

Analyzing the impact of time-correlated noise on zero-noise extrapolation

Kevin Schultz,^{1,*} Ryan LaRose,^{2,3} Andrea Mari,³ Gregory Quiroz,¹
Nathan Shammah,³ B. David Clader,^{1,†} and William J. Zeng^{3,4}

¹*Johns Hopkins University Applied Physics Laboratory, Laurel, MD, 20723, USA*

²*Department of Computational Mathematics, Science, and Engineering,
Michigan State University, East Lansing, MI, 48823, USA*

³*Unitary Fund*

⁴*Goldman, Sachs & Co, New York, NY*

(Dated: September 12, 2022)

Zero-noise extrapolation is a quantum error mitigation technique that has typically been studied under the ideal approximation that the noise acting on a quantum device is not time-correlated. In this work, we investigate the feasibility and performance of zero-noise extrapolation in the presence of time-correlated noise. We show that, in contrast to white noise, time-correlated noise is harder to mitigate via zero-noise extrapolation because it is difficult to scale the noise level without also modifying its spectral distribution. This limitation is particularly strong if “local” gate-level methods are applied for noise scaling. However, we find that “global” noise scaling methods, e.g., global unitary folding, can be sufficiently reliable even in the presence of time-correlated noise. We also introduce gate Trotterization as a new noise scaling technique that may be of independent interest.

I. INTRODUCTION

The theory of fault-tolerant error-corrected quantum computation may result in speed-ups in a number of computations, most notably the factoring of numbers using Shor’s algorithm [1–3], but also in quantum simulation and chemistry [4–8], linear systems [9–12], and other areas [13–18]. While ongoing progress has improved the performance of individual qubits and has allowed quantum computers to scale to larger number of qubits, current systems are not sufficiently performant for useful fault-tolerant operations. Despite this apparent limitation, we are in or rapidly nearing a regime where quantum systems could perform useful computations without (or with less) error correction, the so called noisy intermediate-scale quantum (NISQ) era [19].

In this regime of NISQ computations, it is imperative that any potential errors be reduced or mitigated in order to maximize the utility from these imperfect devices and/or small distance codes. A number of potential techniques have been proposed to mitigate errors in the NISQ regime including quantum control [20–31], decoherence-free subspaces [32–34], readout error mitigation [35–38], Pauli frame randomization [39–41], and optimal compilation [42–47]. One recently-proposed technique motivated by NISQ limitations is zero-noise extrapolation (ZNE) [48–53]. This aims to mitigate the impacts of any errors on a computation by performing a series of computations with *scaled* error levels then post-processing to interpolate to the *zero-noise limit* of the computation.

ZNE techniques have been primarily investigated under the assumption that the errors to be mitigated are uncorrelated in time. On the other hand, time-correlated

noise (in particular $1/f^\alpha$ noise) has been widely observed in physical systems including superconducting devices [54–58], quantum dots [59, 60], and spin qubits [61]. In NISQ devices, such as those offered by the IBM Quantum Experience, evidence of correlated noise has been observed both indirectly through the use of dynamical error suppression [62, 63] and directly through quantum noise spectroscopy (QNS) estimation of the noise [64]. This has been further substantiated by recent studies that have suggested the dynamics of such devices are more accurately captured by non-Markovian models [65, 66].

To estimate the noise present in these real physical systems, one can use QNS [67–69] wherein the outcomes of a set of distinct control pulses or circuits are analyzed. Key to this approach is that while these different probe sequences may in fact represent identical circuits under ideal conditions, they interact with any noise present in different ways. This can be understood through the filter function formalism [70, 71] which describes the “frequency response” of a given probe sequence. Broadly speaking, the impacts of noise (in terms of fidelity) are approximately proportional to the integral of the product of the power spectrum of the noise with the filter function of the control, called an overlap integral. In what follows, we will show how this intuition can also be applied to different ZNE schemes in the presence of temporally correlated dephasing noise.

The recently developed [72] and experimentally validated [64] Schrödinger wave autoregressive moving average (SchWARMA) technique provides a natural mechanism for the exploration of so-called *digital* ZNE techniques [51, 52, 73] that operate at the gate level in a quantum circuit. Building on techniques from classical time-series modeling in statistics and signal processing, SchWARMA was conceived as a highly flexible mechanism for simulating a wide-range of spatiotemporally correlated errors in quantum circuits.

In the following, we first review the SchWARMA

* kevin.schultz@jhuapl.edu

† Current Affiliation: Goldman, Sachs & Co, New York, NY

modeling and simulation formalism as well as a concise overview of ZNE and discuss different methods for scaling noise. Next, we show how these different schemes are impacted by time-correlated dephasing noise despite the fact that they behave equivalently for uncorrelated noise. We then interpret these noise scaling schemes using the language of filter functions and show that these results are well described by the intuition provided by the filter functions. Our findings indicate that, for time-correlated noise, the noise scaling method known as *global unitary folding* [51, 74] produces more accurate noise-scaled expectation values and ZNE results.

II. BACKGROUND

A. Time-correlated noise: The SchWARMA model

Consider a single-qubit Hamiltonian

$$H(t) = H_z(t) + H_c(t) \quad (1)$$

consisting of a semiclassical dephasing noise component $H_z(t)$ along with a deterministic idealized control component $H_c(t)$ corresponding, for example, to the external driving induced by laser pulses. If we further define $H_z(t) = \eta(t)\sigma^z$ with $\eta(t)$ a wide-sense stationary Gaussian stochastic process, we can say that this noise process is not time-correlated if $\mathbb{E}[\eta(t)\eta(t')] = \mathbb{E}[\eta(|t-t'|)\eta(0)] = 0$ for all $t \neq t'$, where $\mathbb{E}(\cdot)$ represents the average over many statistical realizations. σ^i , $i = x, y, z$ are the Pauli matrices. Equivalently, we can say that the noise process is time-correlated if the power spectrum

$$S_\eta(\omega) = \int_0^\infty dt \mathbb{E}[\eta(t)\eta(0)]e^{-i\omega t} \quad (2)$$

is not constant as a function of ω (i.e., not a “white” process). This semiclassical noise setting is the standard setting for QNS [67–69] and is an alternative to general open quantum systems approaches that consider couplings to quantum baths. The semiclassical noise approximation assumes that the bath is in thermal equilibrium and at infinite temperature, yielding regimes with no back action on the environment from the qubits, as well as equal populations of qubit states after long term decay [75–81].

In the SchWARMA modeling approach [72], the impact of the continuous time Hamiltonian in (1) is modeled in a quantum circuit formalism by inserting correlated Z -error operators after each “gate” determined by the control H_c . This is accomplished by generating a time-correlated sequence of rotation angles y_k defined from independent Gaussian inputs x_k using an autoregressive moving average (ARMA) model [82, 83],

$$y_k = \underbrace{\sum_{i=1}^p a_i y_{k-i}}_{AR} + \underbrace{\sum_{j=0}^q b_j x_{k-j}}_{MA}, \quad (3)$$

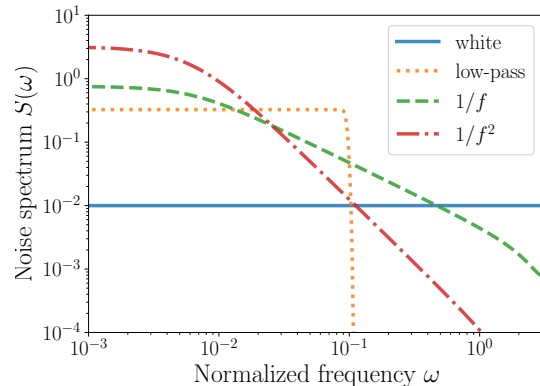


FIG. 1. Noise power spectrum of four different dephasing SchWARMA noise models corresponding to white noise, low-pass noise, $1/f$ noise and $1/f^2$ noise. These noise models are used in Sec. III to test the effect of time-correlated noise on zero-noise extrapolation.

where the set $\{a_i\}$ defines the autoregressive portion of the model, and $\{b_j\}$ the moving average portion with p and $q + 1$ elements of each set respectively. The time correlations are defined via the resulting power spectrum

$$S_y(\omega) = \frac{|\sum_{k=0}^q b_k \exp(-ik\omega)|^2}{|1 + \sum_{k=1}^p a_k \exp(-ik\omega)|^2}, \quad (4)$$

and ARMA models can approximate *any* discrete-time power spectrum to arbitrary accuracy [84]. For the scope of this work we focus on the four paradigmatic noise spectra shown in Fig. 1, namely: white noise, low-pass noise, $1/f$ noise and $1/f^2$ noise.

Dividing the circuit trajectory defined by $H_c(t)$ into consecutive gates G_k , the SchWARMA approach models the impact of correlated noise $H_z(t)$ by adding in a random $Z(\theta_k) = \exp(iy_k\sigma^z)$ after each gate, which can then be Monte Carlo averaged to produce an expectation value. This model can be extended to multi-qubit Hamiltonians

$$H(t) = \sum_{j=1}^n \eta_j(t)\sigma_j^z + H_c(t), \quad (5)$$

by generating independent, yet identically defined, SchWARMA-generated errors on each qubit. In principle, these could of course be heterogeneous and correlated between qubits.

B. Zero-noise extrapolation with colored noise

Zero-noise extrapolation (ZNE) is an error mitigation technique which relies on the ability to increase the noise in a quantum circuit [48, 49, 85]. Like other error mitigation techniques, the target is to estimate an expectation

value

$$E(\lambda) := \text{Tr}[\rho(\lambda)O] \quad (6)$$

at zero noise. The noise scale factor λ dictates how much the base noise level $\lambda = 1$ is scaled in the quantum circuit which prepares the system density matrix ρ , and O is a problem-dependent observable. The key insight of ZNE is to (i) evaluate $E(\lambda)$ at several noise scale factors $\lambda \geq 1$, then (ii) fit a statistical model to the collected data and infer the zero-noise value $E(\lambda \rightarrow 0)$. We refer to these two steps as noise scaling and inference, respectively.

Compared to other error mitigation techniques, zero-noise extrapolation requires very few additional quantum resources. Correspondingly, it has received some attention in recent literature; e.g., it was implemented in Refs. [50, 51, 73, 74, 86, 87] and in [53] on twenty six superconducting qubits to produce results competitive with classical approximation techniques. References [51, 52, 73] formally introduced digital noise scaling, in which noise is scaled at a gate-level without pulse-level control.

While ZNE is straightforward to implement and requires relatively few additional quantum resources, the quality of the solution depends critically on both the inference and noise-scaling method and can be improved by a correct characterization of the hardware noise. In this work, we fix the inference method by assuming a particular noise model and focus on the effects of the noise-scaling method.

1. Noise scaling methods

a. Ideal noise scaling In a purely theoretical setting, the ideal way of scaling the noise would be to multiply the Hamiltonian H_z in Eq. (1) by a constant $\sqrt{\lambda}$:

$$H'(t) = \sqrt{\lambda}H_z(t) + H_c(t). \quad (7)$$

Equivalently, the scale factor can be absorbed into a redefinition of the stochastic noise amplitude: $\eta'(t) = \sqrt{\lambda}\eta(t)$. From Eq. (4), it is evident that the noise power spectrum gets scaled by λ ,

$$S_{\eta'}(\omega) = S_{\sqrt{\lambda}\eta}(\omega) = \lambda S_{\eta}(\omega). \quad (8)$$

If one could directly control the noise, this would be the ideal way of scaling its power and, therefore, the ideal way of applying zero-noise extrapolation. In simulations using SchWARMA, noise can be scaled by transforming the numerator coefficients $b_k \rightarrow \sqrt{\lambda}b_k$. In a typical experimental scenario, of course, one cannot directly control the noise of a quantum device. Even in instances where it is possible to scale the noise spectrum through e.g., manipulating the master clock [64] or flux lines, precise characterization of the native noise spectrum and calibration of the noise injection would be required. Due to these difficulties in directly scaling noise, several indirect noise scaling techniques have been proposed and

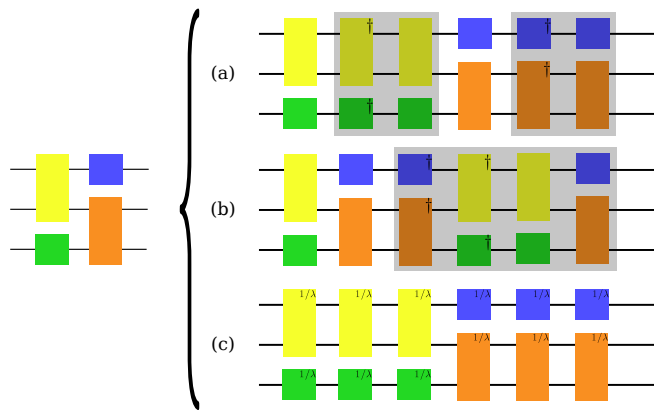


FIG. 2. A sample three-qubit circuit with four gates under the action of three digital noise scaling methods we consider in this work. (a) Local folding, in which each gate G gets mapped to $G(G^\dagger G)^n$ for scale factor $\lambda = 2n - 1$. (b) Global folding, in which the entire circuit C gets mapped to $C(C^\dagger C)^n$. In (a) and (b), grey shading shows the “virtual gates” which logically compile to identity. (c) Gate Trotterization, in which $G \mapsto (G^{1/\lambda})^\lambda$ for each gate G .

applied in recent literature. We define several of these in the following subsections (see Fig. 2 for an overview) in order to analyze their performance in the presence of time-correlated noise in Sec. III.

b. Pulse stretching The intent of pulse stretching is to scale the impacts of the noise on the system by “stretching” the underlying control Hamiltonian, replacing (1) with

$$H(t) = H_z(t) + \frac{1}{\lambda}H_c(t/\lambda), \quad (9)$$

for some dimensionless time-scaling factor λ . In principle, this scales the impacts of the noise by increasing the overall time duration of the circuit. More precisely, if we define $t' = t/\lambda$, the density operator $\rho(t')$ of the system evolves with respect to the effective Hamiltonian:

$$H'(t') = \lambda H_z(\lambda t') + H_c(t'). \quad (10)$$

The corresponding noise power spectrum is:

$$\begin{aligned} S_{\eta'}(\omega) &= \lambda^2 \int_0^\infty dt' \mathbb{E}[\eta(\lambda t')\eta(0)]e^{-i\omega t'} \\ &= \lambda \int_0^\infty dt \mathbb{E}[\eta(t)\eta(0)]e^{-i\omega t/\lambda} = \lambda S_{\eta}(\omega/\lambda). \end{aligned} \quad (11)$$

From the equation above, it is evident that for a white (constant) spectrum, pulse stretching can be used to effectively scale the noise power by λ as in the ideal case defined in Eq. (8). In fact, the equivalence between the ideal noise scaling and the pulse-stretching technique was already shown in Ref. [48], under the hypothesis of a quantum state ρ evolving according to a master equation

with a time-independent noise operator acting as $\mathcal{L}(\rho)$ (more details about the consistency between our findings with the results of Ref. [48] are given in Appendix A). On the other hand, Eq. (11) shows that, for a colored spectrum, pulse-stretching does not exactly reproduce the ideal noise scaling defined in (8). Indeed, on the r.h.s. of Eq. (11) we observe that the original spectrum is also stretched with respect to the frequency variable ω . This fact is a manifestation of the intuitive idea that slowing down the dynamics the system corresponds to effectively speeding up the time scale of the environment. Such frequency stretching, while irrelevant in the white noise limit, becomes relevant for time-correlated noise.

In the SchWARMA formalism, there is not a mechanism for stretching pulses *per se* as it operates at the gate level in a circuit (without pulse-level control on $H_c(t)$). However, as discussed in the supplement to [72], it is possible to manipulate and stretch the spectrum of a SchWARMA model. So, for the task of numerically simulating pulse stretching, instead of implementing equation Eq. (10) one can simply implement Eq. (11) by directly transforming the spectrum of the SchWARMA model.

c. Local unitary folding A possible way of effectively increasing the noise of a circuit is to insert after each noisy CNOT gate, the product of two additional CNOT gates [52, 73]. In this way the ideal unitary is not changed, but the real dynamics are more noisy. More generally, Ref. [51] introduced several digital noise scaling methods that are based on the *unitary folding* replacement rule

$$G \rightarrow G(G^\dagger G)^n, \quad n = 0, 1, 2, \dots, \quad (12)$$

where G is a unitary operation associated to an individual gate. If noise is absent, the replacement rule leaves the operation unchanged since $G^\dagger G$ is equal to the identity. On the contrary, if some base noise is associated to G , the unitary folding operation approximately scales the noise by an odd integer factor $\lambda = 1 + 2n$.

More precisely, by applying the unitary folding replacement to all the gates of an input circuit

$$U = G_d G_{d-1} \dots G_1 \quad (13)$$

which is composed of d gates G_j , we obtain new circuit U' of depth $d' = (1 + 2n)d$ given by

$$U' = G_d (G_d^\dagger G_d)^n G_{d-1} (G_{d-1}^\dagger G_{d-1})^n \dots G_1 (G_1^\dagger G_1)^n. \quad (14)$$

The depth of the new circuit U' is scaled by $\lambda = d'/d = 1 + 2n$ and, similarly, any type of noise which depends on the total number of gates will be effectively scaled by the same constant λ . In Ref. [51], partial folding methods were proposed to obtain arbitrary real values of λ , but for simplicity in this work we only consider odd-integer scale factors. We refer to (14) as *local unitary folding*.

d. Global unitary folding Instead of locally folding all the gates, we can apply Eq. (12) to the entire circuit. In this way, the circuit U defined in Eq. (13) is simply

mapped to

$$U' = U(U^\dagger U)^n. \quad (15)$$

Also in this case the total number of gates of the new circuit U' is multiplied by $\lambda = d'/d = 1 + 2n$ corresponding to an effective scaling of the noise.

e. Gate Trotterization In this work we also introduce another local noise-scaling method, acting at the level of individual gates, that we call *gate Trotterization* since it can be considered as a discretization of the continuous pulse-stretching technique. According to the gate Trotterization technique, each gate of the circuit is replaced as follows:

$$G \rightarrow \left(G^{1/\lambda}\right)^\lambda, \quad \lambda = 0, 1, 2, \dots \quad (16)$$

For example, a Pauli X rotation gate $R_X(\theta)$ is replaced by λ applications of $R_X(\theta/\lambda)$. Eq. (16) is similar to the local version of the unitary folding rule (12) and, indeed, both methods replace a single gate with the product of λ gates. Compared to Eq. (12), the Trotter-like decomposition used in Eq. (16) is more uniform since equal elementary gates are used. On the other hand, a possible drawback of the gate Trotterization method is that $G^{1/\lambda}$ may be compiled by the hardware in different ways depending on λ and, therefore, the circuit depth may not get scaled as expected.

III. RESULTS

In the previous section, we defined several noise-scaling methods that can be used in zero-noise extrapolation. In this section, we study how these different methods affect the performance of ZNE in the presence of time-correlated noise. For all the simulations presented in this section we used the following Python libraries: Mezza [88] for modeling SchWARMA noise, Mezza's TensorFlow Quantum [89] interface for simulating quantum circuits and Mitiq [74] for applying unitary folding and zero-noise extrapolation. Code for specifying the circuits and the dephasing noise spectra used is also available at [88].

A. Zero-noise extrapolation with colored noise

In this section we numerically simulate a simple ZNE experiment with different noise scaling methods and with different noise spectra. The results are reported in Fig. 3 and demonstrate the detrimental effect of time-correlated noise on ZNE. In Fig. 3(a) the noise spectrum is white and all noise scaling methods produce nearly identical expectation values. Correspondingly, the zero-noise limits (marked with stars in the plot) are nearly identical. On the other hand, in Fig. 3(b), the noise is colored (a $1/f$ "pink" spectrum) and different noise-scaling methods produce different expectation values. Correspondingly, the zero-noise limits (marked with stars in the plot)

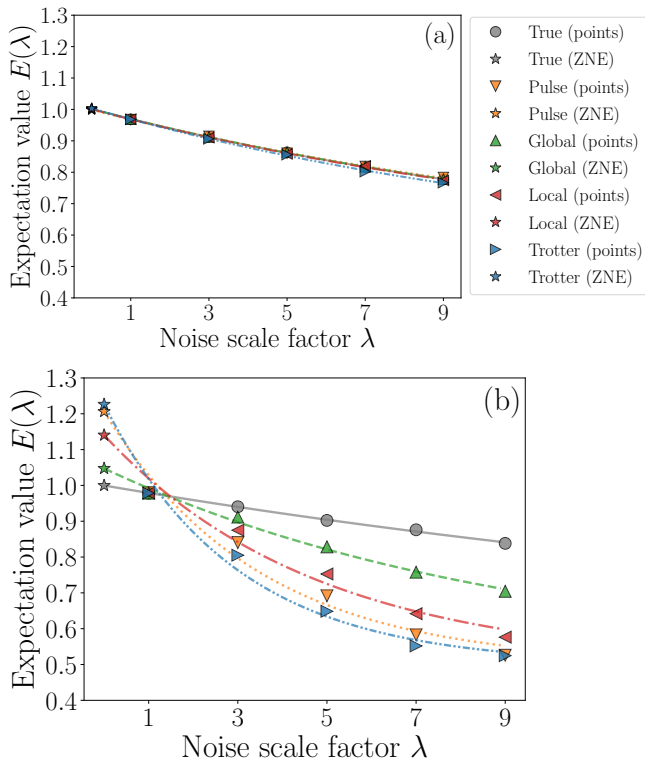


FIG. 3. Comparison of different zero-noise extrapolations obtained with different noise scaling methods. We consider a representative (randomly generated) single-qubit randomized benchmarking circuit affected by dephasing noise of fixed integrated power. The two subfigures correspond to different noise spectra: (a) white noise, (b) $1/f$ pink noise. Both spectra are shown in Fig. 1. The expectation value $E(\lambda) = \text{Tr}[O\rho(\lambda)]$ is associated to the observable $O = |0\rangle\langle 0|$ measured with respect to the noise-scaled quantum state $\rho(\lambda)$. The colored triangles represent the noise-scaled expectation values; the dashed-dotted lines represent the associated exponential fitting curves; the colored stars represent the corresponding zero-noise extrapolations. The figure shows that the zero-noise limit obtained with global unitary folding (green star) is relatively close to the ideal result (gray star) even in the presence of strong time correlations in the noise. The “true” points (gray circles) are obtained by assuming the ideally scaled noise spectrum of Eq. (8), the associated fitting curve (solid gray line) produces the ideal zero-noise extrapolation (gray star).

are also different. This is the main qualitative result that this work aims to highlight: compared to white noise, time-correlated noise can be much harder to mitigate via zero-noise extrapolation.

In the rest of this section, we study this aspect in a more quantitative way. In particular we study the performances of different noise-scaling methods for different types of noise spectra and different types of circuits.

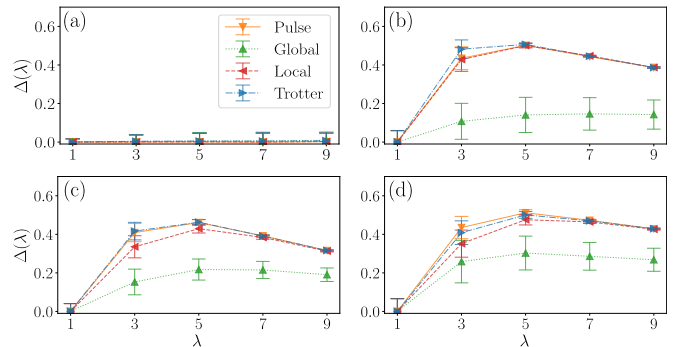


FIG. 4. Average relative errors in noise scaling two-qubit randomized benchmarking circuits with (a) white noise, (b) lowpass noise, (c) $1/f$ noise, and (d) $1/f^2$ noise. Panel (a) shows no significant difference in scaling methods under white noise (no time correlations). Panels (b)-(d) show that global scaling is the lowest-error digital scaling method. The two-qubit randomized benchmarking circuits used here have, on average, 27 single-qubit gates and five two-qubit gates. For each circuit execution, 3000 samples were taken to estimate the probability of the ground state as the observable. Points show the average results over fifty such circuits and error bars show one standard deviation.

B. Comparing noise scaling methods

Observing Fig. 3(b) we notice that, at least for the particular circuit considered in the example, some noise scaling methods perform better than others in the presence of time-correlated noise. In particular the extrapolation based on the global folding technique produces a relatively good approximation of the ideal result even in the presence of time-correlated noise.

To better investigate this phenomenon, we consider the *relative noise-scaling error*

$$\Delta(\lambda) := \left| \frac{E(\lambda) - E^*(\lambda)}{E^*(\lambda)} \right|, \quad (17)$$

as a figure of merit. Here, $E(\lambda)$ is the expectation value of interest evaluated with some particular noise scaling method and scale factor λ , and $E^*(\lambda)$ is the expectation value simulated with a noise spectrum ideally scaled according to Eq. (8). In Fig. 4 we plot the relative error defined in Eq. (17) for each noise-scaling method, after averaging the results over multiple instances of two-qubit randomized-benchmarking circuits. Here the expectation value of the observable $O = |00\rangle\langle 00|$ is considered. The results of Fig. 4 are consistent with those of Fig. 3 discussed in the previous subsection. In fact, even after averaging over multiple random circuits, we observe that in the presence of white noise all noise scaling methods are practically equivalent to each other and are characterized by a small relative noise-scaling error. However, for all colored noise spectra, global folding is optimal when compared to other noise scaling methods.

We repeat the same experiments using mirror circuits [90] and QAOA-like circuits instead of RB circuits.

The former provides another type of randomized circuit structure used for benchmarking, and the latter provides a structured circuit. Fig. 5 shows the results using two-qubit mirror circuits. These circuits have 26 single-qubit gates and eight two-qubit gates on average. As with the randomized benchmarking circuits, 3000 samples were taken when executing each circuit to estimate the probability of sampling the correct bitstring. As shown in Fig. 5, the conclusion that global unitary folding most closely matches true noise scaling holds on average for mirror circuits as well. These results were averaged over fifty random mirror circuits.

Fig. 6 shows the same experiment using QAOA circuits. These $n = 2$ qubit circuits have $p = 2$ QAOA rounds using the standard mixer Hamiltonian $H_M = \sum_{i=1}^n \sigma_i^x$ and driver Hamiltonian $H_C = \sum_{ij} \sigma_i^z \sigma_j^z$. Denoting this circuit as U , we append U^\dagger such that the final noiseless state is $|00\rangle$ independent of the randomly chosen angles β and γ . A total of fifty circuits with random angles were simulated for the final results, again using 3000 samples to estimate the ground state probability for each circuit execution. The results in Fig. 6 have the highest variance of the three circuit types, but on average we still see that global unitary folding is closest to true noise scaling out of all scaling methods considered.

The conclusions of this subsection suggest that, even for different types of circuits, the effect of time-correlated noise on noise scaling methods is qualitatively similar. This intuition is consistent with the theoretical discussion presented in the next section, in which the performances of noise scaling methods are linked to their effective frequency modulation effects.

We emphasize that the comparison considered in this work is focused on one particular figure of merit: the robustness of a noise scaling method with respect to time-correlated noise. Our results suggest that global folding outperforms the other methods considered with respect to this specific figure of merit. In a real-world scenario, the optimal noise-scaling method should be determined according to a more general cost-benefit analysis, e.g., taking into account the sampling cost, coherence time, and other hardware limitations. For instance, it may not be possible to use global noise scaling if the circuit length is comparable to the coherence time of the computer; in such circumstances, pulse stretching can amplify errors via small scale factors [53], although potentially inaccurately in the presence of time-correlated noise as we have shown in this section.

IV. DISCUSSION AND PHYSICAL INTERPRETATION

In classical signal processing and control theory, the frequency response of a linear circuit or filter is used to understand how a circuit interacts with its input in the frequency domain. Using frequency domain techniques, one can understand and design filters that amplify (or

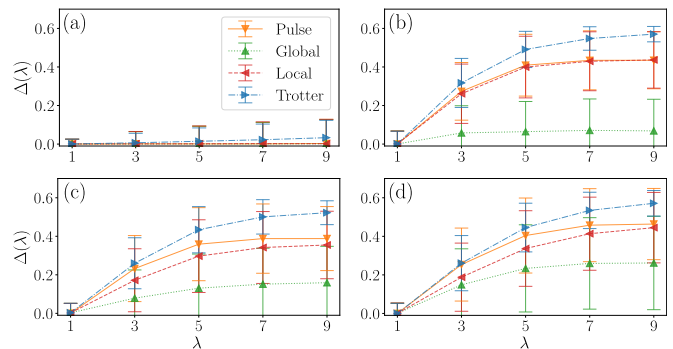


FIG. 5. Relative errors in noise scaling two-qubit mirror circuits with (a) white noise, (b) lowpass noise, (c) $1/f$ noise, and (d) $1/f^2$ noise. Panel (a) shows no significant difference in scaling methods under white noise (no time correlations). Panels (b), (c) and (d) show global scaling is optimal with time-correlated noise. The two-qubit mirror benchmarking circuits used here have, on average, 26 single-qubit gates and eight two-qubit gates. For each circuit execution, 3000 samples were taken to estimate the probability of the correct bitstring (defined by the particular mirror circuit instance) as the observable. Points show the average results over fifty such circuits and error bars show one standard deviation.

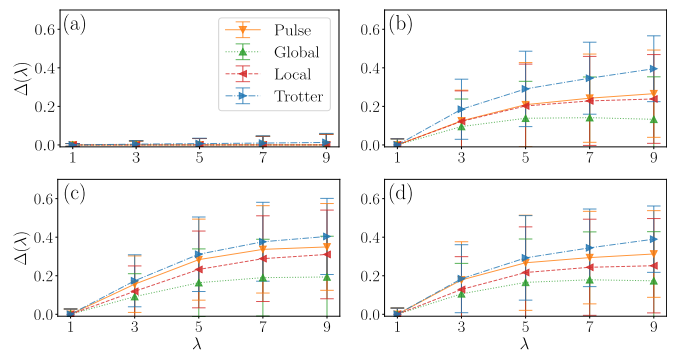


FIG. 6. Relative errors in noise scaling two-qubit $p = 2$ QAOA circuits with (a) white noise, (b) lowpass noise, (c) $1/f$ noise, and (d) $1/f^2$ noise. Panel (a) shows no significant difference in scaling methods under white noise (no time correlations). Panels (b), (c) and (d) show global scaling is optimal with time-correlated noise. The two-qubit $p = 2$ QAOA circuits used here have eight single-qubit gates and four two-qubit gates. For each circuit execution, 3000 samples were taken to estimate the probability of the ground state as the observable. (Note that the QAOA circuit U is echoed such that the total circuit is $UU^\dagger = I$ without noise.) Points show the average results over fifty such circuits and error bars show one standard deviation.

pass through) frequencies that have signal content while attenuating frequencies that contain only noise. In what follows, we introduce an analogous concept for quantum circuits that allows us to approximate the fidelity of a circuit subjected to dephasing noise. This circuit frequency response indicates where a given circuit is particularly sensitive or insensitive to noise in a given frequency range. We then analyze the impact of the various noise

scaling techniques on this circuit frequency response to interpret the results of the previous section.

A. Frequency response of a circuit

The natural extension of the frequency response of a circuit to the quantum context is the so-called filter function formalism [70, 71]. Details of the specific approach used here for multi-qubit, spatiotemporally correlated dephasing noise can be found in Appendix B, but the gist of the technique is that a circuit on n qubits of time duration T defines a set of real-valued switching functions

$$f_{\alpha\beta}(t) = \frac{1}{N} \text{Tr} [U_0(T, t) A_\alpha U_0^\dagger(T, t) A_\beta], \quad (18)$$

defined by the action of a circuit's reverse-time propagator U_0 (see Eq. (B2)) on a set of traceless, Hermitian operators $\{A_\alpha\}$ that satisfy $\langle A_\alpha, A_\beta \rangle = \frac{1}{N} \text{Tr}[A_\alpha A_\beta] = \delta_{\alpha, \beta}$ where $N = 2^n$ (typically, the A_α are multi-qubit Pauli matrices).

The Fourier transforms $F_{\alpha\beta}(\omega, T) = \int_0^T dt f_{\alpha\beta}(t) e^{i\omega t}$ of these switching functions are used to define filter functions

$$\mathcal{F}_{\alpha\beta, \alpha'\beta'}(\omega, T) = \text{Re} [F_{\alpha\beta}(\omega, T) F_{\alpha'\beta'}(-\omega, T)]. \quad (19)$$

Similar to classical frequency domain analysis, these filter functions interact with the dephasing noise spectra $S_{\alpha, \alpha'}(\omega)$ in a multiplicative fashion. Their product forms the integrand of the so-called overlap integral which is a key component of the second cumulant $C_O^{(2)}$:

$$\frac{C_O^{(2)}(T)}{2} = \sum_{\alpha, \beta, \alpha', \beta'} \int_0^\infty \frac{d\omega}{2\pi} S_{\alpha, \alpha'}(\omega) \mathcal{F}_{\alpha\beta, \alpha'\beta'}(\omega, T) \mathcal{A}_{\beta\beta'}. \quad (20)$$

The overlaps between the noise power spectrum and filter functions scale operators $\mathcal{A}_{\beta\beta'}$ that are dependent on the observable O , see Eq. (B9). The magnitude of $C_O^{(2)}$ (and thus the overlaps) can then be used to approximate the expectation of the noisy observable $\overline{\langle O \rangle} \approx \text{Tr}[\exp(-C_O^{(2)}/2) \rho_{S,0}(T) O]$, where $\rho_{S,0}(T)$ is the final state of the ideal noiseless circuit, see discussion around Eq. (B5) in Appendix B.

The expression in Eq. (20) captures potential cross correlations in noise, but here since we consider independent σ^z -dephasing noise on each qubit, $S_{\alpha, \alpha'} = 0$ when $\alpha \neq \alpha'$ and α is not a σ^z operator on a given qubit. Furthermore, for the examples below we compute the filter functions using instantaneous gates as specified by a circuit, but these expressions hold for piecewise constant controls to accommodate pulse shaping. In the context of noise scaling experiments, Eq. (20) provides a mechanism for understanding how the different noise scaling techniques perturb the filter functions to impact the resulting scaled expectations and thus the extrapolation process.

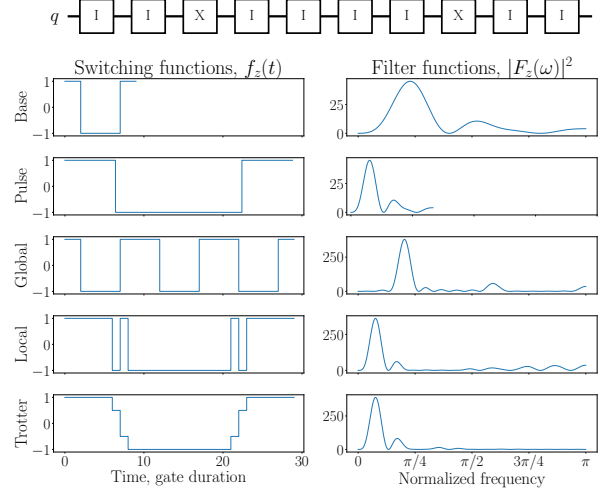


FIG. 7. Filter function analysis of a CPMG circuit with a delay of 2 gate times. Bottom: Plots of switching functions and filter functions of the CPMG circuit and its various scaled versions.

The calculations for the multi-qubit case are quite involved, so in order to build better intuition we will also consider the simpler case of a single qubit subject to dephasing noise and single-axis σ^x control, i.e., $H(t) = \eta(t)\sigma^z + \Omega(t)\sigma^x$. For this case, the filter function formalism can be recast in terms of a single *complex* switching function $f_z(t) = \exp\left(-i \int_0^t d\tau \Omega(\tau)\right)$ with Fourier transform $F_z(\omega)$. The filter function $\mathcal{F}_{zz}(\omega) = |F_z(\omega)|^2$ and power spectrum $S_{z,z}(\omega) = S_\eta(\omega)$ define the overlap integral

$$\chi = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} S_\eta(\omega) |F_z(\omega)|^2. \quad (21)$$

The overlap integral can be used to derive an approximation to the expectation of the noisy circuit states ρ and an observable O , via

$$E[\text{Tr}[\rho O]] \approx A + B \exp(-\chi), \quad (22)$$

where A and B are functions of the ideal final state and observable O . An example Carr-Purcell-Meiboom-Gill (CPMG) [91, 92] circuit and its corresponding f_z and $|F_z(\omega)|^2$ are shown in Fig. 7.

B. Spectral analysis of noise scaling methods

First, we will consider the simpler case of a single qubit subject to dephasing noise and single-axis σ^x control, with the CPMG circuit in Fig. 7 serving as our canonical example. Using the filter function prediction from Eq. (22) we have that direct noise scaling produces states $\rho_{dir}(\lambda)$ with expectation

$$E[\text{Tr}[\rho_{dir}(\lambda) O]] \approx A + B \exp(-\lambda\chi), \quad (23)$$

where χ is the overlap integral of the base circuit. Note that direct noise scaling does not affect the circuit itself, and thus its switching and filter functions are unchanged. Similarly, following Eq. (10), we have that pulse stretching produces the expectation

$$E[\text{Tr}[\rho_{pul}(\lambda) O]] \approx A + B \exp\left(-\lambda \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} S_{z,z}(\omega/\lambda) |F_z(\omega)|^2\right), \quad (24)$$

with similar expressions for Eq. (20), which is clearly not equal to Eq. (23) in general. Equivalently, stretching the pulse amounts to stretching the switching functions and thus “compressing” a filter function response by a factor of λ . This shifts the filter function to lower frequencies, and thus the overlap with low-frequency noise will likely increase by a factor greater than λ . An example of the impact of pulse stretching on a CPMG circuit is shown in Fig. 7, showing that the switching function is perfectly scaled in time, resulting in the corresponding frequency compression.

Gate Trotterization is similar in spirit to pulse stretching, but performed “digitally.” However, repeating a gate’s control waveform λ times with amplitude $1/\lambda$ is in general different from stretching a gate’s control waveform (except in the case of rectangular pulses). Fig. 7 shows a similar qualitative impact of gate Trotterization on the filter function as pulse stretching, in that the filter function is compressed to the low frequencies. Since, unlike pulse stretching, the switching function now has intermediate values between ± 1 the filter function is distorted and not a “perfect” compression.

Like pulse stretching and gate Trotterization, local folding also increases the proportion of the filter function that overlaps with low frequency noise. However unlike pulse stretching and gate Trotterization, local folding also appears to generate response at high frequency. Qualitatively, local folding “pulls” the filter function to the extreme frequencies from the middle of the spectrum. This behavior can be interpreted from the switching function, which now has (brief) oscillations whenever the original switching function had a change, and otherwise remains constant, see Fig. 7. These oscillations increase the high-frequency content of the filter function, whereas the longer duration of constant values increase the low-frequency content. Explicit filter function calculations for local folding can be found in Appendix D that generalize these observations to the multi-qubit case. With these general trends, we would again expect that the overlap integrals produced would not be particularly close to direct noise scaling.

Of the noise scaling methods studied, it appears that global folding preserves the most structure from the unscaled filter function. The circuit response shown in Fig. 7 shows that scaling preserves the qualitative shape of the base circuit’s filter function, but in accordance with well known results about CPMG sequences, the frequency response is sharpened as it is repeated. Quali-

tatively, it looks like the impact of global folding serves to “resolve” a coarse frequency response of the base circuit. Explicit calculations of the filter function (see Appendix C) show that the scaled portion of the circuit dominates the filter function response and approach a common, nontrivial limit. Thus, scaling in this case preserves some structure and produces overlap integrals that are somewhat close to direct noise scaling.

The CPMG sequence considered in the above discussion was chosen as an intuitive example for its well-studied frequency response [67], as well as clarity of exposition. However, the ZNE simulations considered here use multi-qubit random circuits whose frequency response is less well studied. For these more complex circuits, we continue to see the same general trends in the filter function responses, as shown in Fig. 8. These circuits are longer and have greater gate density than the CPMG example, and as such produce switching functions with many transitions that in turn leads to filter functions with many peaks and valleys. The spectral trends for the pulse stretching, local folding, and gate Trotterization methods in Fig. 8 are quite clearly consistent with the CPMG example, and in particular all exhibit increasing low frequency concentration as λ increases (in addition to high frequency concentration for local folding). On the other hand, global folding appears to be approaching some limit that at least somewhat resembles the initial distribution of the frequency response (and can be assessed analytically – see Appendix C). We interpret this as generalization of the sharpening of the spectral features well known for CPMG sequence, and multiple peaks are resolved from initially broad peaks as λ increases.

These observations in the different noise scaling strategies explain the trends in Figs. 3 and 4. As global folding produces scaled filter functions that best preserve the general balance across different frequency ranges, the overlap integrals of the globally folded circuits are the closest to the ideal scaling produced by direct noise scaling. The remaining three scaling approaches all produce some level of concentration at low frequencies, and thus tend to have much greater overlap with the low-frequency noise here. As the pulse stretching and gate Trotterization approaches are very similar in spirit, they produce similar extrapolations. Furthermore, unlike local folding, these two approaches have all their concentration at low frequency, thus producing the most overlap leading to the worst extrapolation error. Local folding, which includes some high frequency content (based on the proportion of the original circuit’s frequency response above $\pi/2$), produces overlaps that lie between the global folding and the stretching/Trotterization approaches.

We note that the trends observed above and the intuition behind them is a direct consequence of the correlated noise classes considered, all of which are fundamentally low frequency. Thus, pulse stretching, gate Trotterization, and local folding produce larger overlaps with the low-frequency noise and drastically bias the noise extrapolation process. In contrast, if the noise was band limited

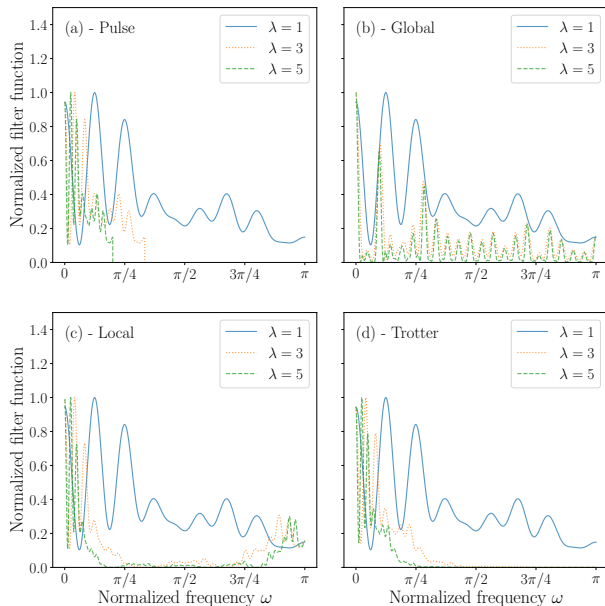


FIG. 8. Largest magnitude filter function of a two-qubit randomized benchmarking circuit of Clifford depth 2 (actual depth 17) for different scale factors λ . All filter functions are normalized by their maximum values (otherwise the integral of the filter function scales by λ). Different subplots correspond to different noise scaling methods. All noise scaling methods change the frequency response of the circuit, however, global folding tends to preserve the qualitative shape of response function and, for this reason, it gives better performances for zero-noise extrapolation with colored noise.

(say between $\pi/4$ and $3\pi/4$ in normalized frequency) we would expect that global folding would continue to track direct noise scaling the best. However, analysis of the other three techniques would be challenging as the overlap integral with these would essentially vanish as the scaling increased. Without knowing the true expectation and the underlying noise spectra, it would be unclear if the leveling out of the scaled expectation values would be due to the overlap integral approaching infinity (i.e., too much noise) or vanishing (i.e., decoupling from the noise). Similarly, if the noise were purely high frequency, we would expect the pulse stretching and gate Trotterization approaches to be insensitive, local folding method to be more sensitive, and global folding between them. Finally, extremely narrow band noise could potentially lie in a “valley” in the scaled response (obviously this is circuit dependent), and thus overlap integrals would vanish for all the noise scaling approaches considered here.

V. CONCLUSION

In this work, we have demonstrated the effect of time-correlated noise on zero-noise extrapolation. Using the SchWARMA technique to model time-correlated dephasing noise, we presented the results of several numerical experiments showing that global unitary folding produces the lowest error relative to direct noise scaling. We analyzed our observed results and provided a physical interpretation in terms of the spectral analysis of the considered noise scaling methods.

Noise injection as a method for noise scaling is a theoretically ideal mechanism for noise scaling, but given the limitations on estimating, emulating, and injecting the native spectra and noise mechanism(s), this is not likely feasible in most situations. On the other hand, global noise folding is broadly applicable and our work suggests its use in global noise scaling for zero-noise extrapolation, if possible, whenever noise may be time-correlated. An obviously important consideration is which quantum computer architectures may have time-correlated noise, a question we do not explicitly consider in this paper, but do note that time-correlated noise has been widely observed in both research-grade qubit experiments and cloud-based NISQ systems, in a variety of platforms [54–66]. We note that global folding is not the only possible noise scaling method suitable for time-correlated noise: other methods could be defined and analyzed, e.g., folding the first half and second half of the gates in a unitary separately. Our work provides the theoretical and practical tools to analyze the performance of such methods under a wide variety of noise models.

DATA AVAILABILITY

Software for reproducing all numerical results is available at <https://github.com/mezze-team/mezze>.

ACKNOWLEDGEMENTS

We thank Sarah Kaiser for helpful discussions. This work was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Accelerated Research in Quantum Computing under Award Number DE-SC0020266 and DE-SC0020316. RL acknowledges support from a NASA Space Technology Graduate Research Fellowship.

[1] P. W. Shor, Algorithms for quantum computation: discrete logarithms and factoring, in *Proceedings 35th annual symposium on foundations of computer science*

(IEEE, 1994) pp. 124–134.

[2] I. L. Chuang, R. Laflamme, P. W. Shor, and W. H. Zurek, Quantum computers, factoring, and decoherence, *Science*

- 270**, 1633 (1995).
- [3] P. W. Shor, Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer, *SIAM review* **41**, 303 (1999).
- [4] J. I. Cirac and P. Zoller, Goals and opportunities in quantum simulation, *Nature Physics* **8**, 264 (2012).
- [5] N. C. Jones, J. D. Whitfield, P. L. McMahon, M.-H. Yung, R. Van Meter, A. Aspuru-Guzik, and Y. Yamamoto, Faster quantum chemistry simulation on fault-tolerant quantum computers, *New J. Phys.* **14**, 115023 (2012).
- [6] A. A. Houck, H. E. Türeci, and J. Koch, On-chip quantum simulation with superconducting circuits, *Nature Physics* **8**, 292 (2012).
- [7] I. M. Georgescu, S. Ashhab, and F. Nori, Quantum simulation, *Rev. Mod. Phys.* **86**, 153 (2014).
- [8] P. J. O’Malley, R. Babbush, I. D. Kivlichan, J. Romero, J. R. McClean, R. Barends, J. Kelly, P. Roushan, A. Tranter, N. Ding, *et al.*, Scalable quantum simulation of molecular energies, *Phys. Rev. X* **6**, 031007 (2016).
- [9] A. W. Harrow, A. Hassidim, and S. Lloyd, Quantum algorithm for linear systems of equations, *Phys. Rev. Lett.* **103**, 150502 (2009).
- [10] B. D. Clader, B. C. Jacobs, and C. R. Sprouse, Preconditioned quantum linear system algorithm, *Phys. Rev. Lett.* **110**, 250504 (2013).
- [11] J. Pan, Y. Cao, X. Yao, Z. Li, C. Ju, H. Chen, X. Peng, S. Kais, and J. Du, Experimental realization of quantum algorithm for solving linear systems of equations, *Phys. Rev. A* **89**, 022313 (2014).
- [12] L. Wossnig, Z. Zhao, and A. Prakash, Quantum linear system algorithm for dense matrices, *Phys. Rev. Lett.* **120**, 050502 (2018).
- [13] P. W. Shor, Introduction to quantum algorithms, in *Proceedings of Symposia in Applied Mathematics*, Vol. 58 (2002) pp. 143–160.
- [14] A. M. Childs and W. van Dam, Quantum algorithms for algebraic problems, *Rev. Mod. Phys.* **82**, 1 (2010).
- [15] S. E. Venegas-Andraca, Quantum walks: a comprehensive review, *Quantum Information Processing* **11**, 1015 (2012).
- [16] M. Schuld, I. Sinayskiy, and F. Petruccione, An introduction to quantum machine learning, *Contemporary Physics* **56**, 172 (2015).
- [17] A. Montanaro, Quantum algorithms: an overview, *npj Quantum Information* **2**, 1 (2016).
- [18] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Quantum machine learning, *Nature* **549**, 195 (2017).
- [19] J. Preskill, Quantum computing in the nisy era and beyond, *Quantum* **2**, 79 (2018).
- [20] L. Viola, E. Knill, and S. Lloyd, Dynamical decoupling of open quantum systems, *Phys. Rev. Lett.* **82**, 2417 (1999).
- [21] L. Viola and E. Knill, Robust dynamical decoupling of quantum systems with bounded controls, *Phys. Rev. Lett.* **90**, 037901 (2003).
- [22] T. Brixner, G. Krampert, T. Pfeifer, R. Selle, G. Gerber, M. Wollenhaupt, O. Graefe, C. Horn, D. Liese, and T. Baumert, Quantum control by ultrafast polarization shaping, *Phys. Rev. Lett.* **92**, 208301 (2004).
- [23] L. M. K. Vandersypen and I. L. Chuang, NMR techniques for quantum control and computation, *Rev. Mod. Phys.* **76**, 1037 (2005).
- [24] N. Khaneja, T. Reiss, C. Kehlet, T. Schulte-Herbrüggen, and S. J. Glaser, Optimal control of coupled spin dynamics: design of nmr pulse sequences by gradient ascent algorithms, *J. Magn. Reson.* **172**, 296 (2005).
- [25] D. d’Alessandro, *Introduction to quantum control and dynamics* (CRC press, 2007).
- [26] M. J. Biercuk, H. Uys, A. P. VanDevender, N. Shiga, W. M. Itano, and J. J. Bollinger, Optimized dynamical decoupling in a model quantum memory, *Nature* **458**, 996 (2009).
- [27] D. Dong and I. R. Petersen, Quantum control theory and applications: a survey, *IET Control Theory & Applications* **4**, 2651 (2010).
- [28] A. M. Souza, G. A. Álvarez, and D. Suter, Robust dynamical decoupling, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **370**, 4748 (2012).
- [29] S. Machnes, E. Assémat, D. Tannor, and F. K. Wilhelm, Tunable, flexible, and efficient optimization of control pulses for practical qubits, *Phys. Rev. Lett.* **120**, 150401 (2018).
- [30] D. Lucarelli, Quantum optimal control via gradient ascent in function space and the time-bandwidth quantum speed limit, *Phys. Rev. A* **97**, 062346 (2018).
- [31] M. Bukov, A. G. R. Day, D. Sels, P. Weinberg, A. Polkovnikov, and P. Mehta, Reinforcement learning in different phases of quantum control, *Phys. Rev. X* **8**, 031086 (2018).
- [32] D. A. Lidar, I. L. Chuang, and K. B. Whaley, Decoherence-free subspaces for quantum computation, *Phys. Rev. Lett.* **81**, 2594 (1998).
- [33] D. Bacon, D. A. Lidar, and K. B. Whaley, Robustness of decoherence-free subspaces for quantum computation, *Phys. Rev. A* **60**, 1944 (1999).
- [34] P. G. Kwiat, A. J. Berglund, J. B. Altepeter, and A. G. White, Experimental verification of decoherence-free subspaces, *Science* **290**, 498 (2000).
- [35] J. M. Chow, L. DiCarlo, J. M. Gambetta, A. Nunnkamp, L. S. Bishop, L. Frunzio, M. H. Devoret, S. M. Girvin, and R. J. Schoelkopf, Detecting highly entangled states with a joint qubit readout, *Phys. Rev. A* **81**, 062325 (2010).
- [36] J. M. Chow, J. M. Gambetta, A. D. Córcoles, S. T. Merkel, J. A. Smolin, C. Rigetti, S. Poletto, G. A. Keefe, M. B. Rothwell, J. R. Rozen, M. B. Ketchen, and M. Steffen, Universal quantum gate set approaching fault-tolerant thresholds with superconducting qubits, *Phys. Rev. Lett.* **109**, 060501 (2012).
- [37] F. B. Maciejewski, F. Baccari, Z. Zimborás, and M. Oszmaniec, Modeling and mitigation of cross-talk effects in readout noise with applications to the quantum approximate optimization algorithm, *Quantum* **5**, 464 (2021).
- [38] S. Bravyi, S. Sheldon, A. Kandala, D. C. McKay, and J. M. Gambetta, Mitigating measurement errors in multiqubit experiments, *Phys. Rev. A* **103**, 042605 (2021).
- [39] O. Kern, G. Alber, and D. L. Shepelyansky, Quantum error correction of coherent errors by randomization, *The European Physical Journal D-Atomic, Molecular, Optical and Plasma Physics* **32**, 153 (2005).
- [40] J. J. Wallman and J. Emerson, Noise tailoring for scalable quantum computation via randomized compiling, *Phys. Rev. A* **94**, 052325 (2016).

- [41] M. Ware, G. Ribeill, D. Ristè, C. A. Ryan, B. Johnson, and M. P. da Silva, Experimental pauli-frame randomization on a superconducting qubit, *Phys. Rev. A* **103**, 042604 (2021).
- [42] D. Venturelli, M. Do, E. Rieffel, and J. Frank, Compiling quantum circuits to realistic hardware architectures using temporal planners, *Quantum Science and Technology* **3**, 025004 (2018).
- [43] A. Zulehner, A. Paler, and R. Wille, An efficient methodology for mapping quantum circuits to the ibm qx architectures, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **38**, 1226 (2018).
- [44] G. W. Dueck, A. Pathak, M. M. Rahman, A. Shukla, and A. Banerjee, Optimization of circuits for ibm's five-qubit quantum computers, in *2018 21st Euromicro Conference on Digital System Design (DSD)* (IEEE, 2018) pp. 680–684.
- [45] D. Venturelli, M. Do, B. O’Gorman, J. Frank, E. Rieffel, K. E. Booth, T. Nguyen, P. Narayan, and S. Nanda, Quantum circuit compilation: An emerging application for automated reasoning, in *Proc. Schedul. Plan. Appl. Workshop* (2019).
- [46] M. G. Davis, E. Smith, A. Tudor, K. Sen, I. Siddiqi, and C. Iancu, Towards optimal topology aware quantum circuit synthesis, in *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)* (IEEE, 2020) pp. 223–234.
- [47] B. Tan and J. Cong, Optimal layout synthesis for quantum computing, in *2020 IEEE/ACM International Conference On Computer Aided Design (ICCAD)* (IEEE, 2020) pp. 1–9.
- [48] K. Temme, S. Bravyi, and J. M. Gambetta, Error mitigation for short-depth quantum circuits, *Physical Review Letters* **119**, 10.1103/physrevlett.119.180509 (2017).
- [49] Y. Li and S. C. Benjamin, Efficient variational quantum simulator incorporating active error minimization, *Phys. Rev. X* **7**, 021050 (2017).
- [50] A. Kandala, K. Temme, A. D. Córcoles, A. Mezzacapo, J. M. Chow, and J. M. Gambetta, Error mitigation extends the computational reach of a noisy quantum processor, *Nature* **567**, 491 (2019).
- [51] T. Giurgica-Tiron, Y. Hindy, R. LaRose, A. Mari, and W. J. Zeng, Digital zero noise extrapolation for quantum error mitigation, *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)*, 306 (2020).
- [52] A. He, B. Nachman, W. A. de Jong, and C. W. Bauer, Zero-noise extrapolation for quantum-gate error mitigation with identity insertions, *Phys. Rev. A* **102**, 012426 (2020).
- [53] Y. Kim, C. J. Wood, T. J. Yoder, S. T. Merkel, J. M. Gambetta, K. Temme, and A. Kandala, Scalable error mitigation for noisy quantum circuits produces competitive expectation values, [arXiv:2108.09197 \[cond-mat, physics:quant-ph\]](https://arxiv.org/abs/2108.09197) (2021), [arXiv: 2108.09197](https://arxiv.org/abs/2108.09197).
- [54] J. Bylander, S. Gustavsson, F. Yan, F. Yoshihara, K. Harrabi, G. Fitch, D. G. Cory, Y. Nakamura, J.-S. Tsia, and W. D. Oliver, Noise spectroscopy through dynamical decoupling with a superconducting flux qubit, *Nature Physics* **7**, 656 (2019).
- [55] F. Yan, S. Gustavsson, J. Bylander, X. Jin, F. Yoshihara, D. G. Cory, Y. Nakamura, T. P. Orlando, and W. D. Oliver, Rotating-frame relaxation as a noise spectrum analyser of a superconducting qubit undergoing driven evolution, *Nature Communications* **4**, 22337 (2013).
- [56] J. Meeson, A. Ya. Tzalenchuk, and T. Lindström, Evidence for interacting two-level systems from the $1/f$ noise of a superconducting resonator, *Nature Communications* **5**, 4119 (2014).
- [57] C. Müller, J. Lisenfeld, A. Shnirman, and P. S., Non-gaussian noise spectroscopy with a superconducting qubit sensor, *Physical Review B* **92**, 035442 (2015).
- [58] J. J. Burnett, A. Bengtsson, M. Scigliuzzo, D. Niepce, M. Kudra, P. Delsing, and J. Bylander, Decoherence benchmarking of superconducting qubits, *npj Quantum Information* **5**, 54 (2019).
- [59] J. Basset, A. Stockklauser, D.-D. Jarausch, T. Frey, C. Reichl, W. Wegscheider, A. Wallraff, K. Ensslin, and I. T., Evaluating charge noise acting on semiconductor quantum dots in the circuit quantum electrodynamics architecture, *Appl. Phys. Lett.* **105**, 063105 (2014).
- [60] K. W. Chan, W. Huang, C. H. Yang, J. C. C. Hwang, B. Hensen, T. Tanttu, F. E. Hudson, K. M. Itoh, A. Laucht, A. Morello, and A. S. Dzurak, Assessment of a silicon quantum dot spin qubit environment via noise spectroscopy, *Phys. Rev. Appl.* **10**, 044017 (2018).
- [61] T. Struck, A. Hollmann, F. Schauer, O. Fedorets, A. Schmidbauer, K. Sawano, H. Riemann, N. V. Abrosimov, L. Cywiński, B. D., and L. R. Schreiber, Low-frequency spin qubit energy splitting noise in highly purified $^{28}\text{Si}/\text{SiGe}$, *npj Quantum Information* **6**, 40 (2020).
- [62] B. Pokharel, N. Anand, B. Fortman, and D. A. Lidar, Demonstration of fidelity improvement using dynamical decoupling with superconducting qubits, *Phys. Rev. Lett.* **121**, 220502 (2018).
- [63] S. Niu and A. Todri-Sanial, Pulse-level noise mitigation on quantum applications, *arXiv preprint arXiv:2204.01471* (2022).
- [64] A. Murphy, J. Epstein, G. Quiroz, K. Schultz, L. Tewala, K. McElroy, C. Trout, B. Tien-Street, J. A. Hoffmann, B. Clader, *et al.*, Universal dephasing noise injection via Schrodinger wave autoregressive moving average models, *arXiv preprint arXiv:2102.03370* (2021).
- [65] V. Tripathi, H. Chen, M. Khezri, K.-W. Yip, E. M. Levenson-Falk, and D. A. Lidar, Suppression of crosstalk in superconducting qubits using dynamical decoupling, *arXiv preprint 10.48550/arxiv.2108.04530* (2021).
- [66] H. Zhang, B. Pokharel, E. M. Levenson-Falk, and D. Lidar, Predicting non-Markovian superconducting qubit dynamics from tomographic reconstruction, *arXiv preprint 10.48550/arxiv.2111.07051* (2021).
- [67] G. A. Álvarez and D. Suter, Measuring the spectrum of colored noise by dynamical decoupling, *Phys. Rev. Lett.* **107**, 230501 (2011).
- [68] P. Szańkowski, G. Ramon, J. Krzywda, D. Kwiatkowski, and L. Cywiński, Environmental noise spectroscopy with qubits subjected to dynamical decoupling, *Journal of Physics: Condensed Matter* **29**, 333001 (2017).
- [69] G. A. Paz-Silva, L. M. Norris, and L. Viola, Multiqubit spectroscopy of gaussian quantum noise, *Phys. Rev. A* **95**, 022121 (2017).
- [70] L. Cywiński, R. M. Lutchyn, C. P. Nave, and S. Das Sarma, How to enhance dephasing time in superconducting qubits, *Phys. Rev. B* **77**, 174509 (2008).
- [71] G. A. Paz-Silva and L. Viola, General transfer-function approach to noise filtering in open-loop quantum control, *Phys. Rev. Lett.* **113**, 250501 (2014).

- [72] K. Schultz, G. Quiroz, P. Titum, and B. D. Clader, SchWARMA: A model-based approach for time-correlated noise in quantum circuits, *Phys. Rev. Research* **3**, 033229 (2021).
- [73] E. F. Dumitrescu, A. J. McCaskey, G. Hagen, G. R. Jansen, T. D. Morris, T. Papenbrock, R. C. Pooser, D. J. Dean, and P. Lougovski, Cloud quantum computing of an atomic nucleus, *Phys. Rev. Lett.* **120**, 210501 (2018).
- [74] R. LaRose, A. Mari, S. Kaiser, P. J. Karalekas, A. A. Alves, P. Czarnik, M. E. Mandouh, M. H. Gordon, Y. Hindy, A. Robertson, P. Thakre, N. Shammah, and W. J. Zeng, Mitiq: A software package for error mitigation on noisy quantum computers, arXiv preprint [10.48550/arxiv.2009.04417](https://arxiv.org/abs/10.48550/arxiv.2009.04417) (2020).
- [75] R. Kubo, Stochastic Liouville equations, *J. of Math. Phys.* **4**, 174 (1963).
- [76] H. Haken and G. Strobl, An exactly solvable model for coherent and incoherent exciton motion, *Z. Phys. A Hadrons Nuclei* **262**, 135 (1973).
- [77] V. Čápek, Haken—strobl—reineker model: its limits of validity and a possible extension, *Chem. Phys.* **171**, 79 (1993).
- [78] Y. Cheng and R. Silbey, Stochastic liouville equation approach for the effect of noise in quantum computations, *Phys. Rev. A* **69**, 052325 (2004).
- [79] C. Gardiner, *Handbook of stochastic methods (for physics, chemistry and the natural sciences)*, 3rd ed (Springer-Verlag, 2004).
- [80] Y. Cheng and R. Silbey, Microscopic quantum dynamics study on the noise threshold of fault-tolerant quantum error correction, *Phys. Rev. A* **72**, 012320 (2005).
- [81] N. G. Van Kampen, *Stochastic processes in physics and chemistry*, Vol. 1 (Elsevier, 1992).
- [82] P. Whittle, *Prediction and regulation by linear least-square methods* (English Universities Press, 1963).
- [83] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control* (John Wiley & Sons, 2015).
- [84] S. H. Holan, R. Lund, G. Davis, *et al.*, The ARMA alphabet soup: A tour of ARMA model variants, *Statistics Surveys* **4**, 232 (2010).
- [85] S. Endo, S. C. Benjamin, and Y. Li, Practical quantum error mitigation for near-future applications, *Phys. Rev. X* **8**, 031027 (2018).
- [86] A. Lowe, M. H. Gordon, P. Czarnik, A. Arrasmith, P. J. Coles, and L. Cincio, Unified approach to data-driven quantum error mitigation, *Phys. Rev. Research* **3**, 033098 (2021).
- [87] A. Mari, N. Shammah, and W. J. Zeng, Extending quantum probabilistic error cancellation by noise scaling, *Phys. Rev. A* **104**, 052607 (2021).
- [88] mezza: a toolox for simulating open quantum system dynamics, <https://github.com/mezza-team/mezza> (2020).
- [89] M. Broughton, G. Verdon, T. McCourt, A. J. Martinez, J. H. Yoo, S. V. Isakov, P. Massey, R. Halavati, M. Y. Niu, A. Zlokapa, E. Peters, O. Lockwood, A. Skolik, S. Jerbi, V. Dunjko, M. Leib, M. Streif, D. V. Dollen, H. Chen, S. Cao, R. Wiersema, H.-Y. Huang, J. R. McClean, R. Babbush, S. Boixo, D. Bacon, A. K. Ho, H. Neven, and M. Mohseni, Tensorflow quantum: A software framework for quantum machine learning (2021), [arXiv:2003.02989 \[quant-ph\]](https://arxiv.org/abs/2003.02989).
- [90] T. Proctor, K. Rudinger, K. Young, E. Nielsen, and R. Blume-Kohout, Measuring the capabilities of quantum computers, [arXiv:2008.11294 \[quant-ph\]](https://arxiv.org/abs/2008.11294) (2020), arXiv: 2008.11294.
- [91] H. Y. Carr and E. M. Purcell, Effects of diffusion on free precession in nuclear magnetic resonance experiments, *Phys. Rev.* **94**, 630 (1954).
- [92] S. Meiboom and D. Gill, Modified spin-echo method for measuring nuclear relaxation times, *Rev. Sci. Instrum.* **29**, 688 (1958).

Appendix A: Consistency between different theories of pulse-stretching

Our work is based on a semiclassical theory of time-correlated noise, according to which, the pulse-stretching technique induces two effective changes on the noise spectrum: (i) it scales the noise level by a constant λ , (ii) it also stretches the noise spectrum on the frequency axis by the same constant. Both effects are formally summarized in Eq. (11) derived in the main text.

In Ref. [48], a different formalism, based on a master equation with a time-independent noise operator, was used to study the pulse-stretching technique. More precisely, a system evolving according to the following master equation was considered:

$$\frac{\partial}{\partial t}\rho(t) = -[K(t), \rho(t)] + \mathcal{L}(\rho(t)), \quad (\text{A1})$$

where $K(t)$ is the system Hamiltonian and \mathcal{L} is a time-independent noise super-operator. As shown in Ref. [48], the effect of pulse stretching (i.e., $K(t) \rightarrow 1/\lambda K(t/\lambda)$) is equivalent to an effective master equation:

$$\frac{\partial}{\partial t'}\rho(t') = -[K(t'), \rho(t')] + \lambda\mathcal{L}(\rho(t')), \quad (\text{A2})$$

where $t' = \lambda t$. In practice pulse-stretching induces a multiplicative scaling of the noise operator $\mathcal{L} \rightarrow \lambda\mathcal{L}$.

The master equation Eq. (A1) is typically used to model Markovian noise (no time-correlations). In this case, the Hilbert space of the environment can be traced out such that ρ represents the reduced state of the system evolving according to the master equation Eq. (A1). In this white-noise regime, also our semiclassical theory of pulse-stretching predicts a simple multiplicative scaling of noise power and this is indeed consistent with Eq. (A2).

What happens for a non-Markovian environment with a colored noise spectrum? In this case, our semiclassical theory suggests that pulse-stretching induces, in addition to a multiplicative scaling, also a scaling of the frequency axis of the noise spectrum (see Eq. (11)). This may seem to contradict the simple multiplicative scaling of the noise $\mathcal{L} \rightarrow \lambda\mathcal{L}$ derived in Ref. [48] and reported in Eq. (A2). However, as explained below, both theoretical derivations are actually consistent with each other.

In principle, the master equation (A1) can be used to model a non-Markovian bath by representing with ρ the

global quantum state (system + bath) instead of the reduced state of the system. In this global picture, a non-Markovian bath can be modeled by a time-independent noise operator $\mathcal{L}(\rho)$ that includes an interaction Hamiltonian term H_{SB} and the bare Hamiltonian H_B acting on the bath only (see Supplemental Material of Ref. [48])

$$\mathcal{L}(\rho(t)) = -i[H_{SB} + H_B, \rho(t)], \quad (\text{A3})$$

which we can split as the sum of two terms $\mathcal{L} = \mathcal{L}_{SB} + \mathcal{L}_B$, where $\mathcal{L}_{SB}(\rho) = -i[H_{SB}, \rho]$ and $\mathcal{L}_B(\rho) = -i[H_B, \rho]$. In this case, the simple multiplicative scaling $\mathcal{L} \rightarrow \lambda\mathcal{L}$ induced by the pulse-stretching technique according to Eq. (A2) has actually two physically different effects: (i) $\mathcal{L}_{SB} \rightarrow \lambda\mathcal{L}_{SB}$ corresponding to a scaling of the noise power and (ii) $\mathcal{L}_B \rightarrow \lambda\mathcal{L}_B$ corresponding to an effective scaling of the all the characteristic frequencies of the bath and, therefore, to a frequency stretching of the noise spectrum. These two effects are consistent with the semiclassical theory of pulse-stretching presented in this work and, in particular, with Eq. (11).

Appendix B: Filter function formalism for quantum circuits

Consider an n -qubit system governed by a Hamiltonian

$$H(t) = H_0(t) + H_E(t), \quad (\text{B1})$$

where $H_0(t)$ defines a sequence of control operations applied to the quantum system and $H_E(t)$ defines the error Hamiltonian. We will assume piecewise-constant evolution such that $H_0(t) = \sum_i s_i(t)H_i$, where $s_i(t) = 1$ when $t \in [t_{i-1}, t_i)$ and $s_i(t) = 0$ otherwise. The resulting pure control evolution is

$$\begin{aligned} U_0(T, 0) &= \mathcal{T}_+ e^{-i \int_0^T H_0(s) ds} \\ &= U_K U_{K-1} \cdots U_1 \\ &= U_{K:1}, \end{aligned} \quad (\text{B2})$$

with $U_j = e^{-i(t_j - t_{j-1})H_j}$ and T designating the total circuit runtime. The last equality is defined in anticipation of the subsequent switching function calculations.

The error Hamiltonian can include anything from systematic control noise to interactions between the system and its environment. Here, we focus on semiclassical, spatiotemporally correlated noise: $H_E(t) = \sum_\alpha b_\alpha(t)A_\alpha$. The operators A_α are Hermitian, traceless and form an operator basis on the system Hilbert space with respect to the Hilbert-Schmidt norm, i.e., $\langle A_\alpha, A_\beta \rangle = 1/N \text{Tr}[A_\alpha A_\beta] = \delta_{\alpha,\beta}$, with $N = 2^n$ denoting the Hilbert space dimension. The noise couples to the system via $b_\alpha(t)$, which are defined as classical wide-sense stationary, Gaussian variables. Hence, the statistical properties of $b_\alpha(t)$ are characterized by the mean $\overline{b_\alpha(t)}$ and the two-point correlation functions $C_{\alpha\beta}(t_1 - t_2) = \overline{b_\alpha(t_1)b_\beta(t_2)}$.

Note that this model can be used to represent both additive control noise and interactions between the quantum system and a classical environment.

Assuming the weak noise limit, i.e., $\|H_E(t)\|T \ll 1$, we can examine the dynamics generated by $H(t)$ perturbatively by moving into an interaction picture with respect to $H_0(t)$. This is performed by representing the full dynamics operator $U(T) = \mathcal{T}_+ e^{-i \int_0^T H(s) ds}$ as $U(T) = \tilde{U}_E(T)U_0(T)$, where $\tilde{U}_E(T) = e^{-i \int_0^T \tilde{H}_E(s) ds}$ with

$$\begin{aligned} \tilde{H}_E(t) &= U_0(T, t)H_E(t)U_0^\dagger(T, t) \\ &= \sum_\alpha U_0(T, t)A_\alpha U_0^\dagger(T, t)B_\alpha(t) \\ &= \sum_{\alpha,\beta} f_{\alpha\beta}(t)A_\beta b_\alpha(t). \end{aligned} \quad (\text{B3})$$

Note that the rotated error Hamiltonian is expressed in terms of a ‘‘reverse’’ interaction picture with respect to the pure control evolution. The functions $f_{\alpha\beta}(t)$ are known as the switching functions and are defined by

$$f_{\alpha\beta}(t) = \frac{1}{N} \text{Tr} \left[U_0(T, t)A_\alpha U_0^\dagger(T, t)A_\beta \right]. \quad (\text{B4})$$

As we will see, the switching functions are an integral part of the filter function formalism.

The dynamics of an observable with respect to $U(T)$ can be written as

$$\begin{aligned} \overline{\langle O \rangle} &= \overline{\text{Tr}[\rho_S(T)O]} \\ &= \overline{\text{Tr}[\tilde{U}_E(T)U_0(T)\rho_S(0)U_0^\dagger(T)\tilde{U}_E^\dagger(T)O]} \\ &= \text{Tr}[\Lambda(T)\rho_{S,0}(T)O], \end{aligned} \quad (\text{B5})$$

where the last equality conveniently illustrates the utility of the particular rotated frame chosen above. The term $\Lambda(T) = \overline{O^{-1}\tilde{U}_E(T)O\tilde{U}_E^\dagger(T)}$ constitutes the error operator, while $\rho_{S,0}(T) = U_0(T)\rho_S(0)U_0^\dagger(T)$ defines the time-evolved state resulting from the noiseless circuit implementation. The error operator can be expressed as a cumulant expansion $\Lambda(T) = \exp\left(\sum_{n=1}^{\infty} (-i)^n \mathcal{C}_O^{(n)}(T)/n!\right)$ that can be truncated to second order if the noise is sufficiently weak and the time is sufficiently short [69].

We will focus on zero-mean noise, and thus, $\overline{b_\alpha(t)} = 0$ and $\mathcal{C}_O^{(1)}(T) = 0$. As a result, the first non-zero term is $\mathcal{C}_O^{(2)}(T)$, which can be written as

$$\frac{\mathcal{C}_O^{(2)}(T)}{2} = \sum_{\alpha,\beta,\alpha',\beta'} \int_0^\infty \frac{d\omega}{2\pi} S_{\alpha,\alpha'}(\omega) \mathcal{F}_{\alpha\beta,\alpha'\beta'}(\omega, T) \mathcal{A}_{\beta\beta'}, \quad (\text{B6})$$

The noise power spectral density $S_{\alpha\alpha'}(\omega)$ is defined via

$$C_{\alpha\alpha'}(t_1 - t_2) = \frac{1}{2\pi} \int_{-\infty}^\infty S_{\alpha\alpha'}(\omega) e^{i\omega(t_1 - t_2)} d\omega, \quad (\text{B7})$$

while the filter functions are

$$\mathcal{F}_{\alpha\beta,\alpha'\beta'}(\omega, T) = \text{Re} [F_{\alpha\beta}(\omega, T)F_{\alpha'\beta'}(-\omega, T)]. \quad (\text{B8})$$

with $F_{\alpha\beta}(\omega, T) = \int_0^T dt f_{\alpha\beta}(t)e^{i\omega t}$. The operator $\mathcal{A}_{\beta\beta'}$ is defined as

$$\begin{aligned} \mathcal{A}_{\beta\beta'} &= A_\beta A_{\beta'} - O^{-1} A_\beta O A_{\beta'} - A_\beta O^{-1} A_{\beta'} O \\ &\quad + O^{-1} A_\beta A_{\beta'} O \end{aligned} \quad (\text{B9})$$

We will now use this representation to derive analytical expressions for the filter functions resulting from various unitary folding techniques.

Appendix C: Filter function perspective on global folding

Consider a global folding protocol defined as $U_{GF}[(2M+1)T, 0] = U_0(T) [U_0^\dagger(T)U_0(T)]^M$, where $U_0(t)$ is the unitary representing the desired quantum algorithm and M is the number of folding repetitions to be performed. The total time required to implement the algorithm is denoted by T .

1. Switching functions

The switching functions resulting from the global folding protocol are given by

$$f_{\alpha\beta}(t) = \begin{cases} f_{\alpha\beta}^{(1)}(t) & : t \in [0, 2MT) \\ f_{\alpha\beta}^{(2)}(t) & : t \in [2MT, (2M+1)T) \end{cases} \quad (\text{C1})$$

We partition them into two terms: the first, denoted with superscript (1), is defined during the global folding as

$$f_{\alpha\beta}^{(1)}(t) = \frac{1}{N} \text{Tr} \left[U_0(T) \Gamma(2MT, t) A_\alpha \Gamma^\dagger(2MT, t) U_0^\dagger(T) A_\beta \right]. \quad (\text{C2})$$

The operator $\Gamma(2MT, t)$ captures the partial (reverse) unitary dynamics of $[U_0^\dagger(T)U_0(T)]^M$. By again expanding into the operator basis of A_α , we define an additional

switching function

$$y_{\alpha\gamma}(t) = \frac{1}{N} \text{Tr} \left[\Gamma(2MT, t) A_\alpha \Gamma^\dagger(2MT, t) A_\gamma \right], \quad (\text{C3})$$

such that

$$\Gamma(2MT, t) A_\alpha \Gamma^\dagger(2MT, t) = \sum_\gamma y_{\alpha\gamma}(t) A_\gamma. \quad (\text{C4})$$

As a result, we can express $f_{\alpha\beta}^{(1)}(t)$ as

$$f_{\alpha\beta}^{(1)}(t) = \frac{1}{N} \sum_\gamma y_{\alpha\gamma}(t) \text{Tr} \left[U_0(T) A_\gamma U_0^\dagger(T) A_\beta \right]. \quad (\text{C5})$$

The advantage of this representation is that we have now defined the switching function $f_{\alpha\beta}^{(1)}(t)$ in terms of a mirror symmetric switching function $y_{\alpha\gamma}(t)$. Mirror symmetric switching functions satisfy the property

$$y_{\alpha\gamma}(t) = y_{\alpha\gamma}(2T - t), \quad (\text{C6})$$

which we will find useful when examining the filter functions of global folding. Lastly, the second term of the switching function $f_{\alpha\beta}(t)$ is defined after the global folding and given by

$$\begin{aligned} f_{\alpha\beta}^{(2)}(t) &= \frac{1}{N} \sum_j g_j(t) \times \\ &\quad \text{Tr} \left[U_{K:j+1} e^{-i(t_j-t)H_j} A_\alpha e^{i(t_j-t)H_j} U_{K:j+1}^\dagger A_\beta \right], \end{aligned} \quad (\text{C7})$$

with $g_j(t) = \Theta(t - t_{j-1}) - \Theta(t - t_j)$, and $\Theta(t)$ denoting the Heaviside function.

2. Filter functions

The filter functions are defined via products of Fourier transforms of switching functions. Since the switching functions can be partitioned into two terms, the filter functions can be partitioned into four terms:

$$F_{\alpha\beta, \alpha'\beta'}(\omega, T) = \sum_{i,j=1}^2 G_{\alpha\beta, \alpha'\beta'}^{(i,j)}(\omega, T) \quad (\text{C8})$$

where

$$G_{\alpha\beta, \alpha'\beta'}^{(1,1)}(\omega, T) = \frac{\sin^2(M\omega T)}{\sin^2(\omega T)} F_{\alpha\beta}^{(1)}(-\omega, T) F_{\alpha'\beta'}^{(1)}(\omega, T) \quad (\text{C9})$$

$$G_{\alpha\beta, \alpha'\beta'}^{(1,2)}(\omega, T) = e^{i(3M+1)\omega T} \frac{\sin(M\omega T)}{\sin(\omega T)} F_{\alpha\beta}^{(1)}(-\omega, T) F_{\alpha'\beta'}^{(2)}(-\omega, T) \quad (\text{C10})$$

$$G_{\alpha\beta, \alpha'\beta'}^{(2,1)}(\omega, T) = [G_{\alpha\beta, \alpha'\beta'}^{(1,2)}(\omega, T)]^* \quad (\text{C11})$$

$$G_{\alpha\beta, \alpha'\beta'}^{(2,2)}(\omega, T) = F_{\alpha\beta}^{(2)}(\omega, T) F_{\alpha'\beta'}^{(2)}(-\omega, T). \quad (\text{C12})$$

The component filter functions are determined by

$$F_{\alpha\beta}^{(1)}(\omega, T) = \int_0^{2T} dt f_{\alpha\beta}^{(1)}(t) e^{i\omega t}, \quad (\text{C13})$$

$$F_{\alpha\beta}^{(2)}(\omega, T) = \int_0^T dt f_{\alpha\beta}^{(2)}(t) e^{i\omega t}. \quad (\text{C14})$$

Note that $G_{\alpha\beta, \alpha'\beta'}^{(1,1)}(\omega, T)$ and $G_{\alpha\beta, \alpha'\beta'}^{(1,2)}(\omega, T)$ exhibit “comb-like” behavior conveyed by the presence of the quotient of sinusoidal functions. These factors appear from the M repetitions of global folding, and they are responsible for the more distinct features in the filter function as M increases.

Appendix D: Filter function perspective on local folding

Local folding is generically described by the total unitary

$$\begin{aligned} U_{LF}(T, 0) &= U_K \left(U_K^\dagger U_K \right)^M \cdots U_1 \left(U_1^\dagger U_1 \right)^M \\ &= U_K \Gamma_K(T_K) \cdots U_1 \Gamma_1(T_1), \end{aligned} \quad (\text{D1})$$

where each gate is subject to a folding interval. Each local folding is equivalent in repetition, occurring M total times for each of the gates U_j . It is assumed that each gate takes an equivalent amount of time τ , and therefore the timing of each gate is given by

$$t_k^{(j)} = [(2M+1)(j-1) + k]\tau, \quad (\text{D2})$$

for $j = 1, \dots, K$ and $k = 1, \dots, 2M+1$. We define the local folding unitary $\Gamma_j(T_j) = (U_j^\dagger U_j)^M$, where the total time $T_j = t_{2M}^{(j)}$, for convenience and in anticipation of the subsequent calculations. At intermediate times, the folding operator is given by

$$\Gamma_j(T_j, t) = \begin{cases} (U_j^\dagger U_j)^{M-k} e^{i(t_k^{(k)} - t)H_j} & : k \text{ odd} \\ (U_j^\dagger U_j)^{M-k} U_j^\dagger e^{-i(t_k^{(j)} - t)H_j} & : k \text{ even} \end{cases}. \quad (\text{D3})$$

Note that we have defined the local folding operator with respect to the total time as this will naturally appear from the “reverse” propagator $U_0(T, t)$; see Eq. (B3). Furthermore, note that for brevity we will use the notation $\Gamma_j(t) = \Gamma_j(T_j, t)$.

1. Switching functions

While local folding utilizes a more complex folding procedure than global folding, the switching functions can still be partitioned into folding and post-folding

terms. Specifically, the folding terms reside in the domain $t \in [t_{2M+1}^{(j-1)}, t_{2M}^{(j)})$ and the post-folding terms can be defined within $t \in [t_{2M}^{(j)}, t_{2M+1}^{(j)})$. This construction allows for the switching function to be expressed as $f_{\alpha\beta}(t) = f_{\alpha\beta}^{(1)}(t) + f_{\alpha\beta}^{(2)}(t)$, where

$$f_{\alpha\beta}^{(\mu)}(t) = \sum_j g_j^{(\mu)}(t) f_{\alpha\beta}^{(\mu,j)}(t). \quad (\text{D4})$$

The indices $\mu = 1, 2$ characterize the folding and post-folding terms, respectively. The functions $g_j^{(1)}(t) = \Theta[t - t_{2M+1}^{(j-1)}] - \Theta[t - t_{2M}^{(j)}]$ and $g_j^{(2)}(t) = \Theta[t - t_{2M}^{(j)}] - \Theta[t - t_{2M+1}^{(j)}]$ capture the piecewise features of the gate folding periods and their final implementation.

During a local folding period, the switching function is defined by

$$f_{\alpha\beta}^{(1,j)}(t) = \frac{1}{N} \sum_\gamma y_{\alpha\gamma}^{(j)}(t) \text{Tr} \left[Q_{K:j} \Gamma_j^\dagger(T_j) A_\gamma \Gamma_j(T_j) Q_{K:j}^\dagger A_\beta \right], \quad (\text{D5})$$

where $Q_{k:j} = U_k \Gamma_k(T_k) \cdots U_j \Gamma_j(T_j)$. As in the global folding case, we introduce an additional switching function

$$y_{\alpha\gamma}^{(j)}(t) = \frac{1}{N} \text{Tr}[\Gamma_j(t) A_\alpha \Gamma_j^\dagger(t) A_\beta] \quad (\text{D6})$$

which satisfies the mirror symmetric condition described in Eq. (C6) for the j th local folding interval. Once again, this property will prove useful during the calculation of the filter functions.

Local folding operations are followed by the implementation of the folded gate. The switching function describing this post-folding period is

$$f_{\alpha\beta}^{(2,j)}(t) = \frac{1}{N} \text{Tr} \left[Q_{K:j+1} e^{-i(t_{2M+1}^{(j)} - t)H_j} A_\alpha \times Q_{K:j+1}^\dagger e^{i(t_{2M+1}^{(j)} - t)H_j} A_\beta \right]. \quad (\text{D7})$$

Partial time evolution during $t \in [t, t_{2M+1}^{(j)})$ captures the switching function dynamics during the gate. Note that this contribution disappears if the gates are assumed to be instantaneous.

2. Filter functions

There are four primary filter function that characterize the self-interference and cross-interference between local folding and post-folding intervals. While this is a similar construction to the global folding case in that the filter function $F_{\alpha\beta, \alpha'\beta'}^{(i,j)}(\omega, \tau)$ is equivalent to Eq. (C8) with $T = \tau$, the distinction lies within the definition of $G_{\alpha\beta, \alpha'\beta'}^{(i,j)}(\omega, \tau)$. Each of the four terms are composed of K^2 terms, where each constituent term captures the interactions between the j th and k th interval. More concretely, the $G_{\alpha\beta, \alpha'\beta'}^{(i,j)}(\omega, \tau)$ filter functions are given by

$$G_{\alpha\beta,\alpha'\beta'}^{(1,1)}(\omega, \tau) = \frac{\sin^2(M\omega\tau)}{\sin^2(\omega\tau)} \sum_{j,k=1}^K e^{i\omega[t_{2M+1}^{(k-1)} - t_{2M+1}^{(j-1)}]} F_{\alpha\beta}^{(1,j)}(-\omega, \tau) F_{\alpha'\beta'}^{(1,k)}(\omega, \tau) \quad (\text{D8})$$

$$G_{\alpha\beta,\alpha'\beta'}^{(1,2)}(\omega, \tau) = e^{i(M+1)\omega\tau} \frac{\sin(M\omega\tau)}{\sin(\omega\tau)} \sum_{j,k=1}^K e^{i\omega[t_{2M}^{(k)} - t_{2M+1}^{(j-1)}]} F_{\alpha\beta}^{(1,j)}(-\omega, \tau) F_{\alpha'\beta'}^{(2,k)}(-\omega, \tau) \quad (\text{D9})$$

$$G_{\alpha\beta,\alpha'\beta'}^{(2,1)}(\omega, \tau) = [G_{\alpha\beta,\alpha'\beta'}^{(1,2)}(\omega, \tau)]^* \quad (\text{D10})$$

$$G_{\alpha\beta,\alpha'\beta'}^{(2,2)}(\omega, \tau) = \sum_{j,k=1}^K e^{i\omega[t_{2M}^{(k)} - t_{2M}^{(j)}]} F_{\alpha\beta}^{(2,j)}(\omega, \tau) F_{\alpha'\beta'}^{(2,k)}(-\omega, \tau). \quad (\text{D11})$$

Each component filter function

$$F_{\alpha\beta}^{(1,j)}(\omega, \tau) = \int_0^{2\tau} dt f_{\alpha\beta}^{(1,j)}(t) e^{i\omega t}, \quad (\text{D12})$$

$$F_{\alpha\beta}^{(2,j)}(\omega, \tau) = \int_0^{\tau} dt f_{\alpha\beta}^{(2,j)}(t) e^{i\omega t}. \quad (\text{D13})$$

denotes the Fourier Transform of the folding and post-folding switching functions, respectively. As in the global folding case, the pure folding filter function $G_{\alpha\beta,\alpha'\beta'}^{(1,1)}(\omega, \tau)$ is proportional to a quotient of sinusoidal functions; thus, exhibiting comb-like behavior with increasing folding repetition M . Despite this similarity, and many others, local folding produces a very distinct folding filter function.

In particular, local folding leads to low and high frequency localization in the filter function as M grows. This behavior can be attributed to the local folding periods, with the contribution of the self-interference term $G_{\alpha\beta,\alpha'\beta'}^{(1,1)}(\omega, \tau)$ being most influential. This term captures two distinct types of interactions between the j th and k th folding periods that are dependent upon the modulation properties of the switching functions. Products of relatively static switching functions will lead to component filter functions that have greatest support at low frequencies, while those that demonstrate rapid fluctuations will produce filter functions that tend towards high frequencies. This effect is exacerbated by the sinusoidal (frequency comb-like) expression as the number of folding repetitions increases.