of Technology).

Gu, Yanwu, Yunheng Ma, Nicolo Forcellini, and Dong E. Liu, 2023, "Noise-Resilient Phase Estimation with Randomized Compiling," Phys. Rev. Lett. **130**, 250601.

Guo, Yuchen, and Shuo Yang, 2022, "Quantum error mitigation via matrix product operators," PRX Quantum **3**, 040313.

Guo, Yuchen, and Shuo Yang, 2023, "Noise effects on purity and quantum entanglement in terms of physical implementability," npj Quantum Inf. **9**, 1–7.

Gutiérrez, Mauricio, and Kenneth R. Brown, 2015, "Comparison of a quantum error-correction threshold for exact and approximate errors," Phys. Rev. A **91**, 022335.

Hakoshima, Hideaki, Yuichiro Matsuzaki, and Suguru Endo, 2021, "Relationship between costs for quantum error mitigation and non-Markovian measures," Phys. Rev. A **103**, 012611.

Hama, Yusuke, and Hirofumi Nishi, 2022, "Quantum error mitigation via quantum-noise-effect circuit groups," arXiv:2205.13907.

Hamilton, Kathleen E., Tyler Kharazi, Titus Morris, Alexander J. McCaskey, Ryan S. Bennink, and Raphael C. Pooser, 2020, "Scalable quantum processor noise characterization," in *Proceedings of the IEEE International Conference on Quantum Computing and Engineering (QCE), Denver, 2020* (IEEE, New York), pp. 430–440.

Haroche, Serge, and Jean-Michel Raimond, 2006, *Exploring the Quantum: Atoms, Cavities, and Photons*, Oxford Graduate Texts (Oxford University Press, Oxford).

He, Andre, Benjamin Nachman, Wibe A. de Jong, and Christian W. Bauer, 2020, "Zero-noise extrapolation for quantum-gate error mitigation with identity insertions," Phys. Rev. A **102**, 012426.

Heinsoo, Johannes, *et al.*, 2018, "Rapid High-Fidelity Multiplexed Readout of Superconducting Qubits," Phys. Rev. Appl. **10**, 034040.

Helgaker, Trygve, Poul Jorgensen, and Jeppe Olsen, 2000, *Molecular Electronic-Structure Theory* (John Wiley & Sons, New York).

Henao, Ivan, Jader P. Santos, and Raam Uzdin, 2023, "Adaptive quantum error mitigation using pulse-based inverse evolutions," npj Quantum Inf. **9**, 1–10.

Hiai, Fumio, Masanori Ohya, and Makoto Tsukada, 2008, "Sufficiency, KMS condition and relative entropy in von Neumann algebras," in *Selected Papers of M. Ohya*, edited by Noburu Watanabe (World Scientific, Singapore), pp. 420–430.

Hicks, Rebecca, Bryce Kobrin, Christian W. Bauer, and Benjamin Nachman, 2022, "Active readout-error mitigation," Phys. Rev. A **105**, 012419.

Hu, Hong-Ye, Ryan LaRose, Yi-Zhuang You, Eleanor Rieffel, and Zhihui Wang, 2022, "Logical shadow tomography: Efficient estimation of error-mitigated observables," arXiv:2203.07263.

Huang, Hsin-Yuan, Richard Kueng, and John Preskill, 2020, "Predicting many properties of a quantum system from very few measurements," Nat. Phys. **16**, 1050–1057.

Huggins, William J., Sam McArdle, Thomas E. O'Brien, Joonho Lee, Nicholas C. Rubin, Sergio Boixo, K. Birgitta Whaley, Ryan Babbush, and Jarrod R. McClean, 2021, "Virtual Distillation for Quantum Error Mitigation," Phys. Rev. X **11**, 041036.

Huggins, William J., Jarrod R. McClean, Nicholas C. Rubin, Zhang Jiang, Nathan Wiebe, K. Birgitta Whaley, and Ryan Babbush, 2021, "Efficient and noise resilient measurements for quantum chemistry on near-term quantum computers," npj Quantum Inf. **7**, 23.

Huo, Mingxia, and Ying Li, 2022, "Dual-state purification for practical quantum error mitigation," Phys. Rev. A **105**, 022427.

Izmaylov, Artur F., Tzu-Ching Yen, and Ilya G. Ryabinkin, 2019, "Revising the measurement process in the variational quantum eigensolver: Is it possible to reduce the number of separately measured operators?," Chem. Sci. **10**, 3746–3755.

Jiang, Jiaqing, Kun Wang, and Xin Wang, 2021, "Physical implementability of linear maps and its application in error mitigation," Quantum **5**, 600.

Jiang, Zhang, Jarrod McClean, Ryan Babbush, and Hartmut Neven, 2019, "Majorana Loop Stabilizer Codes for Error Mitigation in Fermionic Quantum Simulations," Phys. Rev. Appl. **12**, 064041.

Jnane, Hamza, Jonathan Steinberg, Zhenyu Cai, H. Chau Nguyen, and Bálint Koczor, 2023, "Quantum error mitigated classical shadows," arXiv:2305.04956.

Jurcevic, Petar, *et al.*, 2021, "Demonstration of quantum volume 64 on a superconducting quantum computing system," Quantum Sci. Technol. **6**, 025020.

Kandala, Abhinav, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M. Chow, and Jay M. Gambetta, 2017, "Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets," Nature (London) **549**, 242–246.

Kandala, Abhinav, Kristan Temme, Antonio D. Córcoles, Antonio Mezzacapo, Jerry M. Chow, and Jay M. Gambetta, 2019, "Error mitigation extends the computational reach of a noisy quantum processor," Nature (London) **567**, 491–495.

Kastoryano, Michael J., and Kristan Temme, 2013, "Quantum logarithmic Sobolev inequalities and rapid mixing," J. Math. Phys. (N.Y.) **54**, 052202.

Kechedzhi, K., S. V. Isakov, S. Mandrà, B. Villalonga, X. Mi, S. Boixo, and V. Smelyanskiy, 2023, "Effective quantum volume, fidelity and computational cost of noisy quantum processing experiments," arXiv:2306.15970.

Keen, Trevor, Thomas Maier, Steven Johnston, and Pavel Lougovski, 2020, "Quantum-classical simulation of two-site dynamical mean-field theory on noisy quantum hardware," Quantum Sci. Technol. **5**, 035001.

Kelly, J., *et al.*, 2015, "State preservation by repetitive error detection in a superconducting quantum circuit," Nature (London) **519**, 66–69.

Khamoshi, Armin, Francesco A. Evangelista, and Gustavo E. Scuseria, 2021, "Correlating AGP on a quantum computer," Quantum Sci. Technol. **6**, 014004.

Kim, Youngseok, Christopher J. Wood, Theodore J. Yoder, Seth T. Merkel, Jay M. Gambetta, Kristan Temme, and Abhinav Kandala, 2023, "Scalable error mitigation for noisy quantum circuits produces competitive expectation values," Nat. Phys. **19**, 752–759.

Kim, Youngseok, *et al.*, 2023, "Evidence for the utility of quantum computing before fault tolerance," Nature (London) **618**, 500–505.

Kitaev, A. Yu., 1997, "Quantum computations: Algorithms and error correction," Russ. Math. Surv. **52**, 1191.

Kivlichan, Ian D., *et al.*, 2020, "Improved fault-tolerant quantum simulation of condensed-phase correlated electrons via Trotterization," Quantum **4**, 296.

Klco, N., E. F. Dumitrescu, A. J. McCaskey, T. D. Morris, R. C. Pooser, M. Sanz, E. Solano, P. Lougovski, and M. J. Savage, 2018, "Quantum-classical computation of Schwinger model dynamics using quantum computers," Phys. Rev. A **98**, 032331.

Kliesch, Martin, and Ingo Roth, 2021, "Theory of quantum system certification," PRX Quantum **2**, 010201.

Kliuchnikov, Vadym, Dmitri Maslov, and Michele Mosca, 2013, "Asymptotically Optimal Approximation of Single Qubit Unitaries by Clifford and *T* Circuits Using a Constant Number of Ancillary Qubits," Phys. Rev. Lett. **110**, 190502.

Klyachko, Alexander A., 2006, "Quantum marginal problem and *N*-representability," J. Phys. Conf. Ser. **36**, 72–86.

Knill, E., 2004, "Fault-tolerant postselected quantum computation: Threshold analysis," arXiv:quant-ph/0404104.

Koczor, Bálint, 2021a, "The dominant eigenvector of a noisy quantum state," New J. Phys. **23**, 123047.

Koczor, Bálint, 2021b, "Exponential Error Suppression for Near-Term Quantum Devices," Phys. Rev. X **11**, 031057.

Krebsbach, Michael, Björn Trauzettel, and Alessio Calzona, 2022, "Optimization of Richardson extrapolation for quantum error mitigation," Phys. Rev. A **106**, 062436.

Krinner, Sebastian, *et al.*, 2022, "Realizing repeated quantum error correction in a distance-three surface code," Nature (London) **605**, 669–674.

Kueng, Richard, David M. Long, Andrew C. Doherty, and Steven T. Flammia, 2016, "Comparing Experiments to the Fault-Tolerance Threshold," Phys. Rev. Lett. **117**, 170502.

Kwon, Hyeokjea, and Joonwoo Bae, 2021, "A hybrid quantum-classical approach to mitigating measurement errors in quantum algorithms," IEEE Trans. Comput. **70**, 1401–1411.

LaRose, Ryan, *et al.*, 2022, "MITIQ: A software package for error mitigation on noisy quantum computers," Quantum **6**, 774.

Le Cam, Lucien, 1960, "An approximation theorem for the Poisson binomial distribution," Pac. J. Math. **10**, 1181–1197.

Lee, Joonho, Dominic W. Berry, Craig Gidney, William J. Huggins, Jarrod R. McClean, Nathan Wiebe, and Ryan Babbush, 2021, "Even more efficient quantum computations of chemistry through tensor hypercontraction," PRX Quantum **2**, 030305.

Li, Ying, and Simon C. Benjamin, 2017, "Efficient Variational Quantum Simulator Incorporating Active Error Minimization," Phys. Rev. X **7**, 021050.

Lidar, Daniel A., 2014, "Review of decoherence-free subspaces, noiseless subsystems, and dynamical decoupling," in *Quantum Information and Computation for Chemistry*, Advances in Chemical Physics Vol. 327, edited by Sabre Kais (John Wiley & Sons, New York), pp. 295–354.

Lidar, Daniel A., and Todd A. Brun, 2013, "Introduction to decoherence and noise in open quantum systems," in *Quantum Error Correction*, edited by Daniel A. Lidar, and Todd A. Brun (Cambridge University Press, Cambridge), pp. 3–45.

Lin, Junan, Joel J. Wallman, Ian Hincks, and Raymond Laflamme, 2021, "Independent state and measurement characterization for quantum computers," Phys. Rev. Res. **3**, 033285.

Linke, Norbert M., Mauricio Gutierrez, Kevin A. Landsman, Caroline Figgatt, Shantanu Debnath, Kenneth R. Brown, and Christopher Monroe, 2017, "Fault-tolerant quantum error detection," Sci. Adv. **3**, e1701074.

Litinski, Daniel, 2019, "Magic state distillation: Not as costly as you think," Quantum **3**, 205.

Liu, Yi-Kai, Matthias Christandl, and F. Verstraete, 2007, "Quantum Computational Complexity of the *N*-Representability Problem: QMA Complete," Phys. Rev. Lett. **98**, 110503.

Lloyd, Seth, Masoud Mohseni, and Patrick Rebentrost, 2014, "Quantum principal component analysis," Nat. Phys. **10**, 631–633.

Lostaglio, M., and A. Ciani, 2021, "Error Mitigation and Quantum-Assisted Simulation in the Error Corrected Regime," Phys. Rev. Lett. **127**, 200506.

Lowe, Angus, Max Hunter Gordon, Piotr Czarnik, Andrew Arrasmith, Patrick J. Coles, and Lukasz Cincio, 2021, "Unified approach to data-driven quantum error mitigation," Phys. Rev. Res. **3**, 033098.

Lu, Sirui, Mari Carmen Bañuls, and J. Ignacio Cirac, 2021, "Algorithms for quantum simulation at finite energies," PRX Quantum **2**, 020321.

Maciejewski, Filip B., Tomasz Rybotycki, Oskar Slowik, and Jan Tuziemski, 2020, "Quantum readout errors mitigation (QREM)—Open source GitHub repository," https://github.com/fbm2718/QREM.

Maciejewski, Filip B., Zoltán Zimborás, and Michał Oszmaniec, 2020, "Mitigation of readout noise in near-term quantum devices by classical post-processing based on detector tomography," Quantum **4**, 257.

Madjarov, Ivaylo S., Jacob P. Covey, Adam L. Shaw, Joonhee Choi, Anant Kale, Alexandre Cooper, Hannes Pichler, Vladimir Schkolnik, Jason R. Williams, and Manuel Endres, 2020, "High-fidelity entanglement and detection of alkaline-earth Rydberg atoms," Nat. Phys. **16**, 857–861.

Mari, Andrea, Nathan Shammah, and William J. Zeng, 2021, "Extending quantum probabilistic error cancellation by noise scaling," Phys. Rev. A **104**, 052607.

Mazziotti, David A, 2016, "Pure-*N*-representability conditions of two-fermion reduced density matrices," Phys. Rev. A **94**, 032516.

McArdle, Sam, Xiao Yuan, and Simon Benjamin, 2019, "Error-Mitigated Digital Quantum Simulation," Phys. Rev. Lett. **122**, 180501.

McCaskey, Alexander J., Zachary P. Parks, Jacek Jakowski, Shirley V. Moore, Titus D. Morris, Travis S. Humble, and Raphael C. Pooser, 2019, "Quantum chemistry as a benchmark for near-term quantum computers," npj Quantum Inf. **5**, 99.

McClean, Jarrod R., Zhang Jiang, Nicholas C. Rubin, Ryan Babbush, and Hartmut Neven, 2020, "Decoding quantum errors with subspace expansions," Nat. Commun. **11**, 636.

McClean, Jarrod R., Mollie E. Kimchi-Schwartz, Jonathan Carter, and Wibe A. de Jong, 2017, "Hybrid quantum-classical hierarchy

for mitigation of decoherence and determination of excited states," Phys. Rev. A **95**, 042308.
McClean, Jarrod R., Jonathan Romero, Ryan Babbush, and Alán Aspuru-Guzik, 2016, "The theory of variational hybrid quantum-classical algorithms," New J. Phys. **18**, 023023.
McWeeny, R., 1960, "Some recent advances in density matrix theory," Rev. Mod. Phys. **32**, 335–369.
Merkel, Seth T., Jay M. Gambetta, John A. Smolin, Stefano Poletto, Antonio D. Córcoles, Blake R. Johnson, Colm A. Ryan, and Matthias Steffen, 2013, "Self-consistent quantum process tomography," Phys. Rev. A **87**, 062119.
Mi, Xiao, *et al.*, 2021, "Information scrambling in quantum circuits," Science **374**, 1479–1483.
Mitarai, Kosuke, and Keisuke Fujii, 2019, "Methodology for replacing indirect measurements with direct measurements," Phys. Rev. Res. **1**, 013006.
Montanaro, Ashley, and Stasja Stanisic, 2021, "Error mitigation by training with fermionic linear optics," arXiv:2102.02120.
Motta, Mario, Chong Sun, Adrian T. K. Tan, Matthew J. O'Rourke, Erika Ye, Austin J. Minnich, Fernando G. S. L. Brandão, and Garnet Kin-Lic Chan, 2020, "Determining eigenstates and thermal states on a quantum computer using quantum imaginary time evolution," Nat. Phys. **16**, 205–210.
Müller-Hermes, Alexander, Daniel Stilck França, and Michael M. Wolf, 2016, "Entropy production of doubly stochastic quantum channels," J. Math. Phys. (N.Y.) **57**, 022203.
Nachman, Benjamin, Miroslav Urbanek, Wibe A. de Jong, and Christian W. Bauer, 2020, "Unfolding quantum computer readout noise," npj Quantum Inf. **6**, 84.
Nagourney, Warren, Jon Sandberg, and Hans Dehmelt, 1986, "Shelved Optical Electron Amplifier: Observation of Quantum Jumps," Phys. Rev. Lett. **56**, 2797–2799.
Nation, Paul D., Hwajung Kang, Neereja Sundaresan, and Jay M. Gambetta, 2021, "Scalable mitigation of measurement errors on quantum computers," PRX Quantum **2**, 040326.
Neill, C., *et al.*, 2021, "Accurately computing the electronic properties of a quantum ring," Nature (London) **594**, 508–512.
Nielsen, Michael A., and Isaac L. Chuang, 2010, *Quantum Computation and Quantum Information: 10th Anniversary Edition* (Cambridge University Press, Cambridge, England).
Nigg, D., M. Müller, E. A. Martinez, P. Schindler, M. Hennrich, T. Monz, M. A. Martin-Delgado, and R. Blatt, 2014, "Quantum computations on a topologically encoded qubit," Science **345**, 302–305.
O'Brien, T. E., *et al.*, 2023, "Purification-based quantum error mitigation of pair-correlated electron simulations," Nat. Phys. **19**, 1787–1792.
O'Brien, Thomas E., Stefano Polla, Nicholas C. Rubin, William J. Huggins, Sam McArdle, Sergio Boixo, Jarrod R. McClean, and Ryan Babbush, 2021, "Error mitigation via verified phase estimation," PRX Quantum **2**, 020317.
O'Gorman, Joe, and Earl T. Campbell, 2017, "Quantum computation with realistic magic-state factories," Phys. Rev. A **95**, 032338.
Otten, Matthew, and Stephen K. Gray, 2019, "Accounting for errors in quantum algorithms via individual error reduction," npj Quantum Inf. **5**, 11.
Ouyang, Yingkai, David R. White, and Earl T. Campbell, 2020, "Compilation by stochastic Hamiltonian sparsification," Quantum **4**, 235.
Palmieri, Adriano Macarone, Egor Kovlakov, Federico Bianchi, Dmitry Yudin, Stanislav Straupe, Jacob D. Biamonte, and Sergei Kulik, 2020, "Experimental neural network enhanced quantum tomography," npj Quantum Inf. **6**, 20.
Panteleev, Pavel, and Gleb Kalachev, 2022, "Asymptotically good Quantum and locally testable classical LDPC codes," in *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC 2022), Rome, 2022* (Association for Computing Machinery, New York), pp. 375–388.
Peruzzo, Alberto, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, Alán Aspuru-Guzik, and Jeremy L. O'Brien, 2014, "A variational eigenvalue solver on a photonic quantum processor," Nat. Commun. **5**, 4213.
Piveteau, Christophe, David Sutter, Sergey Bravyi, Jay M. Gambetta, and Kristan Temme, 2021, "Error Mitigation for Universal Gates on Encoded Qubits," Phys. Rev. Lett. **127**, 200505.
Piveteau, Christophe, David Sutter, and Stefan Woerner, 2022, "Quasiprobability decompositions with reduced sampling overhead," npj Quantum Inf. **8**, 12.
Polla, Stefano, Gian-Luca R. Anselmetti, and Thomas E. O'Brien, 2023, "Optimizing the information extracted by a single qubit measurement," Phys. Rev. A **108**, 012403.
Postler, Lukas, *et al.*, 2022, "Demonstration of fault-tolerant universal quantum gate operations," Nature (London) **605**, 675–680.
Qin, Dayue, Yanzhu Chen, and Ying Li, 2023, "Error statistics and scalability of quantum error mitigation formulas," npj Quantum Inf. **9**, 1–14.
Qiskit Contributors, 2023c, QISKIT code, https://zenodo.org/records/8190968.
Quek, Yihui, Daniel Stilck França, Sumeet Khatri, Johannes Jakob Meyer, and Jens Eisert, 2022, "Exponentially tighter bounds on limitations of quantum error mitigation," arXiv:2210.11505.
Regula, Bartosz, Ryuji Takagi, and Mile Gu, 2021, "Operational applications of the diamond norm and related measures in quantifying the non-physicality of quantum maps," Quantum **5**, 522.
Rosenberg, Eliott, Paul Ginsparg, and Peter L. McMahon, 2022, "Experimental error mitigation using linear rescaling for variational quantum eigensolving with up to 20 qubits," Quantum Sci. Technol. **7**, 015024.
Ross, Neil J., and Peter Selinger, 2016, "Optimal ancilla-free Clifford + T approximation of z-rotations," Quantum Inf. Comput. **16**, 901–953.
Rubin, Nicholas C., Ryan Babbush, and Jarrod McClean, 2018, "Application of fermionic marginal constraints to hybrid quantum algorithms," New J. Phys. **20**, 053020.
Russo, A. E., K. M. Rudinger, B. C. A. Morrison, and A. D. Baczewski, 2021, "Evaluating Energy Differences on a Quantum Computer with Robust Phase Estimation," Phys. Rev. Lett. **126**, 210501.
Ryan-Anderson, C., *et al.*, 2022, "Implementing fault-tolerant entangling gates on the five-qubit code and the color code," arXiv:2208.01863.
Sagastizabal, R., *et al.*, 2019, "Experimental error mitigation via symmetry verification in a variational quantum eigensolver," Phys. Rev. A **100**, 010302.
Sanders, Yuval R., Joel J. Wallman, and Barry C. Sanders, 2015, "Bounding quantum gate error rate based on reported average fidelity," New J. Phys. **18**, 012002.
Sauter, Th., W. Neuhauser, R. Blatt, and P. E. Toschek, 1986, "Observation of Quantum Jumps," Phys. Rev. Lett. **57**, 1696–1698.
Seif, Alireza, Ze-Pei Cian, Sisi Zhou, Senrui Chen, and Liang Jiang, 2023, "Shadow distillation: Quantum error mitigation with classical shadows for near-term quantum processors," PRX Quantum **4**, 010303.
Setia, Kanav, Sergey Bravyi, Antonio Mezzacapo, and James D. Whitfield, 2019, "Superfast encodings for fermionic quantum simulation," Phys. Rev. Res. **1**, 033033.

Setia, Kanav, and James D. Whitfield, 2018, "Bravyi-Kitaev Superfast simulation of fermions on a quantum computer," J. Chem. Phys. **148**, 164104.

Shor, P., 1999, "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer," SIAM Rev. **41**, 303–332.

Shor, P. W., 1996, "Fault-tolerant quantum computation," in *Proceedings of the 37th Conference on Foundations of Computer Science, Burlington, VT, 1996* (IEEE, New York), pp. 56–65.

Shor, Peter W., 1995, "Scheme for reducing decoherence in quantum computer memory," Phys. Rev. A **52**, R2493–R2496.

Sidi, Avram, 2003, *Practical Extrapolation Methods: Theory and Applications*, Cambridge Monographs on Applied and Computational Mathematics (Cambridge University Press, Cambridge, England).

Smart, Scott E., and David A. Mazziotti, 2019, "Quantum-classical hybrid algorithm using an error-mitigating *N*-representability condition to compute the Mott metal-insulator transition," Phys. Rev. A **100**, 022517.

Smart, Scott E., and David A. Mazziotti, 2020, "Efficient two-electron ansatz for benchmarking quantum chemistry on a quantum computer," Phys. Rev. Res. **2**, 023048.

Song, Chao, Jing Cui, H. Wang, J. Hao, H. Feng, and Ying Li, 2019, "Quantum computation with universal error mitigation on a superconducting quantum processor," Sci. Adv. **5**, eaaw5686.

Stanisic, Stasja, Jan Lukas Bosse, Filippo Maria Gambetta, Raul A. Santos, Wojciech Mruczkiewicz, Thomas E. O'Brien, Eric Ostby, and Ashley Montanaro, 2022, "Observing ground-state properties of the Fermi-Hubbard model using a scalable algorithm on a quantum computer," Nat. Commun. **13**, 5743.

Steane, A. M., 1996, "Error Correcting Codes in Quantum Theory," Phys. Rev. Lett. **77**, 793–797.

Steudtner, Mark, and Stephanie Wehner, 2018, "Fermion-to-qubit mappings with varying resource requirements for quantum simulation," New J. Phys. **20**, 063010.

Steudtner, Mark, and Stephanie Wehner, 2019, "Quantum codes for quantum simulation of fermions on a square lattice of qubits," Phys. Rev. A **99**, 022308.

Strikis, Armands, Dayue Qin, Yanzhu Chen, Simon C. Benjamin, and Ying Li, 2021, "Learning-based quantum error mitigation," PRX Quantum **2**, 040330.

Suchsland, Philippe, Francesco Tacchino, Mark H. Fischer, Titus Neupert, Panagiotis Kl. Barkoutsos, and Ivano Tavernelli, 2021, "Algorithmic error mitigation scheme for current quantum processors," Quantum **5**, 492.

Sun, Jinzhao, Xiao Yuan, Takahiro Tsunoda, Vlatko Vedral, Simon C. Benjamin, and Suguru Endo, 2021, "Mitigating Realistic Noise in Practical Noisy Intermediate-Scale Quantum Devices," Phys. Rev. Appl. **15**, 034026.

Suter, Dieter, and Gonzalo A. Álvarez, 2016, "Colloquium: Protecting quantum information against environmental noise," Rev. Mod. Phys. **88**, 041001.

Suzuki, Masuo, 1990, "Fractal decomposition of exponential operators with applications to many-body theories and Monte Carlo simulations," Phys. Lett. A **146**, 319–323.

Suzuki, Masuo, 1991, "General theory of fractal path integrals with applications to many-body theories and statistical physics," J. Math. Phys. (N.Y.) **32**, 400–407.

Suzuki, Yasunari, Suguru Endo, Keisuke Fujii, and Yuuki Tokunaga, 2022, "Quantum error mitigation as a universal error reduction technique: Applications from the NISQ to the fault-tolerant quantum computing eras," PRX Quantum **3**, 010345.

Tacchino, Francesco, Panagiotis Barkoutsos, Chiara Macchiavello, Ivano Tavernelli, Dario Gerace, and Daniele Bajoni, 2020, "Quantum implementation of an artificial feed-forward neural network," Quantum Sci. Technol. **5**, 044010.

Takagi, Ryuji, 2021, "Optimal resource cost for error mitigation," Phys. Rev. Res. **3**, 033178.

Takagi, Ryuji, Suguru Endo, Shintaro Minagawa, and Mile Gu, 2022, "Fundamental limits of quantum error mitigation," npj Quantum Inf. **8**, 114.

Takagi, Ryuji, Hiroyasu Tajima, and Mile Gu, 2023, "Universal Sampling Lower Bounds for Quantum Error Mitigation," Phys. Rev. Lett. **131**, 210602.

Takeda, Kenta, Akito Noiri, Takashi Nakajima, Takashi Kobayashi, and Seigo Tarucha, 2022, "Quantum error correction with silicon spin qubits," Nature (London) **608**, 682–686.

Takeshita, Tyler, Nicholas C. Rubin, Zhang Jiang, Eunseok Lee, Ryan Babbush, and Jarrod R. McClean, 2020, "Increasing the Representation Accuracy of Quantum Simulations of Chemistry without Extra Quantum Resources," Phys. Rev. X **10**, 011004.

Tannu, Swamit S., and Moinuddin K. Qureshi, 2019, "Mitigating measurement errors in quantum computers by exploiting state-dependent bias," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO '52), Columbus, OH, 2019* (Association for Computing Machinery, New York), pp. 279–290.

Temme, Kristan, Sergey Bravyi, and Jay M. Gambetta, 2017, "Error Mitigation for Short-Depth Quantum Circuits," Phys. Rev. Lett. **119**, 180509.

Terhal, Barbara M., 2015, "Quantum error correction for quantum memories," Rev. Mod. Phys. **87**, 307–346.

Tindall, Joseph, Matt Fishman, Miles Stoudenmire, and Dries Sels, 2023, "Efficient tensor network simulation of IBM's kicked Ising experiment," arXiv:2306.14887.

Tran, Minh C., Yuan Su, Daniel Carney, and Jacob M. Taylor, 2021, "Faster digital quantum simulation by symmetry protection," PRX Quantum **2**, 010323.

Tsubouchi, Kento, Takahiro Sagawa, and Nobuyuki Yoshioka, 2023, "Universal Cost Bound of Quantum Error Mitigation Based on Quantum Estimation Theory," Phys. Rev. Lett. **131**, 210601.

Tsubouchi, Kento, Yasunari Suzuki, Yuuki Tokunaga, Nobuyuki Yoshioka, and Suguru Endo, 2023, "Virtual quantum error detection," Phys. Rev. A **108**, 042426.

Tsuchimochi, Takashi, Yuto Mori, and Seiichiro L. Ten-no, 2020, "Spin-projection for quantum computation: A low-depth approach to strong correlation," Phys. Rev. Res. **2**, 043142.

Unruh, W. G., 1995, "Maintaining coherence in quantum computers," Phys. Rev. A **51**, 992–997.

Urbanek, Miroslav, Daan Camps, Roel Van Beeumen, and Wibe A. de Jong, 2020, "Chemistry on quantum computers with virtual quantum subspace expansion," J. Chem. Theory Comput. **16**, 5425–5431.

Urbanek, Miroslav, Benjamin Nachman, Vincent R. Pascuzzi, Andre He, Christian W. Bauer, and Wibe A. de Jong, 2021, "Mitigating Depolarizing Noise on Quantum Computers with Noise-Estimation Circuits," Phys. Rev. Lett. **127**, 270502.

van den Berg, Ewout, Zlatko K. Minev, Abhinav Kandala, and Kristan Temme, 2023, "Probabilistic error cancellation with sparse Pauli-Lindblad models on noisy quantum processors," Nat. Phys. **19**, 1116–1121.

van den Berg, Ewout, Zlatko K. Minev, and Kristan Temme, 2022, "Model-free readout-error mitigation for quantum expectation values," Phys. Rev. A **105**, 032620.

Vazquez, Almudena Carrera, Ralf Hiptmair, and Stefan Woerner, 2022, "Enhancing the quantum linear systems algorithm using Richardson extrapolation," ACM Trans. Quantum Comput. **3**, 2:1–2:37.

Vovrosh, Joseph, Kiran E. Khosla, Sean Greenaway, Christopher Self, M. S. Kim, and Johannes Knolle, 2021, "Simple mitigation of global depolarizing errors in quantum simulations," Phys. Rev. E **104**, 035309.

Vuillot, Christophe, 2018, "Is error detection helpful on IBM 5Q chips?," Quantum Inf. Comput. **18**, 949–964.

Wallman, Joel, Chris Granade, Robin Harper, and Steven T. Flammia, 2015, "Estimating the coherence of noise," New J. Phys. **17**, 113020.

Wallman, Joel J., and Joseph Emerson, 2016, "Noise tailoring for scalable quantum computation via randomized compiling," Phys. Rev. A **94**, 052325.

Wallraff, A., D. I. Schuster, A. Blais, L. Frunzio, R.-S. Huang, J. Majer, S. Kumar, S. M. Girvin, and R. J. Schoelkopf, 2004, "Strong coupling of a single photon to a superconducting qubit using circuit quantum electrodynamics," Nature (London) **431**, 162–167.

Wang, Samson, Piotr Czarnik, Andrew Arrasmith, M. Cerezo, Lukasz Cincio, and Patrick J. Coles, 2021, "Can error mitigation improve trainability of noisy variational quantum algorithms?," arXiv:2109.01051.

Wang, Samson, Enrico Fontana, M. Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J. Coles, 2021, "Noise-induced barren plateaus in variational quantum algorithms," Nat. Commun. **12**, 6961.

Wang, Zhen, Yanzhu Chen, Zixuan Song, Dayue Qin, Hekang Li, Qiujiang Guo, H. Wang, Chao Song, and Ying Li, 2021, "Scalable Evaluation of Quantum-Circuit Error Loss Using Clifford Sampling," Phys. Rev. Lett. **126**, 080501.

Watanabe, Yu, Takahiro Sagawa, and Masahito Ueda, 2010, "Optimal Measurement on Noisy Quantum Systems," Phys. Rev. Lett. **104**, 020401.

Wecker, Dave, Matthew B. Hastings, and Matthias Troyer, 2015, "Progress towards practical quantum variational algorithms," Phys. Rev. A **92**, 042303.

Wu, Yulin, *et al.*, 2021, "Strong Quantum Computational Advantage Using a Superconducting Quantum Processor," Phys. Rev. Lett. **127**, 180501.

Xue, Xiao, Maximilian Russ, Nodar Samkharadze, Brennan Undseth, Amir Sammak, Giordano Scappucci, and Lieven M. K. Vandersypen, 2022, "Quantum logic with spin qubits crossing the surface code threshold," Nature (London) **601**, 343–347.

Yamamoto, Kaoru, Suguru Endo, Hideaki Hakoshima, Yuichiro Matsuzaki, and Yuuki Tokunaga, 2022, "Error-Mitigated Quantum Metrology via Virtual Purification," Phys. Rev. Lett. **129**, 250503.

Yang, Yongdan, Bing-Nan Lu, and Ying Li, 2021, "Accelerated quantum Monte Carlo with mitigated error on noisy quantum computer," PRX Quantum **2**, 040361.

Yen, Tzu-Ching, Robert A. Lang, and Artur F. Izmaylov, 2019, "Exact and approximate symmetry projectors for the electronic structure problem on a quantum computer," J. Chem. Phys. **151**, 164111.

Yeter-Aydeniz, Kübra, Raphael C. Pooser, and George Siopsis, 2020, "Practical quantum computation of chemical and nuclear energy levels using quantum imaginary time evolution and Lanczos algorithms," npj Quantum Inf. **6**, 63.

Yoshioka, Nobuyuki, Hideaki Hakoshima, Yuichiro Matsuzaki, Yuuki Tokunaga, Yasunari Suzuki, and Suguru Endo, 2022, "Generalized Quantum Subspace Expansion," Phys. Rev. Lett. **129**, 020502.

Zhang, Shuaining, Yao Lu, Kuan Zhang, Wentao Chen, Ying Li, Jing-Ning Zhang, and Kihwan Kim, 2020, "Error-mitigated quantum gates exceeding physical fidelities in a trapped-ion system," Nat. Commun. **11**, 587.

Zhu, D., S. Johri, N. M. Linke, K. A. Landsman, C. Huerta Alderete, N. H. Nguyen, A. Y. Matsuura, T. H. Hsieh, and C. Monroe, 2020, "Generation of thermofield double states and critical ground states with a quantum computer," Proc. Natl. Acad. Sci. U.S.A. **117**, 25402–25406.