Characterization of overparametrization in the simulation of realistic quantum systems

Matthew Duschenes [1,2,3,4,*] Juan Carrasquilla,[1,3,5] and Raymond Laflamme[1,2,4]

[1]*Department of Physics & Astronomy, University of Waterloo, Ontario, Canada N2L 3G1*
[2]*Institute for Quantum Computing, University of Waterloo, Ontario, Canada N2L 3G1*
[3]*Vector Institute, Toronto, Ontario, Canada M5G 1M1*
[4]*Perimeter Institute for Theoretical Physics, Waterloo, Ontario, Canada N2L 2Y5*
[5]*Institute for Theoretical Physics, ETH Zürich, 8093, Switzerland*

Quantum computing devices require exceptional control of their experimental parameters to prepare quantum states and simulate other quantum systems. Classical optimization procedures used to find such optimal control parameters have further been shown in idealized settings to exhibit different regimes of learning. Of interest in this work is the overparametrization regime, where for systems with a sufficient number of parameters, global optima for prepared state and compiled unitary fidelities may potentially be reached exponentially quickly. Here, we study the robustness of overparametrization phenomena in the presence of experimental constraints on the controls, such as bounding or sharing parameters across operators, as well as in the presence of noise inherent to experimental setups. We observe that overparametrization phenomena are resilient in these realistic settings at short times, however fidelities decay to zero past a critical simulation duration due to accumulation of either quantum or classical noise. This critical depth is found to be logarithmic in the scale of noise, and optimal fidelities initially increase exponentially with depth, before decreasing polynomially with depth, and with noise. Our results demonstrate that parametrized *Ansätze* can mitigate entropic effects from their environment, offering tantalizing opportunities for their application and experimental realization in near-term quantum devices.

I. INTRODUCTION

There exist many active avenues and experimental platforms for the quantum information community to explore and advance quantum technologies, including trapped ions [1,2], superconducting qubits [3–5], neutral atoms [6–9], nuclear magnetic resonance [10,11], and several other intriguing approaches. To harness these technologies' full potential for tasks of interest in quantum information [12–14], in particular state preparation or unitary compilation [15], it is imperative to exercise precise control through the manipulation of experimental parameters. This complex, high-dimensional quantum control problem arises in numerous applications [16–20] and is addressed through classical simulation and parameter optimization. Insight into such procedures, in experimentally relevant quantum settings, is therefore necessary to further our ability to rigorously control such systems.

Quantum control shares striking similarities with the field of classical deep learning, where large parametrized models are used to discover and represent complex patterns in large amounts of data. Beyond the resemblance of being variational algorithms concerned with optimizing high-dimensional systems for technological advancement, a series of observations [21–23] has led to direct equivalencies in learning phenomena between variational quantum algorithms and deep learning. A striking example is overparametrization and lazy training

[24,25]. In classical systems, excessively parametrized models can learn efficiently with negligible adjustments to their parameters, leading to improved generalization performance and training efficiency. A similar phenomenon has been anticipated in ideal settings for noise-free variational quantum algorithms [26–28]. It is observed that in the overparametrized regime, the optimization landscape becomes almost free of suboptimal minima, and optimization may converge exponentially quickly.

Here we explore overparametrization via the classical simulation of quantum systems, within experimentally feasible settings. We investigate these phenomena within the quantum optimal control paradigm, where systems evolve under continuous time evolution [10,16,19], and within the variational quantum algorithms paradigm, where systems evolve under discretized sets of operations [15,21,29].

We find that overparametrization phenomena are robust under realistic settings, including constrained parametrizations, and imposing noisy nonunitary *Ansatz*. For a given periodic *Ansatz*, where quantum circuit depth dictates the number of model parameters, we find that inclusion of parameter constraints shifts the overparametrization depth boundary by a system-size-dependent factor. However, the dominant overparametrization phenomenon, of exponential convergence of optimization routines with depth, persists. In noisy settings, we observe that there are different regimes of optimization convergence. For depths beyond the overparametrization depth, but before a noise-induced critical depth, exponential convergence of optimization routines with depth still occurs.

*Contact author: mduschen@uwaterloo.ca

However, beyond this critical depth, an excessive amount of noise accumulates, and the optimization diverges polynomially with depth, and with noise. To complement these numerical findings, analytical investigations into the noise and depth scaling of the infidelity objectives, and other metrics including the entropy and purity of the parametrized states, provide an explanation for these behaviors. Overall, these findings suggest overparametrization is resilient when imposing experimental feasibility, offering opportunities for this phenomenon to be exploited in future simulated and existing quantum experiments.

The work is structured as follows. In Sec. II, we define general parametrized quantum channels and objective tasks of interest, namely noisy infidelities for state preparation and unitary compilation. In Sec. III, we interpret the form of noisy parametrized channels as expectation values of k-error channels, and we perform an analysis on the scaling of infidelity objectives with respect to noise and depth. In Sec. IV, we demonstrate overparametrization and other learning phenomena in constrained and noisy parameterized quantum systems. From numerical studies, we quantify the relationships between noise and depth at optimality. Finally in Secs. V and VI, we discuss the implications of these results and the resulting compromises that occur between numerical and experimental feasibility.

II. BACKGROUND

This work aims to understand the abilities of parametrized quantum systems in realistic, experimentally feasible settings. Critically, noise, resulting from systems interacting with their environment, is well known to be detrimental to quantum computation [15,30,31]. Example effects include noise-induced symmetry breaking [32], fundamental differences in sampling and annealing trajectories [33], and noise-induced barren plateaus [34]. Initial numerical investigations by Fontana *et al.* [35] demonstrate that noise, represented by noise scales or probabilities γ, accumulates with depth M. There are also well-known compromises between expressiveness, i.e., how much of the desired space of solutions can be represented by an *Ansatz* via increased depth [36], and trainability, i.e., the ability of an *Ansatz* to be optimized. However, these precise relationships in noisy settings have yet to be confirmed quantitatively.

The parametrized systems of interest simulated in this work are representative of various noisy intermediate scale quantum devices (NISQ) [3,5,9]. These systems consist of N qubits, each with local space dimension $D = 2$ and total space dimension $d = D^N$. These qubits may have fixed interqubit couplings, however they can be manipulated with external, time-dependent fields over a time T, or equivalently depth of simulation M. Please refer to Appendix A for a complete description of the parametrized *Ansätze* studied in this work. Such experimental parameters are generally constrained due to feasibility [20,30], and this work seeks to quantify the amount with which constraints affect the capabilities of parametrized quantum systems. We assume there are generally on the order of $P = O(\text{poly}(N)M)$ variable parameters in the system, where generally the system size N dependence is held fixed. Thus changes in parameter counts are reflected in

the simulation depth M, and any notion of overparametrization is discussed in the context of depth. We take as an example in this work nuclear magnetic resonance systems where nuclei, acting as $D = 2$ level qubits, are manipulated by time-dependent magnetic pulses [11,20,37,38].

Underlying this analysis are principles from learning theory based on studies of overparametrization in ideal quantum settings, including unitary compilation [26], variational quantum eigensolvers [27,39], and general quantum circuits [28]. These works have subsequently been followed up by initial theoretical studies on the effects of noise. Within an information theoretic context, the quantum Fisher information [40] is used as a metric to determine whether a quantum system is overparametrized. Within a general optimization context to complement neural-tangent kernel approaches for asymptotic learning dynamics [22], Riemannian gradient flow dynamics [39] are used to assert the convergence of overparametrized systems with bounded gradient noise. For our purposes, overparametrization refers to when there are an adequate number of parameters $P > \tilde{P}$ or depths $M > \tilde{M}$, such that the full space of G-dimensional solutions to tasks is spanned by the parametrized *Ansatz*. Within this regime, we investigate the resulting overparametrization phenomena of the possible exponential convergence of noisy optimization procedures with depth. Imposing constraints on the optimization related to experimental feasibility, including restricting individual control of qubits, is also known to decrease convexity in the objective landscape [13,41,42], and it requires many optimization heuristics. Please refer to Appendix B for a more complete description of overparametrization, including quantitative bounds on the overparametrization depth related to the quantum Fisher information.

Based on these studies, we hypothesize that there exists a critical evolution time or depth $M_\gamma > \tilde{M}$ where overparametrization has occurred, however too much noise has also accumulated. This noise is expected to prevent parametrized systems from accomplishing fidelity-based tasks with arbitrary precision. To confirm these predictions, we will consider the average behavior of infidelities, optimized independently over a distribution of tasks. We therefore conjecture that there are convergent and divergent phases of the optimization,

$$\text{Optimization} \sim \begin{cases} \text{Plateau}, & M < \tilde{M}, \\ \text{Convergent}, & \tilde{M} < M < M_\gamma, \\ \text{Divergent}, & M > M_\gamma. \end{cases} \quad (1)$$

This work aims to confirm these conjectures of depth-dependent regimes of learning phenomenon in nonideal settings.

To quantify the effects of noise on the evolution and abilities of parametrized quantum systems to perform tasks of interest, we define parametrized quantum channels as

$$\Lambda_{\theta\gamma} = \mathcal{N}_\gamma \circ \mathcal{U}_\theta, \quad (2)$$

with a unitary channel $\mathcal{U}_\theta$ parametrized by variable parameters θ, and a nonunitary channel $\mathcal{N}_\gamma$ parametrized by constant noise scales γ.

For the unitary channel, we assume the Hamiltonian driving the evolution

$$H_\theta^{(t)} = \sum_\mu H_\mu^{(t)} : \ H_\mu^{(t)} = \theta_\mu^{(t)} G_\mu \tag{3}$$

at a continuous time $t \in [0, T]$. The Hamiltonian is defined by a set of generators $\{G_\mu\}$, which are generally assumed to be acting on at most k qubits. Assuming the evolution is approximately piecewise constant over M time steps $\tau = T/M$ allows for first-order temporal Trotterization of the resulting unitary operator

$$U_\theta = \mathcal{T} e^{-i \int_0^T dt \, H_\theta^{(t)}} = \prod_m^M U_\theta^{(m)} + O(\tau^2). \tag{4}$$

Q order spatial Trotterization of the operator across the N qubits at time index m is also possible. Here Trotterization is represented as a product of a function of lower-order Trotterizations, denoted by the Q superscript,

$$U_\theta^{(m)} = e^{-i\tau H_\theta^{(m)}} = \prod_\mu^{(Q)} U_\mu^{(m,Q)} + O(\tau^{Q+1}). \tag{5}$$

The final first-order temporally localized unitary channel, with Q-order spatial Trotterization, is

$$\mathcal{U}_\theta = \circ_m^M \mathcal{U}_\theta^{(m,Q)} : \mathcal{U}_\theta^{(m,Q)} = \circ_\mu \mathcal{U}_\mu^{(m,Q)}, \tag{6}$$

with resulting gate operators related to the Hamiltonian generators

$$U_\mu^{(m,Q)} = e^{-i\tau H_\mu^{(m,Q)}}. \tag{7}$$

Please refer to Appendix A for a complete description of these schemes.

For the nonunitary channel, we consider temporally and spatially local independent noise acting on the $K = NM$ sites (m, i) of possible errors

$$\mathcal{N}_\gamma = \circ_m^M \left(\circ_i^N \mathcal{N}_{\gamma_i}^{(m)} \right). \tag{8}$$

For our purposes, we decompose each local noise channel into a convex combination of an identity component, and what we refer to as a nonidentity error component,

$$\mathcal{N}_{\gamma_i}^{(m)} \equiv (1 - \gamma)\mathcal{I}_i + \gamma \mathcal{K}_{\gamma_i}^{(m)}. \tag{9}$$

The forms of the nontrivial $\mathcal{K}_\gamma$ depend on the specific noise model of interest. Noise models considered in this work include local dephasing, amplitude damping, and depolarization noise. Such local noise models are known to be relevant in several quantum computing implementations [43,44]. Our analytical and numerical approaches are easily transferable to spatially correlated noise models across multiple qubits, however such studies concerning any nontrivial effect of correlated noise are left for future work. Please refer to Appendix C for a complete description of each noise channel, and to Appendix D for an analytical treatment of noise-induced effects. Given the temporal and spatial locality of the Trotterized unitary and nonunitary channel, we finally reach the explicit *Ansatz* form of interlaced noise and unitary evolution,

$$\Lambda_{\theta\gamma} = \circ_m^M \left(\mathcal{N}_\gamma^{(m)} \circ \mathcal{U}_\theta^{(m)} \right), \tag{10}$$

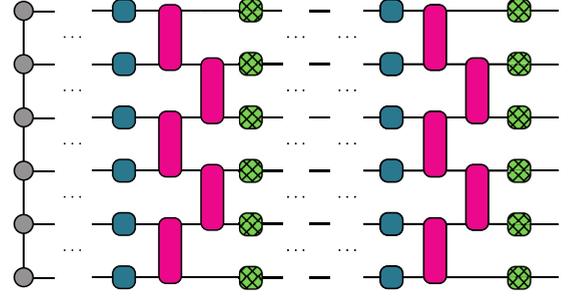leading to our overall channel circuit diagram in Fig. 1.

FIG. 1. Parametrized quantum channel with layers of Trotterized local (blue squares) and two-body (pink rectangles) unitary operators, followed by local noise channels (hatched green squares) after each layer, with an initial state (leftmost gray circles).

In this work, we take as our parametrized unitary *Ansatz* evolution generated by the nuclear magnetic resonance (NMR) Hamiltonian consisting of N, $D = 2$ qubit Pauli operators

$$H_\theta^{(t)} = \sum_i \theta_i^{x(t)} X_i + \sum_i \theta_i^{y(t)} Y_i$$
$$+ \sum_i h_i Z_i + \sum_{i<j} J_{ij} Z_i Z_j. \tag{11}$$

Generally in such systems, we have control over the variable time-dependent local transverse X and Y fields at qubit i, with additional constant time-independent longitudinal local Z at qubit i and nonlocal ZZ fields at qubits $i \neq j$. Such Hamiltonians allow for universal control over qubits, with the local and nonlocal gates allowing entangling gates to be implemented. We use experimentally relevant scales for our *Ansätze* in Table I, and details of the *Ansatz* are discussed in Appendix E. To compare our choice of NMR *Ansatz* to other implementations, we collect rough estimates [45] from recent literature of the 1-qubit T_{U1} and 2-qubit T_{U2} gate times, and decoherence times T_γ, for NMR [10,17,46], trapped ion [1,2,46], superconducting qubit [3,4,46], and neutral atom [6–8] quantum computing experiments in Table III of Appendix E. NMR systems are limited by experimental feasibility. Their small nonlocal coupling constants J limit how much qubits can be correlated at each time step $T_{U1} = O(\tau)$, increasing significantly the necessary 2-qubit gate times $T_{U2} = O(1/J)$. This translates into NMR systems having the

TABLE I. Experimentally relevant constants for constrained NMR *Ansatz*.

N	Number of qubits $\sim 1 - 4$
M	Number of time steps $\sim O(10^0 - 10^4)$
τ	Trotterization time step $\sim O(75 - 100\,\mu s)$
T	Evolution time $= M\tau \sim O(375\,\mu s - 500\,ms)$
Q	Spatial Trotterization order $= 2$
P	Number of parameters $\sim O(poly(N)M)$
J	Constant longitudinal coupling $\sim O(\pi/2 \times 10^2\,Hz)$
h	Constant longitudinal field $\sim O(\pi/2\,kHz)$
θ	Variable transverse field $\sim O(\pi/2\,MHz)$
γ	Noise scale $\sim O(10^{-14} - 10^{-1})$

largest effective depth $T_{U2}/T_{U1} \sim O(100)$ required for each 2-qubit gate, and having the smallest effective maximum depth $T_\gamma/T_{U2} \sim O(10^2)$ before coherence, among considered implementations. We therefore note that any conclusions drawn from this work regarding the explicit scale of noise or depth of models where phenomena occur are specific to this NMR *Ansatz*. However, we believe other similarly universal *Ansätze* should exhibit comparable behavior, at their specific *Ansatz*-dependent scales.

The choice of a universal *Ansatz* spanning the full space of $G = O(d^2)$ unitaries also simplifies transferability to other universal *Ansätze*, and avoids any bias by restricting evolution to being within a subspace. Techniques used other implementations [3] that also generally have origins in NMR techniques and their proof of principle quantum algorithm experiments [18,38], including zero-noise extrapolation [44], dynamical decoupling [47], and refocusing procedures [38]. These approaches are highly relevant to interpretations of the phenomena observed.

Given this parametrized *Ansatz*, we wish to assess its ability to represent targets of interest, such as operator compilation, where sequences of operators are optimized to approach a target operator, and state preparation, where initial states transformed by sequences of operators are optimized to approach a target state [48]. Such tasks arise often in quantum algorithms, and they depend crucially on targets being within the span of the *Ansatz*. Here we focus on unitary compilation and pure state preparation, given a parametrized *Ansatz* with universal control over the full space of unitary operators. The parameters θ of the *Ansätze* are optimized via optimization routines, and given an initial pure state σ, and target unitaries U or target pure states ρ, we assess the abilities of the respective parametrized unitaries and states,

$$U_{\theta\gamma} \approx U, \tag{12}$$

$$\rho_{\theta\gamma} = \Lambda_{\theta\gamma}(\sigma) \approx \rho = \mathcal{U}(\sigma). \tag{13}$$

Objective metrics of infidelities with respect to the given unitary compilation and state preparation tasks [49] are chosen to quantify these *Ansätze* through optimization,

$$\mathcal{L}_{\theta\gamma}^U = 1 - (1/d^2)|\mathrm{tr}(U^\dagger U_{\theta\gamma})|^2, \tag{14}$$

$$\mathcal{L}_{\theta\gamma}^\rho = 1 - \mathrm{tr}(\rho\rho_{\theta\gamma}). \tag{15}$$

We also define the impurity, von-Neumann entropy, and relative entropy divergence, relative to a pure state ρ as

$$\mathcal{I}_{\theta\gamma} = 1 - \mathrm{tr}(\rho_{\theta\gamma}^2), \tag{16}$$

$$\mathcal{S}_{\theta\gamma} = -\mathrm{tr}(\rho_{\theta\gamma} \ln(\rho_{\theta\gamma}))/\ln(d), \tag{17}$$

$$\mathcal{D}_{\theta\gamma}^\rho = -\mathrm{tr}(\rho \ln(\rho_{\theta\gamma}))/\ln(d), \tag{18}$$

for later interpretations of noise-induced phenomena.

In general, the optimization of objectives $\mathcal{L}_{\theta\gamma} \to \mathcal{L}_{\theta\gamma}^*$, particularly in noisy settings to determine the optimal parametrization θ_γ^*, has no closed forms [50]. Whether there are similarities between optimal noisy and noiseless quantities, such as infidelities $\mathcal{L}_{\theta\gamma}^* \approx \mathcal{L}_\theta^*$, or even more strongly between parameters $\theta_\gamma^* \approx \theta^*$, remains an open question [32]. In our subsequent analysis, in addition to numerical studies,

we derive analytically the leading-order scalings of the discussed quantities of interest, given our variables of the depth M and the noise scale γ.

We also note that all plotted statistics in this work reflect average behaviors of the *Ansatz* across independent optimizations with respect to Haar random initial pure states and Haar random targets. The minimum infidelity reached for each independent optimization for a given fixed depth and noise scale is used for statistics across samples. Error bars and shaded regions represent one standard deviation from the mean of samples. Lower error bars are plotted equal in length to the upper error bar on a log scale, and nonvisible error bars can be considered to be equal in scale to any plot markers. Error bars in this work generally appear to be relatively orders of magnitude smaller than their corresponding average values, and $S = 50 \leqslant O(100)$ samples are deemed adequate to capture all behaviors well in practice for $N \leqslant 4$ qubits. We note, however, that optimization hyperparameters, as per Appendix F, in particular the use of a modified conjugate-gradient-based optimizer with a Wolfe condition line search and appropriate learning rates, must be carefully selected. Our initial studies indicated that at larger depths and larger small noise scales, after an initially smooth convergence in infidelity, optimization routines can oscillate rapidly between local minima, as also observed in previous optimization studies [27]. A point concerning notation used in this work: quantities computed within a noisy context generally have γ subscripts, such as parameters obtained in a noisy setting θ_γ. Otherwise in noiseless settings, any noise subscripts are dropped. The Trotterization order Q superscripts are also generally dropped for simplicity as $Q = 2$ is held fixed.

III. METHODS

We now develop formalisms to understand the scaling of channel-dependent quantities with respect to depth and noise scales. We first develop methods for comparing constrained versus unconstrained noiseless *Ansätze*, as detailed in Appendix B. We follow the formalism developed by Larocca *et al.* [28], which sets bounds on the rank $R = R(P) \leqslant P$ of the quantum Fisher information $\mathcal{F}_{\theta_{\mu\nu}}$ to determine whether a quantum system with P parameters is overparametrized. At an overparametrization limit $\tilde{P} = O(G)$, defined in terms of the dimension of the space spanned by the *Ansätze*, this metric's rank is shown to transition from being full rank $R = P$ for underparametrized $P < \tilde{P}$, to saturating at this limit $R = \tilde{P}$ for adequately or overparametrized $P \geqslant \tilde{P}$. Underlying these definitions of overparametrization is the quantum Fisher information's rank generally reflecting how many directions a parametrized *Ansatz* may span in its space of solutions. We seek to generalize these intuitions from state-dependent *Ansätze* to unitary-dependent *Ansätze*, independent of the initial state being transformed. The conventional state Fisher information $\mathcal{F}_\theta^\rho$ is defined in terms of states ρ_θ, given a state preparation objective $\mathcal{L}_\delta^\rho \sim \mathrm{tr}(\rho_\theta \rho_{\theta+\delta})$. We define a generalized unitary Fisher information $\mathcal{F}_\theta^U$, developed concurrently by Haug *et al.* [51], that is strictly dependent on the unitary *Ansatz* U_θ. Given our unitary compilation objective $\mathcal{L}_\delta^U \sim \mathrm{tr}(U_\theta^\dagger U_{\theta+\delta})$, we may derive the unitary Fisher information as the leading-order deviation, in the perturbing parameters

$\delta \to 0$, of the objective

$$\mathcal{F}^U_{\theta_{\mu\nu}} = \frac{1}{d^2} \text{Re}(d\,\text{tr}(\partial_\mu U^\dagger_\theta \partial_\nu U_\theta)$$
$$- \text{tr}(\partial_\mu U^\dagger_\theta U_\theta)\text{tr}(U^\dagger_\theta \partial_\nu U_\theta)), \qquad (19)$$

which reduces to the state Fisher information definition in the limit of tracing over $d = 1$-dimensional states

$$\mathcal{F}^\rho_{\theta_{\mu\nu}} = \text{Re}(\langle \partial_\mu \rho_\theta | \partial_\nu \rho_\theta \rangle - \langle \rho_\theta | \partial_\mu \rho_\theta \rangle \langle \partial_\nu \rho_\theta | \rho_\theta \rangle). \qquad (20)$$

The rank of the unitary quantum Fisher information subsequently offers insight into the capabilities of an *Ansatz* to span a set of unitaries, and to potentially become overparametrized for compilation tasks. We also may derive expressions for the resulting noise-dependent quantities that allow for an easier interpretation for noise-induced phenomena. Crucially, we differentiate states by their number of errors due to local noise,

$$\mathcal{N}_\gamma = \sum_{\substack{k \ \chi_k \in [2]^K \\ |\chi_k| = k}}^K \circ_m^M \otimes_i^N \gamma^{\chi_{k_i}^m}(1-\gamma)^{1-\chi_{k_i}^m} \mathcal{K}^{(m)\chi_{k_i}^m}_{\gamma_i}. \qquad (21)$$

Here, we refer to an error as any single nonidentity noise operation $\mathcal{K}_\gamma$ acting locally at any of $K = NM$ possible qubits and time indices, interlaced within a unitary *Ansatz* $\mathcal{U}_\theta$. We then may define k-error channels $\Lambda_{\theta\gamma_k}$ that are convex combinations of channels consisting of all possible locations of $k \leqslant K$ errors, represented by multi-indices $\chi_k \in [2]^K : |\chi_k| = k$, each with probability γ^k. The channels that generate at most K-error noisy states, can therefore be seen to have an interesting form of an expectation value over a distribution of k-error channels,

$$\Lambda_{\theta\gamma} = \left\langle \Lambda_{\theta\gamma_k} \right\rangle_{k \sim p_{K\gamma}}. \qquad (22)$$

For noise models defined in terms of strictly identity or nontrivial errors, the exact distribution over the errors

$$p_{K\gamma}(k) = \binom{K}{k}\gamma^k(1-\gamma)^{K-k} \qquad (23)$$

is the binomial distribution with mean $K\gamma$. As discussed in Appendix C, other interpretations for noise can arise due to the binomial distribution being equivalent to other distributions in various limits of γ and K. Grouping the K, γ-dependent terms yields the leading-order scaling

$$\Lambda_{\theta\gamma} - \Lambda_\theta = \sum_{k>0}^K \binom{K}{k}\gamma^k \Lambda_{\theta\gamma \leqslant k}. \qquad (24)$$

We denote k-error channels as the uniform convex combination of all possible error locations

$$\Lambda_{\theta\gamma_k} = \frac{1}{\binom{K}{k}} \sum_{\chi_k} \Lambda^{\chi_k}_{\theta\gamma}, \qquad (25)$$

and we denote at-most-k-error operators as the nonuniform combination of all possible $l \leqslant k$ error locations

$$\Lambda_{\theta\gamma \leqslant k} = \sum_l^k (-1)^{k-l} \binom{k}{l} \Lambda_{\theta\gamma_l}, \qquad (26)$$

where specific k-error channels with errors at locations χ_k are denoted as

$$\Lambda^{\chi_k}_{\theta\gamma} = \circ_m^M \left[\left[\circ_i^N \mathcal{K}^{(m)\chi_{k_i}^m}_{\gamma_i} \right] \circ \mathcal{U}^{(m)}_\theta \right]. \qquad (27)$$

For example, the $k = 1$-order deviation from the noiseless channel

$$\Lambda_{\theta\gamma \leqslant 1} = \frac{1}{K} \sum_{m,i}^{M,N} \mathcal{U}^{(>m)}_\theta \circ \left(\mathcal{K}^{(m)}_{\gamma_i} - \mathcal{I}_i \right) \circ \mathcal{U}^{(\leqslant m)}_\theta \qquad (28)$$

depends on the nontriviality of the K possible single $\mathcal{K}^{(m)}_i \neq \mathcal{I}_i$ local channels. Generalizing this interpretation, the k-order deviation depends on how the $\binom{k}{l}$ possible $l \leqslant k$-error channels combine, weighted by their binomial coefficients, to cancel or deviate from the trivial noiseless channel. Channels consisting of multiple q types of errors at indices $\chi_k \in [q]^K$ may also be written as an expectation over a generalized multinomial distribution of channels.

As derived in Appendix C, gradients along parameter directions μ of objectives, which are linear in the parametrized channels with constant noise, follow parameter shift rules [52]. We denote such gradients with perturbing parameter angles φ and coefficients α_φ, which are dependent on the spectrum of the *Ansatz* generators

$$\partial_\mu \Lambda_{\theta\gamma} = \sum_\varphi \alpha^\mu_\varphi \Lambda_{\theta+\varphi\,\gamma}. \qquad (29)$$

These parameter shift rules indicate that the linear nature of the noise interlaced with the parametrized unitary *Ansatz* perturbatively affects the noisy quantities.

Whether classical sources of noise impose similar noise-induced phenomena to the quantum noise sources studied is a separate important question when determining the robustness of parametrized systems. As developed in Appendix D, we introduce the notion of classical noise due to numerical precision, or floating point error scale ϵ, and we determine its similarity to the quantum noise scale γ. This error affects both the representation and numerical operations of floating point scalars, which we extend to matrices $A \in \mathcal{M}(d)$ and matrix multiplication

$$\prod_\mu^k A_\mu \to \prod_\mu^k {}_\epsilon A_\mu = \sum_{\chi \in [2]^k} \prod_\mu^k A_\mu \Sigma^{\chi_\mu} \qquad (30)$$

for error matrices $\Sigma \in \mathcal{M}(d)$, interlaced at locations χ. We then relate classical $\epsilon = \epsilon(d) = \|\Sigma\|$, and quantum $\gamma = O(\|\mathcal{K}\|)$ noise phenomena, given the number of $k = O(K) = O(\text{poly}(N)M)$ matrix multiplication operations that are subject to classical error, are related to the number of K possible quantum errors in the *Ansatz*. We are thus able to derive exact bounds on the deviations of noisy infidelities from their noiseless counterparts

$$|\mathcal{L}_{\theta\epsilon} - \mathcal{L}_\theta| \leqslant |1 - (1+\epsilon)^{2k}| \sim O(2k\epsilon), \qquad (31)$$

$$|\mathcal{L}_{\theta\gamma} - \mathcal{L}_\theta| \leqslant 2|1 - (1-\gamma)^K| \sim O(2K\gamma), \qquad (32)$$

with the use of various properties of norms such as Holder and von-Neumann's inequalities for Schatten matrix norms [53]. Both classical and quantum noisy infidelities scale similarly: polynomially in the noise scale, exponentially in the number

of errors, and identically in the limit of small local error scales. However, the constant classical error factors depend on the dimension of the space of interest and the number of parameters, indicating that larger classically simulated systems experience potentially greater error. This scaling of infidelities analysis is *Ansatz*-independent and provides valuable insight into general deviations of noisy quantities from their noiseless counterparts. However tighter bounds are potentially possible if there are symmetries or relationships between the noise and dynamics, or if the optimization is analytically tractable, and if an exact form for $\mathcal{L}_{\theta_\gamma^* \gamma}$ may be derived. These upper bounds, as will be discussed in Sec. IV, also relate to the critical point at which noise effects begin to dominate for greater numbers of operations, and optimization enters a divergent regime.

IV. RESULTS

Classical simulations and analytical calculations are performed to quantify changes in the behavior of parametrized systems and their optimization, with depth and with noise scale. NMR *Ansatz* and simulation details are described in Appendix E and Appendix F, and all data can be found in a repository [54].

A. Unconstrained versus constrained optimization

We first investigate the effects of constraints on noiseless Haar random unitary compilation, with respect to the number of optimization iterations and various depths M. The controllable transverse fields $\theta_i^{x,y(m)}$ at each time step $m \in [M]$ and qubit $i \in [N]$ are constrained due to it being difficult to exercise individual control over qubits, $\theta_i^{x,y(m)} = \theta^{x,y(m)}$, pulse amplitudes are bounded $|\theta_i^{x,y(m)}| \leqslant \bar{\theta}$, and are generally turned off at the start and end of experiments, $\theta_i^{x,y(0,M-1)} = 0$.

Given the trends throughout optimization for the $N = 4$ NMR *Ansatz* and Haar random targets, as well as the rank saturation of the generalized unitary Fisher information metric in Fig. 2, we observe exponential infidelity convergence beyond a depth $M > \tilde{M} \approx G \sim O(256)$. We thus claim numerically that in noiseless settings, constrained optimization converges as

$$\mathcal{L}_\theta \sim e^{-\alpha M} : \quad M > \tilde{M} \tag{33}$$

for some rate of convergence α. In practice, $M \approx O(1000) \gg \tilde{M}$ is necessary for adequate convergence within a feasible number of optimization iterations. We also note that, in particular at larger depths, constrained tasks exhibit greater variance relative to unconstrained tasks, however both types of tasks have generally low relative variance, suggesting an appropriate choice of optimization hyperparameters, as per Appendix F.

We note that the nonlocal coupling $J \ll h, \theta$ is much smaller than other scales, and for experimentally realistic time steps, $\tau J \ll 1$. This limits how much of a nonlocal entangling gate can be implemented in a single time step, an essential part of generating Haar random unitaries. From these simulations, we conjecture that these constraints necessitate that the minimum depth where overparametrization can occur is increased
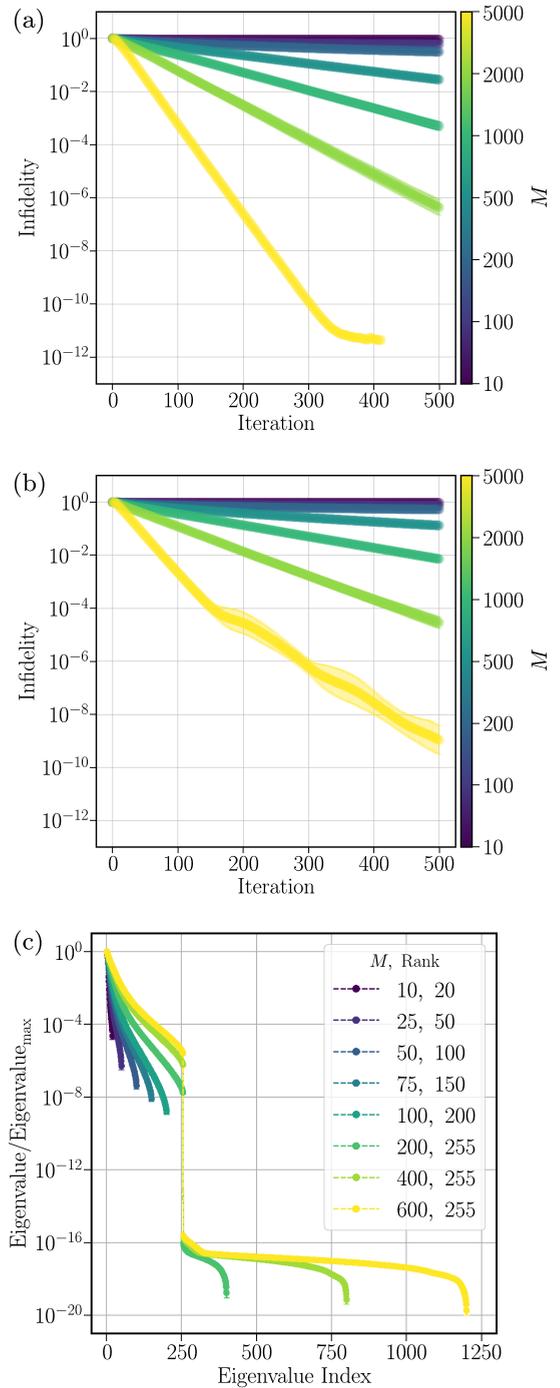

FIG. 2. Convergence of unitary compilation infidelity with respect to optimization iteration and depth M (colored/gradient) for the $N = 4$ NMR *Ansatz*. (a) Unconstrained parametrization with independent qubit parameters, and no boundary conditions. (b) Constrained parametrization with shared qubit parameters, and zero-field temporal boundary conditions. Constrained tasks require comparable iterations or depth to converge comparably to unconstrained tasks, and they exhibit exponential convergence beyond $M > \tilde{M} \approx O(d^2) = O(256)$. (c) Constrained quantum Fisher information eigenvalue spectrum at optimality. The spectrum is full-rank $R = P$ for $M < \tilde{M}$ before saturating at rank $R = \tilde{P} \sim O(d^2) \leqslant P$ for $M \geqslant \tilde{M}$, indicating overparametrization.

by a subexponential factor in the number of qubits,

$$\tilde{M}_{\text{Constrained}} \sim O\left(\frac{1}{\tau J}\text{poly}(N)\right)\tilde{M}_{\text{Unconstrained}}. \qquad (34)$$

Other than a shift in the overparametrization boundary, the given constraints do not appear to fundamentally affect the exponential infidelity convergence. We also note that even for relatively small system sizes of $N \leqslant 4$, several orders of magnitude deeper circuits than are typically classically simulated [26–28], with up to $O(10^5)$ gates, are necessary to study such realistic systems.

B. Noisy state preparation

We now investigate the effects of local noise on Haar random pure state preparation. Here parameters are unconstrained according to experimental feasibility to ensure all observed phenomena are due to noise. Given our findings on the effects of constraints, to leading order, imposing constraints should only shift the depth-dependent results by a noise-independent factor.

To demonstrate the interplay between depth and noise scales on optimized infidelities, we plot infidelities with respect to each independent quantity in Fig. 3. Here we display unital dephasing noise, and other unital and nonunital noise models are shown to exhibit similar behavior in Appendix C. When varying the noise scale in Fig. 3(a), for small noise scales the average optimal infidelity [solid lines in Fig. 3(a)] strictly scales as expected with depth. Increasing the noise scale past a depth-dependent critical noise scale γ_M causes infidelities to increase polynomially, between linearly and quadratically, until the infidelities plateau at their maximum value. We also investigate trends when inserting parameters learned in the noisy setting into an identical, but noiseless unitary *Ansatz*, yielding infidelities [dashed lines in Fig. 3(a)]. Such noiseless, tested infidelities are superior than their corresponding trained noisy infidelities in terms of their greater critical noise scale $\gamma_{M_{\text{Noiseless}}} > \gamma_{M_{\text{Noisy}}}$: the noiseless infidelities increase in the divergent regime at much larger noise scales. This suggests the optimization is learning about the underlying unitary dynamics, and is not just preparing another mixed state that happens to be close to the target pure state. The error bars of both the noisy trained and noiseless tested infidelities are equal in the convergent regime with $\gamma < \gamma_M$ [overlapped markers in the left of Fig. 3(a)], and they are orders of magnitude smaller in the divergent regime with $\gamma > \gamma_M$. Overall, the trends in error suggest the optimization and dynamics are most uncertain within the regime around and past the noise scale γ_M, indicating increased complexity near this transition.

The noiseless behavior can partly be explained from parameter shift rules for gradients of parametrized channels with constant noise, as derived in Appendix C. The noisy state is a convex combination of parametrized pure states, each with identical gradient directions to the noiseless case, albeit with magnitudes that are scaled by polynomials of the noise scale. Therefore, the trajectory of the gradient-based optimization remains similar at small noise scales in both noisy and noiseless cases.

When varying the depth, we observe a critical noise-dependent depth M_γ that occurs in Fig. 3(b). Previously decreasing infidelities with depth in a convergent regime
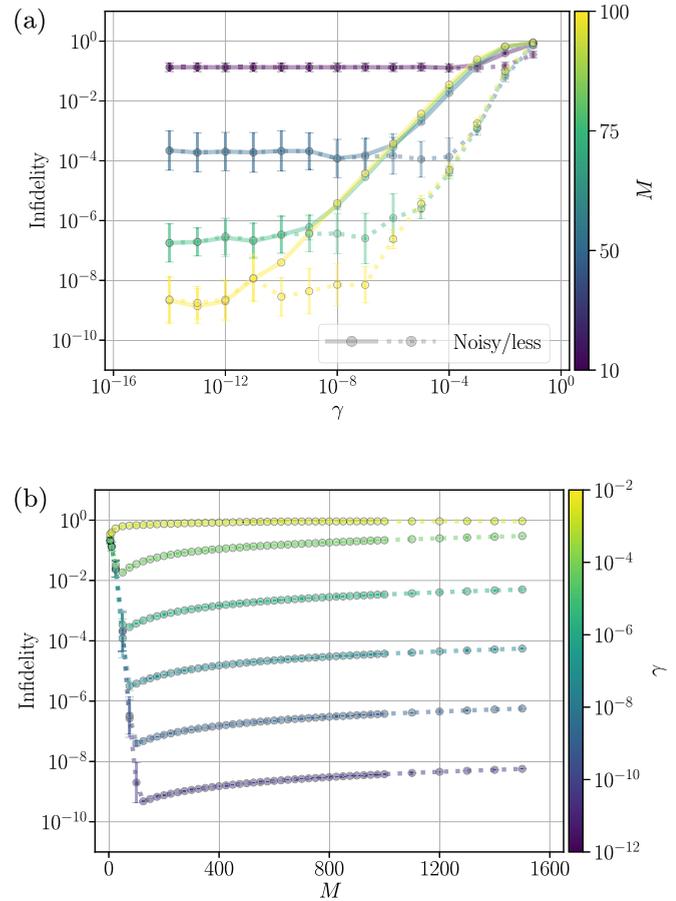

FIG. 3. Behavior of state preparation infidelity with respect to unital dephasing noise γ and depth M for the $N = 4$ NMR *Ansatz*. (a) Trained noisy infidelity (solid), and tested infidelity of noisy parameters in noiseless *Ansatz* (dashed), with respect to noise scale, for various depths M (colored/gradient). Infidelities are depth-dependent and noise-independent for small noise scales, before universally increasing polynomially with noise. Tested noiseless infidelities indicate that the underlying unitary dynamics are being learned resiliently. (b) Critical depth for noisy infidelity for various noise scales (colored/gradient). Infidelities improve exponentially with depth, up until a noise-induced critical depth, where entropic effects worsen infidelities polynomially.

for $M < M_\gamma$ increase uniformly with depth in a divergent regime for $M > M_\gamma$. Beyond this depth, the increase in expressiveness of the *Ansatz* to prepare arbitrary states from increasing the number of variable parameters is outweighed by the accumulated noise from the increased sources of error. From a trainability standpoint, beyond this critical depth, noise-induced barren plateaus may be occurring, leading to a decrease in trainability where the gradients are unable to find the ideal trajectory to reach optimality. Alternatively, from an expressiveness standpoint, noise potentially increases in an uncontrolled manner the number of directions permitted to be explored in the objective landscape [40]. Future work should investigate the density of mixed states that are perturbatively away from a given pure state [15,55]. As discussed below, through analytical calculations, we offer
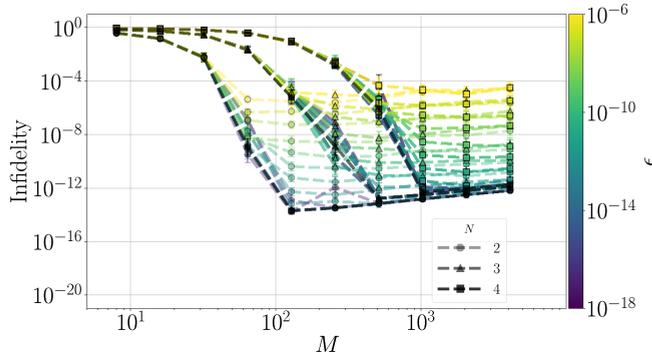
FIG. 4. Behavior of unitary compilation infidelity with respect to depth M and classical floating point noise scale ϵ (colored/gradient) for various N NMR *Ansatz*, relative to the noiseless case (black). Decreasing error scale verifies the classical noise model's suitability and estimates the architecture-dependent machine precision of $\varepsilon \sim O(10^{-16})$. Classical noise is also shown to exhibit a critical depth M_ϵ and divergent regime.

complementary interpretations into exactly how entropic effects begin to dominate infidelity behaviors.

C. Universal effects of classical and quantum sources of noise

These studies bring to mind the question of whether noise-induced critical depth phenomena can be attributed to strictly quantum or potentially classical noise phenomena. For very deep ideal noiseless *Ansatz* with $M > O(1000)$, infidelities approach machine precision and start to increase with depth. This suggests floating point errors accumulate for large numbers of simulated operations. As per the noise models derived in Appendix D, we insert artificial classical floating point error of different scales into simulations, as per Fig. 4. Adding zero-mean random errors with standard deviation proportional to error scales to results of floating point operations shows similar trends in noise-induced convergent and divergent regimes. Since unitary compilation tasks have quadratically more degrees of freedom than state compilation tasks, their infidelities decrease slower with depth. Furthermore, infidelity curves reach their critical depth at earlier depths for smaller system sizes due to the exponentially smaller spaces to search.

Infidelities are shown to increase and enter the divergent regime when they reach a scale proportional to that of the error scale. This error scale may be approximately upper bounded by the derived deviations of the noisy infidelities from their noiseless counterparts. By decreasing the artificial floating point error, the curves also converge to the supposedly noiseless case, offering an estimate for the machine precision of $\varepsilon \sim O(10^{-16})$. These trends open many questions on the viability of large-scale simulations close to machine precision with finite floating point architectures. Can arbitrarily large systems be accurately simulated without resorting to inefficient arbitrary precision arithmetic, or error mitigation or correction approaches [14,44]?

V. DISCUSSION

From numerical experiments, we find for a given scale of local noise γ that there is a critical depth of circuit M_γ,

beyond which optimal infidelities increase with depth due to an accumulation of noise. From fitting procedures discussed in Appendix C, we are able to determine the critical depth to be logarithmic in the noise scale

$$M_\gamma \sim \ln(1/\gamma). \tag{35}$$

The optimal infidelity is therefore approximately

$$\mathcal{L}_{\theta^*\gamma|M_\gamma} \sim e^{-\alpha M_\gamma} \sim \gamma^\alpha, \tag{36}$$

with $1 \leqslant \alpha \leqslant 2$, confirming previous conjectures of linear or quadratic scaling of infidelity with noise [35]. The interpretation of a noisy channel being a binomial distribution of k-error channels also suggests that parametrized quantum channels can mitigate approximately

$$\bar{K}_\gamma \sim \gamma \ln(1/\gamma) \tag{37}$$

errors. Determining whether the optimization is finding a parametrization that explicitly performs error mitigation [30,44,56], or even error correction through a parametrized encoding [57], would constitute important future contributions [14,58]. The presence of a noise-induced depth is also reminiscent of weak measurement-induced phase transitions [59]. However, these noise-induced effects intuitively should be apparent at all system sizes, and do not seem to be related to typical indicators of phase transitions such as scale invariance.

As derived in Appendix C, the Bloch representation allows explicit leading-order scaling of quantities with respect to depth and noise to be derived. Here, we represent states as $\rho = (1/d)(I + \lambda \cdot \omega)$, with Bloch coefficients λ associated with a set of $d^2 - 1$ nonidentity trace orthogonal basis operators ω. Channels Λ may then be represented as affine transformations $\lambda \to \Gamma\lambda + \upsilon$. Parametrized noisy channels $\Lambda_{\theta\gamma}$ with K possible errors may be decomposed into strictly unitary u_θ and unital $u_{\theta\gamma}$ and nonunital $\eta_{\theta\gamma}$ noise-dependent components

$$\Gamma_{\theta\gamma} = (1 - \gamma)^K u_\theta + [1 - (1 - \gamma)^K]u_{\theta\gamma}, \tag{38}$$

$$\upsilon_{\theta\gamma} = [1 - (1 - \gamma)^K]\eta_{\theta\gamma}. \tag{39}$$

In this Bloch representation, we may then express a parametrized noisy state as

$$\rho_{\theta\gamma} = (1 - \gamma)^K \rho + [1 - (1 - \gamma)^K]\epsilon_{\theta\gamma} + \Delta_{\theta\gamma}. \tag{40}$$

This decomposition expresses the interplay of the parametrized unitary and noise-induced nonunitary components of the channel. The unitary component rotates what we refer to as the pure component of the state. This component consists of a superposition of the pure target state ρ, with associated coefficients λ, and an orthogonal pure state with orthogonal associated coefficients $\zeta \perp \lambda$, represented within the traceless deviation term $\Delta_{\theta\gamma}$. The noise component of the channel scales the pure component of the state with the noise scale $1 - \gamma$, plus it shifts the state by what we refer to as the mixed component of the state $\epsilon_{\theta\gamma}$, with associated Bloch coefficients $\varepsilon_{\theta\gamma}$. In the limit of the optimization reaching optimality in the noisy setting $\theta \to \theta^*_\gamma$, the pure component of the state approaches the pure target state, the deviation term $\Delta_{\theta^*_\gamma\gamma} \to 0$ approaches zero, and there only remains an inherent noise-dependent mixed component $\epsilon_{\theta^*_\gamma\gamma}$. We may

then assume that optimality is reached in this noisy setting, and channel-dependent quantities may be expanded in the number of errors and noise scale. At optimality, we find our quantities of interest analytically scale similarly to leading order in K, γ, namely

$$\mathcal{L}^\rho_{\theta\gamma} \sim K\gamma \frac{d-1}{d}\left(1 - \frac{\lambda \cdot \varepsilon_{\theta\gamma}}{\lambda^2}\right) + O\left(\binom{K}{2}\gamma^2\right), \qquad (41)$$

$$\mathcal{I}_{\theta\gamma} \sim 2K\gamma \frac{d-1}{d}\left(1 - \frac{\lambda \cdot \varepsilon_{\theta\gamma}}{\lambda^2}\right) + O\left(\binom{K}{2}\gamma^2\right), \qquad (42)$$

$$\mathcal{S}_{\theta\gamma} \sim O(K\gamma), \quad \mathcal{D}^\rho_{\theta\gamma} \sim O(K\gamma). \qquad (43)$$

As derived in Appendix C, the Bloch representation allows exact leading-order terms to be derived for quantities that are strictly polynomial functions of the Bloch coefficients. Quantities that are more complex, for example logarithmic, functions of the Bloch coefficients require knowledge of the algebras governing the specific choice of basis ω, yielding in principle calculable [60] but nonintuitive forms. Importantly, when the mixed components $\varepsilon_{\theta\gamma} \parallel \lambda$ become pure, or aligned with the pure target coefficients, quantities differ from their noiseless values by strictly higher-order terms. This purification of the mixed component could be due to the lack of noise, or the specific combination of parametrizations and noise models forcing the system towards a pure state, and it describes general error mitigation. The use of Block representations thus allows simplified and occasionally *Ansatz*-independent depictions of noise-induced phenomena. Beyond the critical depth, infidelities appear both analytically and numerically to be linear functions of entropy and impurity, and they scale with the overlap of the pure target state with the mixed component. In particular, at low noise scales in the divergent noisy regime, analytical predictions and numerical simulations of the discussed quantities all correspond precisely, as demonstrated in Fig. 9 in Appendix C. Further, from Fig. 5, all quantities appear to collapse together with increasing noise scales. Reasoning that noise phenomena dominate at $M = M_\gamma$, when the scale of the optimization-driven decreasing optimized infidelity $\mathcal{L}^\rho_{\theta^*\gamma}$ reaches the scale of the entropic-driven increasing analytical infidelity $\mathcal{L}^\rho_{\theta^*\gamma}$,

$$\mathcal{L}^\rho_{\theta^*\gamma} \sim e^{-\alpha M}\big|_{M_\gamma} \approx \mathcal{L}^\rho_{\theta^*\gamma} \sim NM\gamma\big|_{M_\gamma}, \qquad (44)$$

and we recover our numerically predicted noise-induced critical depth $M_\gamma \sim \ln(1/\gamma)$. We note that bipartite (entanglement) entropy, between a system and its environment, is generally bounded strictly by the system size $N = K/M$. However, noisy quantities are polynomials of k-error factors $p_{K\gamma}(k) \sim O(K^k \gamma^k)$. The additional degree of freedom representing the strength of system-environment interactions γ appears to suppress the higher-order polynomial factors.

Universal behaviors in the divergent regime can be attributed to entropy-increasing phenomena, once the parametrized channel has rotated the state to within a depth-dependent distance from the target state. However, at large noise scales there are important distinctions between unital versus nonunital types of noise. For unital noise, such as dephasing noise in Fig. 3, a potential explanation
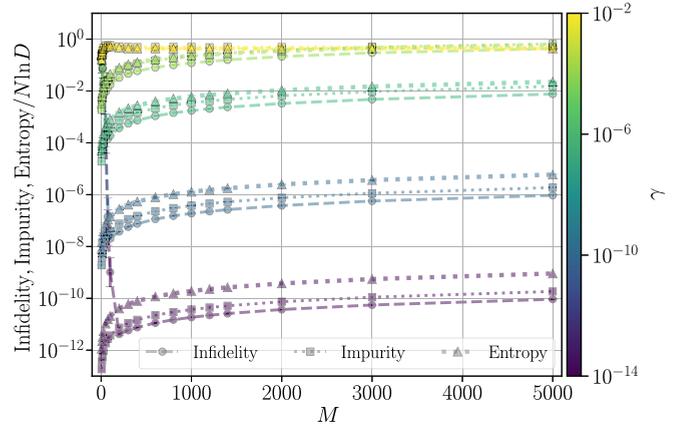

FIG. 5. Behavior of infidelity, entropy, and impurity with respect to depth M and nonunital amplitude damping noise γ (colored/gradient) for the $N = 4$ NMR *Ansatz*. At small noise scales, quantities scale identically linearly with depth and with noise in the divergent, entropic driven regime. At large noise scales, unlike unital noise, nonunital noise infidelities decrease polynomially with depth once the parametrized unitary aligns the state towards the target pure state.

is that the combination of the potentially close to Haar random parametrized unitaries, and the accumulated noise, induces depolarization. Entropy is shown to increase linearly with depth, at practically all noise scales, and dominates the infidelity behaviors. For nonunital noise, such as amplitude damping noise in Fig. 5, at small noise scales the behavior appears qualitatively similar to unital noise of linearly increasing infidelity along with impurity and entropy. However, nonunital noise appears to have fundamentally different nonuniversal behavior at large noise scales. Nonunital noise forces the state into a specific (pure) state, which appears to improve the infidelities. Unlike in the overparametrized regime, however, infidelities appear to decrease polynomially with depth. Adequately parametrized unitaries at depths far beyond the typical overparametrization bound appear necessary to rotate some components of this forced state towards the target state. Ultimately, optimization routines in a noisy setting are shown to be capable of rotating the pure components of states towards target pure states, and they exhibit overparametrization phenomena. Once objectives approach a noise-induced entropy-dependent scale, entropic effects then dominate objective behavior with increasing depth. Finally, fundamental differences between unital and nonunital noise at large depths and large noise scales are relevant when considering NISQ applications.

VI. CONCLUSION

Through this work's classical simulation and analytical treatments, overparametrization phenomena for quantum systems are shown to be robust under realistic settings. Infidelities decrease exponentially with depth in the convergent regime before increasing polynomially with depth and with noise in the divergent regime. These scalings provide essential data for the experimental design of variational quantum algorithms.

When assessing the relevance of this work, given we simulate specifically NMR systems with depths $M \in [10, 5000]$ and noise scales $\gamma \in [10^{-18}, 10^{-2}]$, we must assess the depth and noise scales of other implementations, such as trapped ions, superconducting qubits, and neutral atoms in Table III. Given our derived logarithmic dependence with noise scales of the noise-induced critical depth, we conjecture that NMR systems' robustness deteriorates at depths $M_\gamma \sim O(200)$ and noise scales of $\gamma_M < O(10^{-3})$ if we require infidelities $\mathcal{L} < O(10^{-4})$. Although the full span of considered depths and noise scales exceeds currently experimentally feasible regimes for NISQ implementations, the derived limits where robustness deteriorates are still approaching currently feasible scales. Other implementations are expected to show identical phenomena, with some *Ansatz*-dependent shift in these depth and noise scales. It should also be noted that nonunital noise appears to allow for reimproved infidelities in the divergent regime at exceptionally large depths of $M \sim O(5000)$ and large noise $\gamma \sim O(10^{-2})$. There may be intriguing nontrivial, noise-type induced emergent phenomena at these large depths, even if these regimes are currently impractical experimentally. This work further serves as studies of general noise phenomena, which are likely to be encountered when existing implementations are scaled to address practical problems. The conclusions drawn, therefore, support the necessity of quantum error correction, and they challenge aspirations [61,62] of existing NISQ applications being scaled to thousands of qubits and gates.

Finally, we remark that entropic effects appear to dominate only beyond a critical number of errors in the system. General parametrized systems are thus shown to be capable of suppressing entropic behavior imposed by their environment. By locating the noise-induced critical depth, problems can be optimized to their best-case objectives across all depths, for example in coveted quantum control problems [19]. This opens up intriguing applications [63] for variational *Ansätze*, both classical and quantum, and we are excited about their potential.

APPENDIX A: BACKGROUND

In these Appendixes, we elaborate on the specific parametrized quantum channel *Ansätze* studied in this work. We discuss the unitary and nonunitary components of the channel, and their approximations and derivations via Trotterization. These Trotterized forms correspond to quantum circuit models of continuous evolution of quantum systems up to a specified order of precision, and they are used for classical simulation.

1. Unitary evolution

The unitary evolution operator takes the form of the time ordered matrix exponential of the Hamiltonian

$$U_\theta = \mathcal{T} e^{-i \int_0^T dt\, H_\theta^{(t)}}. \tag{A1}$$

Here the time-dependent Hamiltonian driving the evolution takes the parametrized form

$$H_\theta^{(t)} = \sum_\mu H_\mu^{(t)}, \tag{A2}$$

with a set of poly(N) parameters $\theta^{(t)}$ at each time $t \in [0, T]$. Each term in the Hamiltonian is parametrized with fixed generators $\mathcal{G} = \{G_\mu\}$ as

$$H_\mu^{(t)} = \theta_\mu^{(t)} G_\mu. \tag{A3}$$

The set of operators $\{G_\mu\}$ may contain local or nonlocal operators, and generally is at least partially noncommuting.

This continuous evolution generated by exponential maps of Hamiltonians must be discretized temporally and spatially across the space of subsystems for feasible classical simulation. This discretization further allows for comparison with the variational quantum circuit paradigm. Depending on the control problem of interest, there are several choices for the specific discretization scheme. Further, given constraints placed on the parameters, the explicitly optimized parameters may take various functional forms.

2. Trotterization

To classically simulate such time-dependent systems, unitaries are Trotterized both temporally and spatially across the space. To first-order in time,

$$U_\theta \approx \prod_m^M U_\theta^{(m)} + O(\tau^2), \tag{A4}$$

where the time has been discretized into M time steps of size $\tau = T/M$, and evolution at each time step $m \in [M]$ is

$$U_\theta^{(m)} = e^{-i\tau H_\theta^{(m)}}. \tag{A5}$$

Further, depending on the commutation relations between terms in the Hamiltonian to Q-order in space across the qubits, at a given instance m in time,

$$U_\theta^{(m)} \approx \prod_\mu^{(Q)} U_\mu^{(m,Q)} + O\big([H_\mu^{(m)}, H_\nu^{(m)}]_{Q+1} \tau^{Q+1}\big). \tag{A6}$$

The product $\prod_\mu^{(Q)}$ represents a product over some function of lower-order Trotterizations, denoted by $U_\mu^{(m,Q)}$, with the error

being proportional to the $Q + 1$-deep nested commutator of Hamiltonian terms at time m [65]. For $Q = 2$, the product is equal to the forward and backward ordering of first-order Trotterized operators

$$U_\mu^{(m,Q)} \overset{Q=2}{=} U_\mu^{(m)1/Q}, \tag{A7}$$

with a factor of $1/Q$ in the generators to ensure consistency, and

$$U_\theta^{(m)} \overset{Q=2}{\approx} \overrightarrow{\prod_\mu} U_\mu^{(m)/Q} \overleftarrow{\prod_\mu} U_\mu^{(m)/Q} + O(\tau^3). \tag{A8}$$

The resulting evolution can be directly described by a parametrized quantum circuit, consisting of operators $U_\mu^{(m)}$ with locality of the corresponding Hamiltonian generator G_μ. Therefore, the following results, up to the precision of these discretizations, are relevant both in the discrete gate-based circuit model of quantum systems and the continuous time evolution model, more generally seen in pulse-level and quantum control problems [19,66].

3. Parametrized quantum channels

To understand the effects of noise on the evolution and abilities of such a system to represent various targets, we describe the evolution with parametrized quantum channels acting on states as

$$\Lambda_{\theta\gamma} = \mathcal{N}_\gamma \circ \mathcal{U}_\theta \tag{A9}$$

such that the evolved parametrized state from an initial state σ is

$$\rho_{\theta\gamma} = \Lambda_{\theta\gamma}(\sigma). \tag{A10}$$

The channel is composed of a noiseless unitary channel with variable parameters θ,

$$\mathcal{U}_\theta(\cdot) = U_\theta \cdot U_\theta^\dagger, \tag{A11}$$

and a noisy nonunitary Kraus operator channel with fixed noise parameters γ,

$$\mathcal{N}_\gamma(\cdot) = \sum_\alpha K_\gamma^{(\alpha)} \cdot K_\gamma^{(\alpha)\dagger}. \tag{A12}$$

All channels must be normalized such that they are trace-preserving.

Given the Trotterization of the continuous time evolution, we define the channel as a composition of M layers of channels at each time step,

$$\Lambda_{\theta\gamma} = \circ_m^M \left(\mathcal{N}_\gamma^{(m)} \circ \mathcal{U}_\theta^{(m)} \right). \tag{A13}$$

We also make use of the notation for the composition of channels before or after an index m as

$$\Lambda_{\theta\gamma}^{\leqq m} = \circ_{l \leqq m} \left(\mathcal{N}_\gamma^{(l)} \circ \mathcal{U}_\theta^{(l)} \right). \tag{A14}$$

We also write the Q-order spatial Trotterization of the unitary part of the channel as a composition over the Trotterized unitaries, with identical notation to the products of unitaries when deriving their Trotterization,

$$\mathcal{U}_\theta^{(m)} = \circ_\mu^{(Q)} \mathcal{U}_\mu^{(m,Q)}, \tag{A15}$$

and we can similarly define the notation for composition of channels before or after an index μ as

$$\mathcal{U}_{\leqq\mu}^{(m)} = \circ_{\nu \leqq \mu}^{(Q)} \mathcal{U}_\nu^{(m,Q)}. \tag{A16}$$

These notations may be combined for partial channels relative to before or after an index (μ, m), such that

$$\Lambda_{\theta\gamma} = \Lambda_\gamma^{(>m)} \circ \mathcal{N}_\gamma^{(m)} \circ \mathcal{U}_{>\mu}^{(m)} \circ \mathcal{U}_\mu^{(m)} \circ \mathcal{U}_{<\mu}^{(m)} \circ \Lambda_\gamma^{(<m)}. \tag{A17}$$

The corresponding partially forward evolved state relative to a state σ from the action of the channel before an index (μ, m) may then be denoted as

$$\rho_{<\mu\gamma}^{(<m)} = \Lambda_{<\mu\gamma}^{(<m)}(\sigma), \tag{A18}$$

and the corresponding backward evolved operator relative to an input operator O from the adjoint action of the channel after an index (μ, m) may also be denoted as

$$O_{>\mu\gamma}^{(>m)} = \Lambda_{>\mu\gamma}^{(>m)\dagger}(O). \tag{A19}$$

This notation is used to understand parameter shift rules for channels, and to derive the scaling of objectives with noise and with depth. Here we have dropped the (m, Q) super-script notation for the Q-order spatial Trotterization in favor of (m) superscript notation for simplicity, and it is assumed that operators at spatial indices are implicit functions of the Trotterization scheme.

APPENDIX B: LEARNING PHENOMENA

In these Appendixes, we give an overview of learning phenomena in quantum settings, in particular overparametrization phenomena, and we discuss their relevance to this work. We derive a unitary version of the quantum Fisher information, and we show numerically that its rank acts as an indicator of overparametrization identically to the state quantum Fisher information studied in previous works. Finally, we use these results to confirm that overparametrization phenomena occur in realistic settings, as per established definitions.

A key aspect of this work involves understanding whether in realistic settings indicators of overparametrization, in particular exponential convergence of the optimization with the number of parameters, still occur. We follow the approaches by Larocca *et al.* [28], later followed up by Garcia-Martin *et al.* [40], which set bounds on the rank of the quantum Fisher information to determine whether a quantum system is overparametrized, both in noiseless and then noisy settings.

Overparametrization in this context, as per the Fisher information definition [28], refers to when there are adequate number P parameters such that the model *Ansatz* can span the space [22,26] of the dynamical Lie algebra $\mathcal{G}$ formed by its generators $\{G_\mu\}$. This parametrization generally translates to the optimization procedure being able to converge exponentially with the number of optimization iterations. This may occur due to a fundamental change in the objective landscape where it becomes much more convex in this regime. It may also be accompanied by what is known as lazy training, where the optimal parameters are negligibly different from their random initial values.

In the continuous time evolution, or gate-based circuit formalisms, these generators are the noncommuting terms in

the underlying Hamiltonian that drives the evolution, and the dynamical Lie algebra is formed by the Lie closure

$$\mathcal{G} = \langle \{G_\mu\} \rangle_{\text{Lie}} \tag{B1}$$

of all linearly independent nested commutators of these generators.

This dynamical Lie algebra span forms a subspace $\mathcal{G} \subseteq \mathcal{H}$ of the full space $\mathcal{H}$ where the operators act, and it has dimension

$$G = |\mathcal{G}|. \tag{B2}$$

Within the context of spaces with dimension $d = D^N$, example algebras associated with the Lie closure that arise in quantum control settings include the special unitary algebra $su(D^N)$ with dimension $|su(D^N)| = D^{2N} - 1$, or the symplectic algebra $sp(2N)$ with dimension $|sp(2N)| = 2N^2 + N$. Generally, the dynamical Lie algebra dimension is either a polynomial or an exponential function of the system size N, and one-dimensional Lie algebras have been classified in [67]. The number of parameters P for a fixed periodic *Ansatz* repeated for M layers generally scales as

$$P \sim O(\text{poly}(N)M) \tag{B3}$$

and the depth itself may depend on the system size, depending on the *Ansatz*. For overparametrization to occur, the number of parameters must be of at least similar order to this dimension,

$$P > \tilde{P} \sim O(G), \tag{B4}$$

and in ideal settings, overparametrization occurs when exactly $\tilde{P} = G$.

A key indicator [28,40] of overparametrization is whether the rank of the quantum Fisher information $\mathcal{F}_\theta$,

$$R_\theta^{\mathcal{F}} = \text{rank}(\mathcal{F}_\theta) \leqslant G, \tag{B5}$$

saturates at this dimension. It can be shown that this saturation is independent of where you are in the objective or parameter landscapes, and it does not occur only at optimality.

Similar bounds for the Hessian $\mathcal{H}_\theta$ of the objective, being objective- and target-dependent, only occur at optimality,

$$R_{\theta*}^{\mathcal{H}} = \text{rank}(\mathcal{H}_\theta) \leqslant G, \tag{B6}$$

and otherwise the Hessian is generally full rank $R_\theta^{\mathcal{H}} = P$.

We should also note recent developments [40] in understanding the effects of noise on variational quantum circuits in the context of the spectrum of the quantum Fisher information. Small amounts of local noise are shown to allow more directions to be searched [55]. These additional directions in the state space may increase or decrease the parametrized state's purity. However, for increasing noise scales the system becomes exponentially less sensitive to its parameters, and it can search fewer and fewer directions. As shown in Sec. IV, we also observe different regimes depending on the noise scales, which confirm this recent analysis. We observe that the quasilinear scaling of infidelity with noise is at the boundary of the convergent and divergent regimes. There is finite noise that suppresses the system's abilities to achieve perfect infidelity, however not so much as to reach the divergent regime.

1. Unitary quantum Fisher information

We now study the form of the quantum Fisher information, which can be shown [68] to be proportional to the second-order correction to the Bures metric, also referred to as the Fubini-Study metric in the case of pure states. This quantity offers insight into which directions of the space can be reached, given the variational *Ansatz*.

Here, we generalize the definition of the Fisher information from states to a state-independent definition that only depends on the underlying parametrized unitary, similar in form to another recent study of generalized metrics [51].

We now investigate the second-order term, or Fisher-like information of distance measures $\mathcal{L}_\theta^U$ between a parametrized unitary U_θ with parameters $\theta = \{\theta_\mu\}$, and a fixed reference unitary U over a d-dimensional space. We define $\mathcal{L}_{\theta*}^U = \mathcal{L}_\theta^U |_{U_\theta = U}$ to be the parametrized unitary evaluated exactly at the reference unitary.

Let a distance $\mathcal{L}_\theta^U$ between unitaries (that is not a proper distance metric as it does not satisfy the triangle inequality) be related to the absolute trace overlap

$$\mathcal{L}_\theta^U = 1 - \frac{1}{d^2} |\text{tr}(U^\dagger U_\theta)|^2, \tag{B7}$$

with derivatives with respect to θ of

$$\partial_\mu \mathcal{L}_\theta^U = -2\frac{1}{d^2} \text{Re}(\text{tr}(U_\theta^\dagger U)\text{tr}(U^\dagger \partial_\mu U_\theta)), \tag{B8}$$

$$\partial_{\mu\nu} \mathcal{L}_\theta^U = -2\frac{1}{d^2} \text{Re}(\text{tr}(U_\theta^\dagger U)\text{tr}(U^\dagger \partial_{\mu\nu} U_\theta)$$
$$- \text{tr}(\partial_\mu U_\theta^\dagger U)\text{tr}(U^\dagger \partial_\nu U_\theta)). \tag{B9}$$

At optimality where $U_\theta = U$,

$$\mathcal{L}_{\theta*}^U = 0, \quad \partial_\mu \mathcal{L}_{\theta*}^U = 0,$$

$$\partial_{\mu\nu} \mathcal{L}_{\theta*}^U = 2\frac{1}{d^2} \text{Re}(d\,\text{tr}(\partial_\mu U_\theta^\dagger \partial_\nu U_\theta)$$
$$- \text{tr}(\partial_\mu U_\theta^\dagger U_\theta)\text{tr}(U_\theta^\dagger \partial_\nu U_\theta)). \tag{B10}$$

To define the Fisher information metric, we define it as the leading-order behavior of the objective given a perturbation of parameters $\theta \to \theta + \delta$, evaluated at $U = U_\theta$, yielding

$$\mathcal{L}_{\theta+\delta}^U = \mathcal{F}_{\theta_{\mu\nu}}^U \delta_\mu \delta_\nu + O(\delta^3), \tag{B11}$$

with the Fisher information metric being

$$\mathcal{F}_{\theta_{\mu\nu}}^U = \frac{1}{d^2} \text{Re}(d\,\text{tr}(\partial_\mu U_\theta^\dagger \partial_\nu U_\theta)$$
$$- \text{tr}(\partial_\mu U_\theta^\dagger U_\theta)\text{tr}(U_\theta^\dagger \partial_\nu U_\theta)). \tag{B12}$$

This state-independent quantity, identical to other Fisher information metrics, contains a term that reflects the change in the *Ansatz*, plus a corrective term to ensure gauge invariance, with additional dimensionality d factors to reflect that the action of the *Ansatz* with respect to specific states is not being considered.

We can also use a proper distance metric in terms of the Frobenius norm $\|A\|^2 = \text{tr}(A^\dagger A)$ [69], such that a proper objective is invariant up to phases between the operators

$$\tilde{\mathcal{L}}_\theta^U = 1 - \sqrt{1 - \mathcal{L}_\theta^U} = \frac{1}{2}\frac{1}{d} \max_\phi \|U_\theta - e^{i\phi}U\|^2, \tag{B13}$$
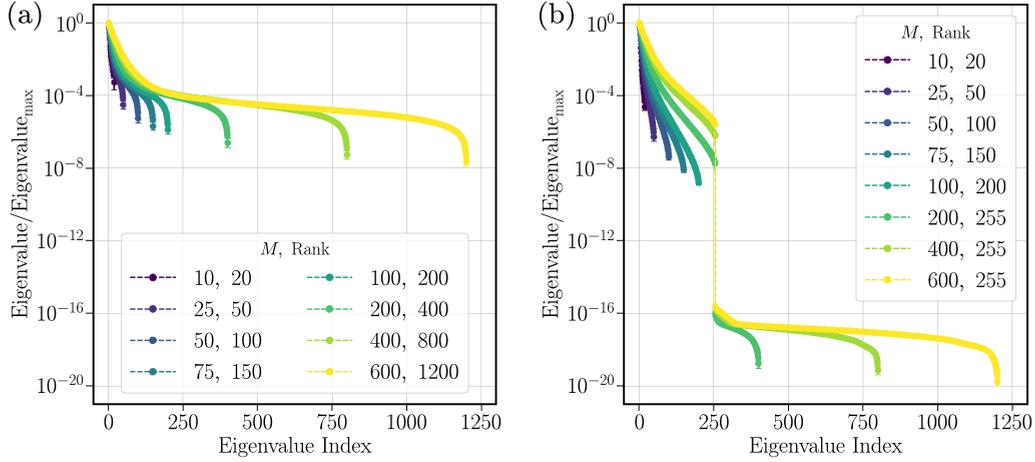
FIG. 6. Metrics of overparametrization for constrained unitary compilation of an $N = 4$ qubit NMR *Ansatz* for various depths M (colored/gradient). (a) Hessian spectrum of eigenvalues. The Hessian is shown to be full-rank at optimality for depths below the constrained overparametrization depth. Except at optimality in the overparametrized regime, all directions of the span of the *Ansatz* are able to be reached during optimization with respect to a target unitary. (b) Quantum Fisher information spectrum of eigenvalues. The quantum Fisher information is shown to be full-rank at optimality for depths below the overparametrization depth. It then saturates at the dynamical Lie algebra dimension for depths above the overparametrization depth, indicating the capabilities of the *Ansatz*, independent of the specific target unitary.

with derivatives with respect to θ of

$$\partial_\mu \tilde{\mathcal{L}}_\theta^U = \frac{1}{2} \frac{1}{1 - \tilde{\mathcal{L}}_\theta^U} \partial_\mu \mathcal{L}_\theta^U, \tag{B14}$$

$$\partial_{\mu\nu} \tilde{\mathcal{L}}_\theta^U = \frac{1}{2} \frac{1}{1 - \tilde{\mathcal{L}}_\theta^U} \left(\partial_{\mu\nu} \mathcal{L}_\theta^U + 2 \partial_\mu \mathcal{L}_\theta^U \partial_\nu \mathcal{L}_\theta^U \right), \tag{B15}$$

such that the optimal proper quantities are identical up to constant scalings to the improper quantities

$$\tilde{\mathcal{L}}_{\theta*}^U = 0, \quad \partial_\mu \tilde{\mathcal{L}}_{\theta*}^U = 0, \quad \partial_{\mu\nu} \tilde{\mathcal{L}}_{\theta*}^U = \frac{1}{2} \partial_{\mu\nu} \mathcal{L}_{\theta*}^U. \tag{B16}$$

Therefore, the proper Fisher information metric definition is identical up to a constant scaling to the improper definition

$$\tilde{\mathcal{F}}_{\theta_{\mu\nu}}^U = \frac{1}{2} \mathcal{F}_{\theta_{\mu\nu}}^U. \tag{B17}$$

For example, when the trace is over $d = 1$ states, $|\rho_\theta\rangle = U_\theta |\sigma\rangle$ from an initial fixed state $|\sigma\rangle$, or equivalently the unitaries are projected onto a $d = 1$-dimensional subspace, the Fisher information reduces to the standard definition

$$\mathcal{F}_{\theta_{\mu\nu}}^\rho = \text{Re}(\langle \partial_\mu \rho_\theta | \partial_\nu \rho_\theta \rangle - \langle \rho_\theta | \partial_\mu \rho_\theta \rangle \langle \partial_\nu \rho_\theta | \rho_\theta \rangle). \tag{B18}$$

2. Numerical overparametrization

We now investigate the effects of overparametrization numerically via the quantum Fisher information $\mathcal{F}_\theta$ and the objective Hessian $\mathcal{H}_\theta$. Here, we optimize $N = 4$ qubit, constrained, noiseless unitary compilation tasks. For the NMR *Ansatz* there are $D^2 - 2 = 2$ variable parameters per time step, meaning there are $P = 2M$ variable parameters per *Ansatz*. Given the universal *Ansatz*, $G = D^N - 1$, we observe in Fig. 6 that the expected rank saturation occurs at $R = P = G = 255$ for $N = 4$ for the Fisher information, and full rank saturation occurs at likely $R = P \gg G$ for the Hessian.

As per previous studies [28], the Fisher metric exhibits saturation behavior and indicates overparametrization at any point in the objective or parameter landscape. However, the

Hessian rank does not saturate, and remains full rank at this point in the landscape achieved by the optimizer. This point in the constrained landscape, even for $P \geqslant G$, is likely not optimal, and therefore not saturating the Hessian rank. This nonoptimality is attributed to the previous studies indicating that $M \sim O(1000)$ depth is necessary for this constrained *Ansatz* to achieve infidelities close to machine precision. Due to the quadratic scaling of the memory requirements for computing these $P \times P$ dimensional matrices, plus determining their spectrum, only $M \sim O(600)$ are currently feasible to compute in the current implementation.

We should note that for finite machine precision simulations, there is not a definitive method of determining the rank or number of nonzero eigenvalues of matrices. Here we choose the heuristic when there is an obvious visual distinction between the set of zero and nonzero spectra. In the case in which there is not an obvious cutoff, we choose a relative precision of $\lambda/\lambda_{\max} > P\varepsilon$ for P parameters and machine precision ε. It remains an interesting question whether there is a more principled and physics-informed approach for determining the rank.

APPENDIX C: BEHAVIOR OF MULTIPLE LAYER NOISE CHANNELS

In these Appendixes, we investigate properties of quantum channels consisting of parametrized unitary layers interlaced with noise. We show that many local noise models form a binomial distribution over the number of errors or nonidentity noise operations applied throughout the layers. We perform additional numerical studies to show differences in the behavior of infidelities with depth for unital versus nonunital noise. We extract from piecewise fitting the scaling of the noise-induced critical depth for infidelities, to be logarithmic in the noise scale. Finally, we perform analytical calculations of the leading-order scaling, with depth and noise scale, of infidelities, impurities, entropies, and relative entropy

divergences. These scalings match the numerical studies in this work exactly, and confirm that the divergent regime of optimization is driven by entropic effects.

There are many choices in the exact noise model $\mathcal{N}_\gamma$ used, and whether the noise acts globally or locally, both temporally and spatially. We assume that all noise acts independently in time and locally on each qubit in this work. The total noise channel is chosen to take the form of temporally local noise

$$\Lambda_{\theta\gamma} = \circ_m^M \Lambda_{\theta\gamma}^{(m)}, \tag{C1}$$

where the unitary and nonunitary components are separable,

$$\Lambda_{\theta\gamma}^{(m)} = \mathcal{N}_\gamma^{(m)} \circ \mathcal{U}_\theta^{(m)}. \tag{C2}$$

For this work, we also consider spatially local noise models to represent the nonunitary components of our channels,

$$\mathcal{N}_\gamma^{(m)} = \circ_i^N \mathcal{N}_{\gamma_i}^{(m)}, \tag{C3}$$

which act identically, and independently, on qubit i at time step m, with noise scale $\gamma_i^{(m)} = \gamma$.

We will discuss noise models in terms of the number of errors, or nonidentity operations that they apply to the evolution. The temporally (and spatially across qubits) noise models in this work imply that there are

$$K = NM \tag{C4}$$

independent sites of possible errors at each site in time and in space.

As a key assumption for our analysis, we explicitly assume that the local nonunitary components of the channel can be written as a convex combination of an identity channel and a nontrivial (nonidentity) error channel

$$\mathcal{N}_\gamma \equiv (1 - \gamma)\mathcal{I} + \gamma\mathcal{K}_\gamma. \tag{C5}$$

Here, there is a scale γ of a nontrivial operation $\mathcal{K}_\gamma$ applied, causing the evolution to deviate from being strictly unitary, resulting in what we refer to as an error. The nontrivial error channel is normalized identically as $\mathcal{N}_\gamma$. If the explicit noise model $\mathcal{N}_\gamma$ does not strictly contain an identity Kraus operator, for example if it is nonunital, then the nontrivial error is defined as

$$\mathcal{K}_\gamma = \frac{1}{\gamma}\mathcal{N}_\gamma - \frac{1 - \gamma}{\gamma}\mathcal{I}. \tag{C6}$$

This noise model for the nonunitary component of the channel forms a binomial distribution over K possible errors defined by $\mathcal{K}_\gamma$. Errors can occur at spatial and temporal sites $i \in [N]$ and $m \in [M]$, and we use the multi-index $\chi_k = \{\chi_{k_i}^m \in [2], k \in [K]\}$ as an indicator function for where the $|\chi_k| = k \leqslant K$ errors occur across the sites. Given this binomial distribution description, and using the binomial expansion of the noise scale factors, the channel may be written as a convex combinations of k-error, or at most k-error channels

$$\Lambda_{\theta\gamma} = \sum_k^K \binom{K}{k}\gamma^k(1 - \gamma)^{K-k}\Lambda_{\theta\gamma_k} \tag{C7}$$

$$= \sum_k^K \sum_l^{K-k} \binom{K}{k}\binom{K-k}{l}(-1)^l\gamma^{k+l}\Lambda_{\theta\gamma_k} \tag{C8}$$

$$= \sum_k^K \sum_{l=k}^K \binom{K}{k}\binom{K-k}{l-k}(-1)^{l-k}\gamma^l\Lambda_{\theta\gamma_k} \tag{C9}$$

$$= \sum_k^K \sum_l^k \binom{K}{k}\binom{k}{l}(-1)^{k-l}\gamma^k\Lambda_{\theta\gamma_l} \tag{C10}$$

$$= \sum_k^K \binom{K}{k}\gamma^k\left[\sum_l^k \binom{k}{l}(-1)^{k-l}\Lambda_{\theta\gamma_l}\right] \tag{C11}$$

$$= \sum_k^K \binom{K}{k}\gamma^k\Lambda_{\theta\gamma_{\leqslant k}}. \tag{C12}$$

We denote trace-preserving channels with k nontrivial errors as the uniform convex combination of all possible locations of the $k \leqslant K$ errors,

$$\Lambda_{\theta\gamma_k} = \frac{1}{\binom{K}{k}}\sum_{\chi_k}\Lambda_{\theta\gamma}^{\chi_k}, \tag{C13}$$

and we denote traceless operators with at most k nontrivial errors as the nonuniform combination of all possible locations of the $l \leqslant k$ errors,

$$\Lambda_{\theta\gamma_{\leqslant k}} = \sum_l^k (-1)^{k-l}\binom{k}{l}\Lambda_{\theta\gamma_l}, \tag{C14}$$

where specific k-error channels with k nontrivial errors at indices χ_k are denoted as

$$\Lambda_{\theta\gamma}^{\chi_k} = \circ_m^M\left[\left[\circ_i^N \mathcal{K}_{\gamma_i}^{(m)\chi_{k_i}^m}\right] \circ \mathcal{U}_\theta^{(m)}\right]. \tag{C15}$$

Therefore, to leading order, the noisy and noiseless channels differ as per the nontriviality $\mathcal{K}_{\gamma_i}^{(m)} \neq \mathcal{I}_i$ of K possible local errors,

$$\Lambda_{\theta\gamma} - \Lambda_\theta = \left(\sum_m^M \sum_i^N \mathcal{U}_\theta^{(>m)} \circ \mathcal{K}_{\gamma_i}^{(m)} - \mathcal{I}_i \circ \mathcal{U}_\theta^{(\leqslant m)}\right)\gamma$$
$$+ O\left(\binom{K}{2}\gamma^2\right). \tag{C16}$$

1. Noise models

In this work, we consider several noise models $\mathcal{N}_\gamma = (1 - \gamma)\mathcal{I} + \gamma\mathcal{K}_\gamma$ that are relevant to existing quantum devices, namely independent, local, dephasing, amplitude damping, and depolarizing noise. These noise models all belong to the class of unital or nonunital Pauli noise that transform local Pauli operators $\mathcal{P}_D$. For the case of $D = 2$ qubits, $\mathcal{P}_2 = \{I, Z, X, Y\}$. The forms of the models are identified by their nontrivial error component $\mathcal{K}_\gamma$ and are described as follows.

Unital local dephasing noise for all inputs may be written as

$$\mathcal{K}_\gamma{}^{\text{dephase}}(\cdot) = Z \cdot Z^\dagger, \tag{C17}$$

where the nontrivial channel is unitary.

Unital local depolarizing noise for all inputs may be written as

$$\mathcal{K}_\gamma{}^{\text{depolarize}}(\cdot) = \frac{\text{tr}(\cdot)}{D}I, \tag{C18}$$

where the nontrivial channel is maximally depolarizing.

Nonunital amplitude damping noise for Pauli inputs $P \in \mathcal{P}_2$ may be written as

$$\mathcal{K}_\gamma{}^{\text{amplitude}}(P) = \begin{cases} P + \gamma Z, & P \in \{I\}, \\ (1-\gamma)P, & P \in \{Z\}, \\ \sqrt{1-\gamma}P, & P \in \mathcal{P}_2 \backslash \{I, Z\}, \end{cases} \quad \text{(C19)}$$

and the nontrivial channel is complicated by the lack of an inherent identity Kraus operator.

In the case of the nontrivial components of the noise channel being single unitary operators V such that $\mathcal{K}_\gamma(\cdot) = V \cdot V^\dagger$, each k-error channel is unitary, and the original unitary channel is interlaced with unitaries V at indices χ_k where errors occur.

In the case of depolarizing noise where $\mathcal{K}_\gamma(\cdot) = [\text{tr}(\cdot)/D]\mathcal{I}$, partial traces remove any information about the state at the local indices in χ_k where errors occur.

2. Probabilistic interpretation of noise channels

Within this formalism, the channel can be represented as an expectation over a distribution of the number of $k \leqslant K$ possible errors,

$$\Lambda_{\theta\gamma} = \langle \Lambda_{\theta\gamma_k} \rangle_{k \sim p_{K\gamma}}, \quad \text{(C20)}$$

where for moments of channels with k nontrivial errors,

$$\Lambda_{\theta\gamma_k} = \frac{1}{\binom{K}{k}} \sum_{\chi_k} \Lambda_{\theta\gamma}^{\chi_k}. \quad \text{(C21)}$$

The exact distribution with this local noise model used in this work is the binomial distribution

$$p_{K\gamma}(k) = p_{K\gamma}^{\text{Binomial}}(k) = \binom{K}{k}\gamma^k(1-\gamma)^{K-k}, \quad \text{(C22)}$$

which has mean and variance

$$\mu_{K\gamma} = K\gamma, \quad \Sigma_{K\gamma} = K\gamma(1-\gamma). \quad \text{(C23)}$$

Such interpretations are also used in error mitigation approaches [63]. In the limit $K \to \infty$, this distribution tends to the Gaussian distribution

$$p_{K\gamma}(k) \to p_{K\gamma}^{\text{Gaussian}}(k) = \sqrt{\frac{1}{2\pi\Sigma_{K\gamma}}} e^{-\frac{1}{2}\frac{(k-\mu_{K\gamma})^2}{\Sigma_{K\gamma}}}, \quad \text{(C24)}$$

which can be shown through the De Moivre–Laplace theorem [70], a central-limit version of the standardized binomially distributed variables $\Sigma_{K\gamma}^{-1/2}(k - \mu_{K\gamma})$.

In the limit $K \to \infty$, $\gamma \to 0$, and the finite limit $K\gamma \to \lambda_{K\gamma}$, the distribution tends to the Poisson distribution

$$p_{K\gamma}(k) \to p_{K\gamma}^{\text{Poisson}}(k) = \frac{1}{k!}(K\gamma)^k e^{-K\gamma}, \quad \text{(C25)}$$

which can be shown from using Sterling's approximation for the binomial coefficient, and equating the generating functions for the distributions.

3. Noisy state preparation

We conduct studies in Fig. 7 of the noise and depth dependence on the optimal infidelities for each of the unital dephasing and depolarizing and nonunital amplitude damping noise models. For unital noise, we observe that the critical depth phenomena seem to be consistent across noise models at all noise scales, and entropic effects dominate. For nonunital noise, we observe that the critical depth phenomena seem to be consistent with unital noise models at all small noise scales. However, for large noise scales, nonunital noise forces the state into a specific (pure) state, which dominates over entropic effects. The infidelities then appear to converge to nonunity values or potentially decrease polynomially to zero. The unitary component of the channel appears to be able to slowly rotate this noise-induced pure state towards the correct target pure state. In both unital and nonunital noise models, inserting parameters learned in the noisy setting into the corresponding noiseless unitary *Ansatz* indicates that the noiseless infidelities have greater resilience to increasing noise scales. This suggests the underlying unitary transformations are being learned in the noisy settings, however with greater variance.

4. Noise-induced critical depth

To develop a relationship between noise and an induced critical depth beyond which infidelities no longer converge exponentially, we perform piecewise fits, as per Fig. 8. For each noise scale, we perform exponential and then polynomial fits, respectively, before and after an approximate location of the critical depth suggested by the finite amount of data points available. We are then able to approximate where the piecewise curves intersect, indicating the location of the critical depth as a function of noise. Plotting this relationship suggests a logarithmic relationship between noise and critical depth. This leads to the optimal infidelity scaling approximately polynomially, between linearly and quadratically with noise, and in agreement with previous conjectures about these relationships [35].

5. Parameter shift rules for quantum channels

Here, we derive parameter shift rules for the gradients of multiple layer quantum channels. This generalizes previous results, and provides some explanation for why the optimization routines in noisy settings generally appear to converge to the noiseless optima at small noise scales. Given our composite channel definition over M layers, written decomposed into channels before and after an index (μ, m), our state preparation objectives of the trace infidelity with respect to a pure state ρ and initial state σ may be written as

$$\mathcal{L}_{\theta\gamma} = 1 - \text{tr}\big(\mathcal{U}_\mu^{(m)}\big(\sigma_{<\mu\gamma}^{(<m)}\big)\mathcal{U}_{>\mu}^{(m)\dagger}\big[\mathcal{N}_\gamma^{(m)\dagger}\big(\rho_{>\mu\gamma}^{(>m)}\big)\big]\big). \quad \text{(C26)}$$

These objectives are crucially linear in the states, and therefore gradients of the objectives with respect to parameters at the index (μ, m) are linear functions of the gradients of the state.

Let $\mathcal{U}_\theta$ be a single parameter unitary channel, with the corresponding unitary operator $U_\theta = e^{-i\theta G}$, where the Hermitian generator is involutory up to a factor $G^2 = \zeta^2 I$. The gradient of such a unitary channel for an arbitrary input follows the parameter shift rule

$$\partial_\theta \mathcal{U}_\theta = -i[G, \mathcal{U}_\theta] = \zeta(\mathcal{U}_{\theta+\varphi} - \mathcal{U}_{\theta-\varphi}), \quad \text{(C27)}$$
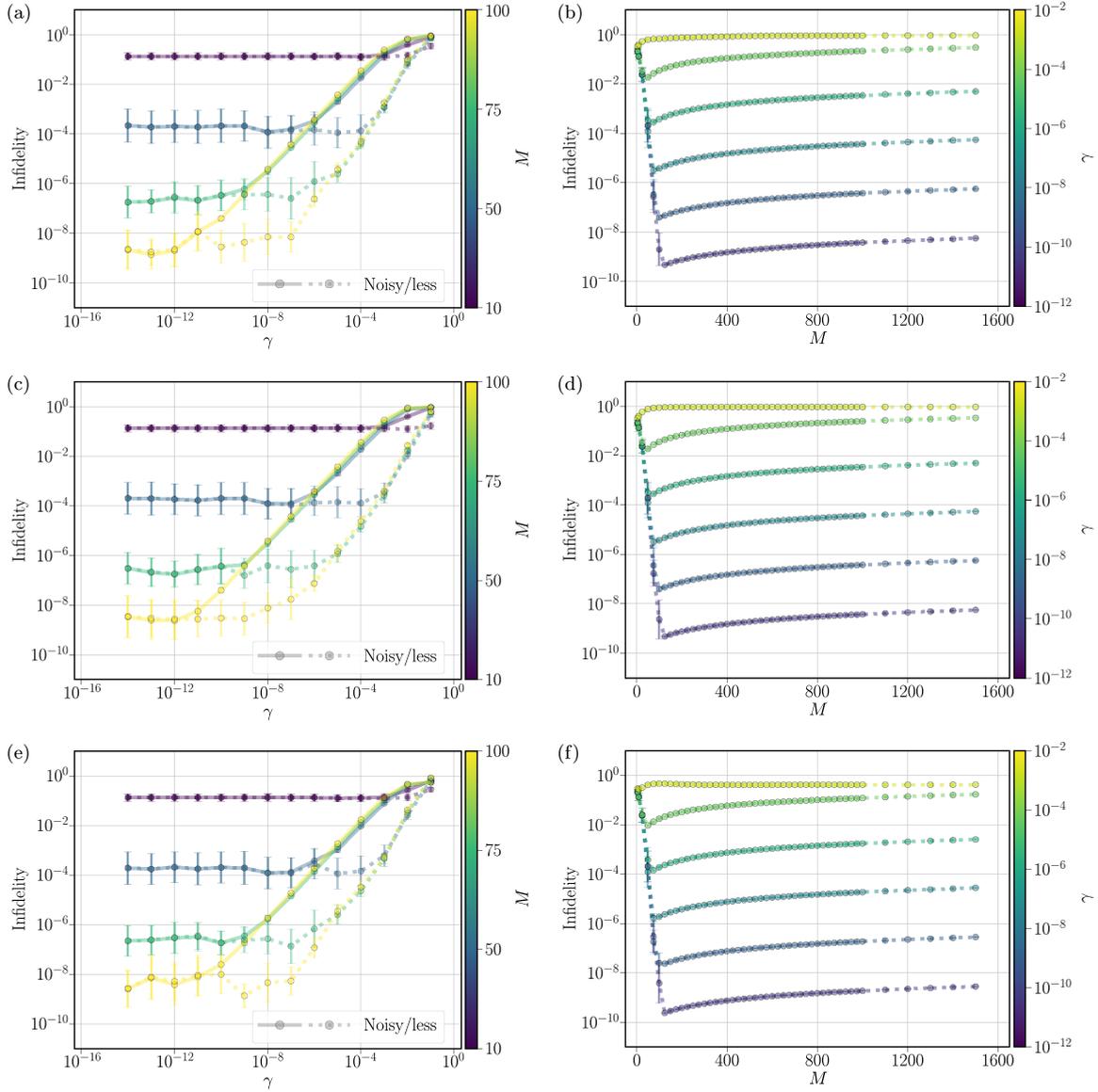
FIG. 7. Behavior of infidelity objectives with respect to noise and depth for the $N = 4$ NMR *Ansatz*. (a),(b) Dephasing, (c),(d) depolarizing, (e),(f) amplitude damping. Left: Trained noisy infidelity (solid) and tested infidelity of noisy parameters in noiseless *Ansatz* (dashed) for various depths M (colored/gradient). Infidelities of noisy *Ansatz* are shown to be depth-dependent and noise-independent for small noise scales, before universally increasing polynomially with noise. Inserting parameters learned in the noisy setting into an identical noiseless *Ansatz* indicates that the underlying unitary evolution is being learned, and is resilient to noise. Right: Critical depth for noisy infidelity for various noise scales γ (colored/gradient). Infidelities are shown to improve exponentially with depth, up until a noise-induced critical depth, where entropic effects worsen infidelities polynomially with depth. Nonunital noise is also shown to be less dominated be entropic effects at large noise scales.

where we denote $\varphi = \pi/4\zeta$, and other choices of φ are also possible depending on experimental feasibility. This can be extended to the k-order gradient of this channel as the k-order nested commutator

$$\partial_\theta^k \mathcal{U}_\theta = (-i)^k [G, \mathcal{U}_\theta]_k = \zeta^k \sum_l^k (-1)^l \binom{k}{l} \mathcal{U}_{\theta+(k-2l)\varphi}. \quad (C28)$$

Unitary channels with generators with more complicated spectra can be expressed as a linear combination of unitary channels, with perturbative angles φ, weighted by coefficients α_φ,

$$\partial_\theta^k \mathcal{U}_\theta = \sum_\varphi \alpha_\varphi \mathcal{U}_{\theta+\varphi}. \quad (C29)$$

Thus gradients of general parametrized channels for constant noise scales γ have the form

$$\partial_\mu \Lambda_{\theta\gamma} = \sum_\varphi \alpha_{\varphi_\mu} \Lambda_{\theta+\varphi_\mu \gamma}, \quad (C30)$$

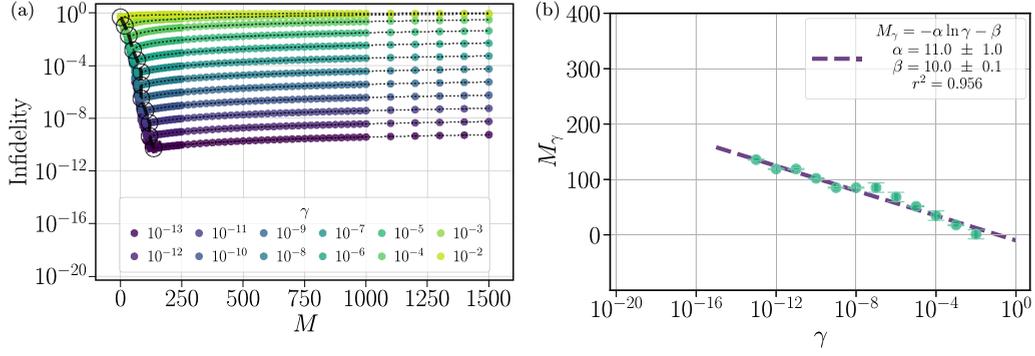and similarly any linear objectives such as $\partial_\mu^{(m)} \mathcal{L}_{\theta\gamma}$ have an identical form.

FIG. 8. Fit of noise-induced critical depth. (a) Piecewise fits for various noise scales γ (colored/gradient). Objectives decrease exponentially with depth for depths less than the critical depth (black circles), and objectives increase polynomially with depth for depths greater than the critical depth. (b) Uncertainty propagation is used to estimate the error of the linear-log fit (dashed line) between the estimated critical depths where the piecewise curves intersect, and the associated noise scales.

6. Relationship between infidelity, impurity, and von-Neumann entropy

To better understand the phenomena dictating the behavior of infidelities with respect to noise scales and depth, we also investigate the scaling of the impurity and entropy of the noisy parametrized states. For completeness, we restate the impurity, von-Neumann entropy, conditional entropy, and relative entropy divergence between a d-dimensional state ρ and a pure state ρ', as

$$\mathcal{L}'_\rho = 1 - \mathrm{tr}(\rho'\rho), \tag{C31}$$

$$\mathcal{I}_\rho = 1 - \mathrm{tr}(\rho^2), \tag{C32}$$

$$\mathcal{S}_\rho = -\mathrm{tr}(\rho\ln(\rho))/\ln(d), \tag{C33}$$

$$\mathcal{S}^{\rho'}_\rho = -\mathrm{tr}(\rho'\ln(\rho))/\ln(d), \tag{C34}$$

$$\mathcal{D}^{\rho'}_\rho = \mathcal{S}^{\rho'}_\rho - \mathcal{S}_{\rho'}. \tag{C35}$$

We also note that when we use the definitions of parametrized states $\rho_{\theta\gamma}$ and target reference states ρ in place of ρ and ρ', we instead use the subscripts $\theta\gamma$ and superscripts ρ for these quantities.

For this analysis, we use the Bloch representation [71], which describes operators in terms of a trace orthogonal basis $\mathcal{P}_d$, such that $\mathrm{tr}(\alpha^\dagger\beta) = d\delta_{\alpha\beta}$ for all operators $\alpha, \beta \in \mathcal{P}_d$. This basis contains the identity, is of size $|\mathcal{P}_d| = d^2$, and may represent an algebra, with structure constants defining their commutation relations. Let $\omega = \{P : P \in \mathcal{P}_d \backslash \{I\}\}$ represent the vector of all nonidentity operators such that quantum states with unit traces have the form

$$\rho = \frac{I + \lambda \cdot \omega}{d}. \tag{C36}$$

Here the $|\mathcal{P}_d| - 1$ Bloch coefficients λ fully describe quantum states. These coefficients are constrained to represent positive-semidefinite operators, and their magnitude is bounded by the pure-state boundary described by

$$\lambda^2 \leqslant d - 1. \tag{C37}$$

The action of arbitrary trace-preserving channels on quantum states can further be described by the affine linear transformation on the coefficients,

$$\Lambda : \lambda \to \Gamma\lambda + \upsilon. \tag{C38}$$

This affine linear transformation represents a rotation and scaling of λ by Γ, plus a translation to a different axis by υ, and it can be thought of as a transformation on the generalized d-dimensional Bloch sphere of radius $d - 1$. The primary constraints on the transformations are to remain within this boundary such that $(\Gamma\lambda + \upsilon)^2 \leqslant d - 1$.

If the channel is unital, then the transformation is strictly linear and $\upsilon = 0$, otherwise the channel is nonunital. Finally, if the channel is unitary, $\Gamma = u$ is an orthogonal transformation that preserves the length of λ^2.

We also note a useful identity when λ is associated with a pure state with zero entropy, given the expansion for the matrix logarithm

$$\ln(I + \lambda \cdot \omega) = \sum_{k>0} \frac{(-1)^{k+1}}{k}(\lambda \cdot \omega)^k \tag{C39}$$

and the tracelessness of $\lambda \cdot \omega$, meaning

$$\frac{1}{d\ln(d)} \sum_{k>1} \frac{(-1)^k}{k(k-1)} \mathrm{tr}((\lambda \cdot \omega)^k) = 1. \tag{C40}$$

This identity avoids complicated expressions for powers of Bloch vectors [72].

We may now define quantities in terms of the Bloch coefficients. Similarity between quantum states ρ, ρ' can be described either by their trace inner products

$$\mathrm{tr}(\rho\rho') = \frac{1 + \lambda \cdot \lambda'}{d} = 1 - \mathcal{L}^{\rho'}_\rho, \tag{C41}$$

or by their cosine similarity with an angle $\phi^{\rho'}_\rho$ between their coefficients,

$$\cos\phi^{\rho'}_\rho = \frac{\lambda \cdot \lambda'}{\sqrt{\lambda^2\lambda'^2}}$$

$$= \frac{d(1 - \mathcal{L}^{\rho'}_\rho) - 1}{\sqrt{[d(1 - \mathcal{I}_\rho) - 1][d(1 - \mathcal{I}_{\rho'}) - 1]}}. \tag{C42}$$

Similarly for other functions of interest with respect to a pure state ρ',

$$\mathcal{L}_\rho^{\rho'} = 1 - \frac{1 + \lambda \cdot \lambda'}{d}, \tag{C43}$$

$$\mathcal{I}_\rho = 1 - \frac{1 + \lambda^2}{d}, \tag{C44}$$

$$\mathcal{S}_\rho = 1 - \frac{1}{d \ln(d)} \sum_{k>1} \frac{(-1)^k}{k(k-1)} \text{tr}((\lambda \cdot \omega)^k)$$

$$= 1 - \frac{1}{2} \frac{1}{\ln(d)} \lambda^2 + O(\lambda^3), \tag{C45}$$

$$\mathcal{D}_\rho^{\rho'} = 1 - \frac{1}{d \ln(d)} \sum_{k>0} \frac{(-1)^{k+1}}{k} \text{tr}((I + \lambda' \cdot \omega)(\lambda \cdot \omega)^k)$$

$$= 1 - \frac{1}{2} \frac{1}{\ln(d)} \left(2 \frac{\lambda \cdot \lambda'}{\lambda^2} - 1\right) \lambda^2 + O(\lambda^3). \tag{C46}$$

We can also define the generalized α-Renyi infidelities, impurities, entropies, and divergences for $\alpha \geqslant 1/2$,

$$\mathcal{L}_\rho^{\rho'(\alpha)} = \text{tr}\left(\left(\rho'^{\frac{1-\alpha}{2\alpha}} \rho \rho'^{\frac{1-\alpha}{2\alpha}}\right)^\alpha\right), \tag{C47}$$

$$\mathcal{I}_\rho^{(\alpha)} = 1 - \text{tr}(\rho^\alpha), \tag{C48}$$

$$\mathcal{S}_\rho^{(\alpha)} = \frac{1}{\alpha - 1} \frac{1}{\ln(d)} \ln\left(1 - \mathcal{I}_\rho^{(\alpha)}\right), \tag{C49}$$

$$\mathcal{D}_\rho^{\rho'(\alpha)} = \frac{1}{\alpha - 1} \frac{1}{\ln(d)} \ln\left(1 - \mathcal{L}_\rho^{\rho'(\alpha)}\right). \tag{C50}$$

Using these definitions, we can show [73] that the entropies and divergences are monotonically increasing with decreasing α, and we can identify $\mathcal{D}_\rho^{\rho'}$ with $\alpha \to 1$ and related divergences and infidelities with $\mathcal{D}_\rho^{\rho'(1/2)} = -2\ln(1 - \mathcal{L}_\rho^{\rho'})/\ln(d)$. Based on these relationships, we have the bounds relating the conditional entropy and the infidelity,

$$\mathcal{S}_\rho^{\rho'} \geqslant \frac{2}{\ln(d)} \mathcal{L}_\rho^{\rho'}, \tag{C51}$$

however we have not found any known similar bounds relating entropy and infidelity or impurity.

Given our previous expressions for channels consisting of layers of nonunitary noise with scale γ, interlaced by unitary

channels with parameters θ, resulting in a binomial distribution over states with at most K errors, we may represent such parametrized quantum channels in this formalism as

$$\Lambda_{\theta\gamma} : \Gamma_{\theta\gamma} = (1 - \gamma_K)u_\theta + \gamma_K u_{\theta\gamma}, \quad \upsilon_{\theta\gamma} = \gamma_K \eta_{\theta\gamma}. \tag{C52}$$

Here, we have explicitly separated the noiseless unitary rotation from the other nonunitary and nonunital transformations, and we defined the error-dependent noise scale as

$$(1 - \gamma_K) = (1 - \gamma)^K. \tag{C53}$$

Therefore, the transformed coefficients, given an initial state σ with coefficients ξ, are transformed by unitary and noise components of the transformation

$$\lambda_{\theta\gamma} = (1 - \gamma_K)u_\theta \xi + \gamma_K u_{\theta\gamma} \xi + \gamma_K \eta_{\theta\gamma} \tag{C54}$$

$$= (1 - \gamma_K)\lambda_{\theta|\gamma} + \gamma_K \varepsilon_{\theta\gamma}. \tag{C55}$$

Here, we decompose the transformed state into what we refer to as the pure and mixed components. The pure component of the state may be written in terms of a reference pure state ρ with coefficients λ, and a pure state with orthogonal coefficients $\zeta \perp \lambda$ as

$$\lambda_{\theta|\gamma} = u_\theta \xi = (1 - \alpha_{\theta|\gamma})\lambda + \beta_{\theta|\gamma} \zeta, \tag{C56}$$

where the coefficients implicitly contain a dependence on the noise from optimization, and they are constrained such that $(1 - \alpha_{\theta|\gamma})^2 + \beta_{\theta|\gamma}^2 = 1$. The mixed components of the state can be written in terms of its unital and nonunital components,

$$\varepsilon_{\theta\gamma} = u_{\theta\gamma} \xi + \eta_{\theta\gamma}. \tag{C57}$$

This decomposition allows us to understand how the unitary and noise components transform the state. The unitary component, independent of the noise, rotates the state into the pure components parallel and perpendicular to ρ. The noise component then scales both the pure and mixed components with γ_K, and performs an affine translation with $\eta_{\theta\gamma}$. In the limit that $\gamma \to 0$, and assuming converged optimization to the optimal $\theta = \theta^*$ such that $\rho_{\theta^*} \to \rho$, the coefficients should then reduce to $\alpha_{\theta^*}, \beta_{\theta^*} \to 0$, and the nonunital affine translation should also vanish, $\eta_\theta \to 0$.

Due to the nontrivial optimization, we do not have a general closed-form expression depicting the θ, γ, K dependence of the pure state components $\alpha_{\theta|\gamma}, \beta_{\theta|\gamma}$ or the mixed component $\varepsilon_{\theta\gamma}$. However, by expanding out the quantities of interest in terms of these expressions for the coefficients $\lambda_{\theta\gamma}$, we can obtain the leading-order scaling of quantities in terms of K, γ.

The inner products of the coefficients are

$$\lambda_{\theta\gamma}^2 = (1 - \gamma_K)^2 \lambda^2 + \gamma_K^2 \varepsilon_{\theta\gamma}^2 + 2\gamma_K(1 - \gamma_K)^2 [(1 - \alpha_{\theta|\gamma})\lambda + \beta_{\theta|\gamma}\zeta] \cdot \varepsilon_{\theta\gamma} \tag{C58}$$

$$= \lambda^2 - 2K\gamma\left(1 - (1 - \alpha_{\theta|\gamma})\frac{\lambda \cdot \varepsilon_{\theta\gamma}}{\lambda^2} - \beta_{\theta|\gamma}\frac{\zeta \cdot \varepsilon_{\theta\gamma}}{\lambda^2}\right)\lambda^2 + O\left(\binom{K}{2}\gamma^2\right), \tag{C59}$$

$$\lambda_{\theta\gamma} \cdot \lambda = (1 - \gamma_K)(1 - \alpha_{\theta|\gamma})\lambda^2 + \gamma_K \lambda \cdot \varepsilon_{\theta\gamma} \tag{C60}$$

$$= \lambda^2 - \alpha_{\theta|\gamma}\lambda^2 - K\gamma\left(1 - \alpha_{\theta|\gamma} - \frac{\lambda \cdot \varepsilon_{\theta\gamma}}{\lambda^2}\right)\lambda^2 + O\left(\binom{K}{2}\gamma^2\right). \tag{C61}$$
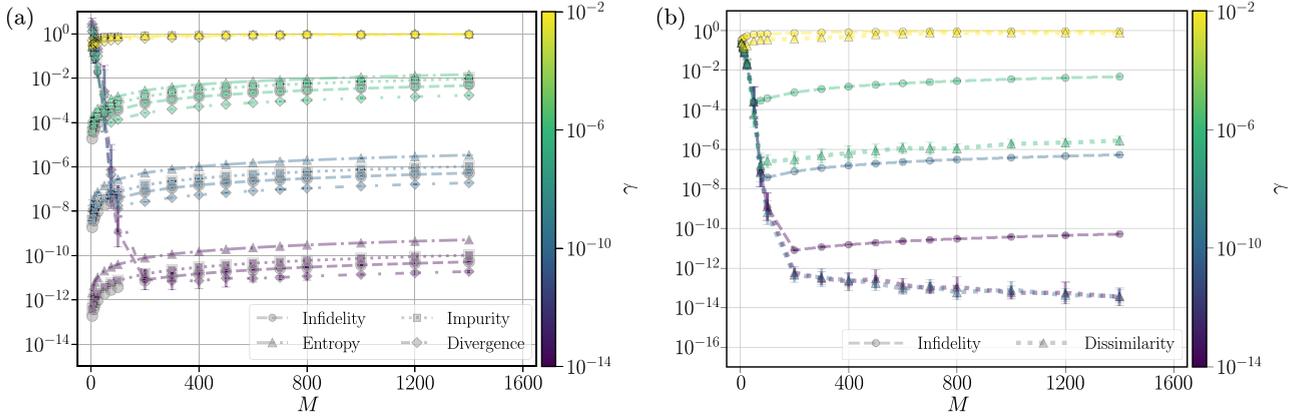
FIG. 9. Behavior of infidelity, impurity, entropy, relative entropy divergence, and cosine dissimilarity of parametrized states relative to pure target states, with respect to dephasing noise γ (colored/gradient) and depth for the $N = 4$ NMR *Ansatz*. (a) Impurity, entropy, relative entropy divergence. Once the optimized parametrized *Ansatz* achieves infidelities or relative entropy divergences of the order of the impurity and entropies, the system is dominated by entropic effects. In this divergent regime, all quantities (colored markers) scale identically, and leading-order infidelity and impurity analytical values (enlarged gray markers) show exact agreement. (b) Cosine dissimilarity. The unitary component of the channel is able to align the parametrized state with the target state. For sufficiently low noise scales, this dissimilarity remains constant, and infidelities increase in the divergent regime strictly due to being scaled by the depth-dependent noise scale.

Powers of Bloch vectors in terms of the target coefficients, even when $\alpha_{\theta|\gamma} = \beta_{\theta|\gamma} = 0$, scale with γ as

$$(\lambda_{\theta\gamma} \cdot \omega)^k = (1 - \gamma_K)^k (\lambda \cdot \omega)^k + O(\gamma). \qquad \text{(C62)}$$

All powers k of Bloch vectors, such as those that occur in infinite series expansions for logarithmic functions appearing in entropies, therefore contribute leading-order terms in K, γ. It does not then suffice to retain only some $O(\lambda^k)$-order terms in expansions of such functions to capture the coefficients of their leading-order behavior with γ. Obtaining these coefficients to all orders, for all dimensions d, is possible, however it requires opaque, recursive expressions for products $\mathrm{tr}((\lambda \cdot \omega)^k)$ in terms of the structure constants, such as those found for the Pauli basis by Sarkar [60].

Therefore, to leading order in K, γ, the quantities of interest show similar scaling with overlaps of the parametrized state with the target pure state,

$$\mathcal{L}^{\rho}_{\theta\gamma} = \frac{d-1}{d} \alpha_{\theta|\gamma} + K\gamma \frac{d-1}{d} \left((1 - \alpha_{\theta|\gamma}) - \frac{\lambda \cdot \varepsilon_{\theta\gamma}}{\lambda^2} \right)$$
$$+ O\left(\binom{K}{2}\gamma^2\right), \qquad \text{(C63)}$$

$$\mathcal{I}_{\theta\gamma} = 2K\gamma \frac{d-1}{d} \left(1 - (1 - \alpha_{\theta|\gamma}) \frac{\lambda \cdot \varepsilon_{\theta\gamma}}{\lambda^2} - \beta_{\theta|\gamma} \frac{\zeta \cdot \varepsilon_{\theta\gamma}}{\lambda^2} \right)$$
$$+ O\left(\binom{K}{2}\gamma^2\right), \qquad \text{(C64)}$$

$$\mathcal{S}_{\theta\gamma} = K\gamma \frac{d-1}{\ln(d)} \left(1 - (1 - \alpha_{\theta|\gamma}) \frac{\lambda \cdot \varepsilon_{\theta\gamma}}{\lambda^2} - \beta_{\theta|\gamma} \frac{\zeta \cdot \varepsilon_{\theta\gamma}}{\lambda^2} \right)$$
$$+ O(\lambda^3_{\theta\gamma}) \sim O(K\gamma), \qquad \text{(C65)}$$

$$\mathcal{D}^{\rho}_{\theta\gamma} = \frac{d-1}{\ln(d)} \alpha_{\theta|\gamma} + K\gamma \frac{d-1}{\ln(d)} \left(\alpha_{\theta|\gamma} \left(1 + \frac{\lambda \cdot \varepsilon_{\theta\gamma}}{\lambda^2} \right) \right.$$
$$\left. -\beta_{\theta|\gamma} \frac{\zeta \cdot \varepsilon_{\theta\gamma}}{\lambda^2} \right) + O(\lambda^3_{\theta\gamma}) \sim O(K\gamma). \qquad \text{(C66)}$$

If the unitary component of the channel transforms the pure component of the state to exactly the target state such that $\alpha_{\theta|\gamma} = \beta_{\theta|\gamma} = 0$, then expressions simplify further. A hierarchy between quantities can be partially observed to leading order in terms of the overlap $\lambda \cdot \varepsilon_{\theta\gamma}$ between the mixed component of the state and the target state,

$$\mathcal{L}^{\rho}_{\theta\gamma} = K\gamma \frac{d-1}{d} \left(1 - \frac{\lambda \cdot \varepsilon_{\theta\gamma}}{\lambda^2} \right) + O\left(\binom{K}{2}\gamma^2\right), \qquad \text{(C67)}$$

$$\mathcal{I}_{\theta\gamma} = 2K\gamma \frac{d-1}{d} \left(1 - \frac{\lambda \cdot \varepsilon_{\theta\gamma}}{\lambda^2} \right) + O\left(\binom{K}{2}\gamma^2\right), \qquad \text{(C68)}$$

$$\mathcal{S}_{\theta\gamma} = O(K\gamma), \qquad \text{(C69)}$$

$$\mathcal{D}^{\rho}_{\theta\gamma} = O(K\gamma). \qquad \text{(C70)}$$

Plotting each of these quantities from numerical optimizations in Fig. 9 for unital dephasing noise, we observe that once the infidelities or divergences are less than the impurities or entropies, there is a transition from the convergent to the divergent regime. This transition occurs at a critical depth, before which increasing depth allows the dominant unitary component of the channel to rotate the state to converge to the target state. Beyond this critical depth, the noise component of the channel dominates, and the parametrized state scales towards the mixed state. This divergence of all quantities is shown numerically at small noise scales to be linear in depth and noise scale. We also plot the derived analytical leading-order scaling of the infidelity and impurities with gray enlarged markers, which are in excellent agreement with the numerical results. In this divergent regime at small noise scales, we also

observe the numerical hierarchy of the quantities

$$\mathcal{D}^{\rho}_{\theta\gamma} \leqslant \mathcal{L}^{\rho}_{\theta\gamma} \leqslant \mathcal{I}_{\theta\gamma} \leqslant \mathcal{S}_{\theta\gamma}, \tag{C71}$$

which is in agreement with our infidelity and impurity analytical results. All quantities also numerically appear to converge together as noise increases. In the case of unital noise, all quantities diverge to the worst-case unity values as the noise and depth increase. In the case of nonunital noise, quantities potentially converge to optimality polynomially with depth as the noise and depth increase. This contrasting behavior to the unital noise case suggests the nonunital terms $\eta_{\theta\gamma} \neq 0$ dominate the leading-order scaling in this regime.

We also note that the relative entropy divergences show identical convergent and divergent overparametrized regimes of optimization as the infidelities. The numerically found hierarchy of quantities also suggests that infidelity is potentially lower bounded by the divergence. This appears to be reasonable due to both quantities reflecting distances between distributions corresponding to the quantum states. Given that the divergence is used in classical learning tasks, it is also suitable as an objective function for quantum fidelity-based tasks, although it is more demanding to compute.

To assess the validity of this assumption of the unitary component aligning the pure component of the state with the target state, we note that for small noise scales, the cosine dissimilarity

$$1 - \left|\cos \phi^{\rho}_{\theta\gamma}\right| \sim O(|\alpha_{\theta|\gamma}|) \tag{C72}$$

is of order of the alignment of the pure component with the target state. Plots of the cosine dissimilarity in Fig. 9 indicate that for sufficiently low noise scales, the dissimilarity remains constant, or even decreases to machine precision scales. This suggests that divergences in the infidelities at large depths are strictly due to the noise component of the channel scaling the pure component by the depth-dependent noise scales. At larger noise scales, the cosine dissimilarity is larger, but still orders of magnitude smaller than the infidelities, and diverges similarly to the infidelities.

APPENDIX D: CLASSICAL AND QUANTUM ERROR ANALYSIS

In these Appendixes, we derive and compare the noise-induced bias of the parameterized noisy states from their noiseless values, for both classical and quantum noise. We show both noise types have biases that scale polynomially with the noise scale, and exponentially with the number of errors induced by the noise. Finally, we discuss the implications of classical floating point error on the viability of large-scale simulations without error mitigation.

To bound generally independent errors in simulations, we use Schatten norms $\|A\| = \|A\|_p, \quad p \in [1, \infty]$ for d-dimensional matrices $A, B \in \mathcal{M}(d)$. Such norms satisfy the convenient properties of monotonicity $\|A\|_p \leqslant \|A\|_q, \quad q \leqslant p$, subadditivity $\|A + B\|_p \leqslant \|A\|_p + \|B\|_p$, submultiplicativity $\|AB\|_p \leqslant \|A\|_p \|B\|_p$, and being invariant under unitaries $\|UAV^\dagger\|_p = \|A\|_p$ for $U, V \in \mathcal{U}(d)$. Unitary matrices have constant norm $\|U\|_p = d^{1/p}$ for $U \in \mathcal{U}(d)$, and we denote the $\lim_{p\to\infty} \|A\|_p = \lambda(A)$ norm as the largest singular value. Finally, Schatten norms satisfy Holder's $\|AB\|_s \leqslant$

$\|A\|_p \|B\|_q, \quad 1/p + 1/q = 1/s$, and von-Neumann $|\text{Tr}[AB]| \leqslant \|A\|_p \|B\|_q, \quad 1/p + 1/q = 1$ inequalities [53].

1. Classical error analysis

We first investigate the effect of classical floating point errors on scalar operations, generalizing the results of [74] to the case of an arbitrary number of successive matrix multiplications.

a. Scalar floating point error

Let us assume that we are performing binary floating point operations, with the exact, ideal operations denoted by $\circ \in \{+, -, \times, /\}$. We denote floating point representations of any exact operations $\circ$ with subscripts $\circ_\epsilon$.

For operations between scalars, and scalar values themselves, we assume they may only be represented with a relative error, upper bounded by ε, sometimes referred to as machine precision.

For vectorized operations comprised of vectors of dimension d, let the relative error be upper bounded by $\epsilon = \epsilon(d, \varepsilon)$. This dependence of the error on the number of operations d and the machine precision ε may be polynomial, or even exponential. This general definition includes the case of more sophisticated computer architectures that perform operations such as fused-addition-multiplications.

The scalar product between scalars x, y, when $d = 1$, therefore has error

$$x \circ_\epsilon y - x \circ y = (x \circ y)\epsilon, \tag{D1}$$

and in this scalar case, the error is generally proportional to the machine precision

$$\epsilon = \varepsilon. \tag{D2}$$

It should be noted that this error could be considered as deterministic error, where a fixed magnitude error ϵ is associated with the operations. Instead, stochastic error could be considered,

$$x \circ_\epsilon y - x \circ y = (x \circ y)\xi, \tag{D3}$$

where the error is assumed to be the random variable

$$\xi \sim \Omega_\epsilon \tag{D4}$$

from a distribution Ω_ϵ, generally with zero mean and generally a standard deviation that is proportional to the scale ϵ.

For successive additions, the order of the operations affects which terms have greater error, as initial terms accumulate more error over the course of successive additions. The sum of d scalars $\{a_\mu\}$ has error

$$\sum_\mu^d {}_\epsilon a_\mu - \sum_\mu^d a_\mu = \sum_\mu^d a_\mu \epsilon_\mu, \tag{D5}$$

where the accumulated error from addition is

$$\epsilon_\mu = (1 + \varepsilon)^{d-\mu} - 1. \tag{D6}$$

The error of summations of d scalars $a = \{a_\mu\}$ can be represented as inner products with a ones vectors 1, and error vectors $\epsilon = \{\epsilon_\mu\}$ to account for the accumulated error from

addition

$$1 \cdot_\epsilon a - 1 \cdot a = \epsilon \cdot a. \tag{D7}$$

b. Matrix floating point error

Generalizing beyond scalars, the deterministic error of inner products of d-dimensional matrices A, B can therefore be represented as an inner product with respect to an error matrix

$$AB_\epsilon - AB = A\Sigma B. \tag{D8}$$

Similarly, the stochastic error of inner products of d-dimensional matrices A, B can therefore be represented as an inner product with respect to an error matrix,

$$AB_\epsilon - AB = A\Xi B, \tag{D9}$$

where the error is assumed to be the random variable

$$\Xi \sim \Omega_\epsilon^{d \times d} \tag{D10}$$

from a distribution Ω_ϵ, generally with zero mean and generally a standard deviation that is proportional to the scale ϵ.

The relative error matrix is arbitrary depending on the model of computation, and we choose it to be diagonal,

$$\Sigma = \{\epsilon_\mu = (1 + \varepsilon)^{d - \mu} - 1\}, \tag{D11}$$

with norm

$$\|\Sigma\|_p = \left(\sum_\mu^d [(1 + \varepsilon)^{d - \mu} - 1]^p\right)^{1/p} \approx \left(\sum_\mu^d (d - \mu)^p\right)^{1/p} \varepsilon$$

$$\approx O\left(d^{\frac{p+1}{p}}\right)\varepsilon \sim O(\text{poly}(d))\varepsilon \equiv \epsilon^{(p)} \leqslant \epsilon, \tag{D12}$$

which we define to be upper bounded by a p-norm independent error ϵ.

This can be extended to a product of k matrices $\prod_\mu^k A_\mu$. We assume the error acts recursively, propagating from the initial to final matrix multiplications. Therefore, the matrix multiplication operation becomes a sum over all possible locations of the error matrix interlaced within the products, denoted with a multi-index $\chi_l = \{\chi_{l_\mu} \in [2], \mu \in [k]\}$, with the number of errors $|\chi_l| = l \leqslant k$. We assume that errors also exist on the representation of the matrix elements, yielding an extra error matrix factor at the end of the product, and errors from additions of the error terms are a secondary effect.

The error of the product of k matrices $\prod_\mu^k A_\mu$ therefore takes the form

$$\prod_\mu^k{}_\epsilon A_\mu - \prod_\mu^k A_\mu = \sum_{l>0} \sum_{\chi_l} \prod_\mu^k A_\mu \Sigma^{\chi_{l_\mu}}. \tag{D13}$$

We then use norm inequalities, and define an upper bound on the norms values $\|A_\mu\|_p \leqslant \|A\|_p$ to bound the norm of the absolute matrix multiplicative error,

$$\left\|\prod_\mu^k{}_\epsilon A_\mu - \prod_\mu^k A_\mu\right\|_s = \left\|\sum_{l>0} \sum_{\chi_l} \prod_\mu^k A_\mu \Sigma^{\chi_{l_\mu}}\right\|_s \tag{D14}$$

$$\leqslant \sum_{l>0} \sum_{\chi_l} \left\|\prod_\mu^k A_\mu \Sigma^{\chi_{l_\mu}}\right\|_s \tag{D15}$$

$$\leqslant \sum_{l>0} \sum_{\chi_l} \min_{\sum_\mu^k \frac{1}{p_\mu} + \frac{\chi_{l_\mu}}{q_\mu} = \frac{1}{s}} \prod_\mu^k \|A_\mu\|_{p_\mu} \|\Sigma\|_{q_\mu}^{\chi_{l_\mu}} \tag{D16}$$

$$\leqslant \sum_{l>0} \min_{\frac{k}{p} + \frac{l}{q} = \frac{1}{s}} \binom{k}{l} \|A\|_p^k \|\Sigma\|_q^l \tag{D17}$$

$$\leqslant \sum_{l>0} \binom{k}{l} \lambda(A)^k \epsilon^l \tag{D18}$$

$$= \lambda(A)^k [(1 + \epsilon)^k - 1]. \tag{D19}$$

Therefore, the relative multiplicative error is upper bounded for a general norm by

$$\frac{\left\|\prod_\mu^k{}_\epsilon A_\mu - \prod_\mu^k A_\mu\right\|}{\left\|\prod_\mu^k A_\mu\right\|} \leqslant \frac{\lambda(A)^k}{\left\|\prod_\mu^k A_\mu\right\|} ((1 + \epsilon)^k - 1)$$

$$\xrightarrow{A_\mu \in \mathcal{U}(d)} \frac{1}{\text{poly}(d)} ((1 + \epsilon)^k - 1), \tag{D20}$$

where in the case of unitary matrices with unit singular values, the bound is simplified.

2. Classical error scaling

Let us define a unitary evolution as a product of k, d-dimensional unitaries $U = \prod_\mu^k U_\mu$ that transforms an initial state as $\sigma \to \rho = U\sigma U^\dagger$. Suppose the evolution is subject to classical floating point error $U \to U_\epsilon$ and $\sigma \to \rho_\epsilon$. An interesting interpretation is that this noisy evolution with classical floating point error can be represented by unnormalized Kraus-like operators, $\{I, \Sigma\}$, where $I + \Sigma^\dagger \Sigma \neq I$, meaning the operation is not trace-preserving.

The composite adjoint action of unitaries with k matrix multiplications with classical error is therefore

$$\rho_\epsilon = U_\epsilon \sigma U_\epsilon^\dagger = \rho + \delta_\epsilon, \tag{D21}$$

where the perturbation is

$$\delta_\epsilon = \sum_{l+l'>0} \sum_{\chi_l, \chi_{l'}} \left[\prod_\mu^k U_\mu \Sigma^{\chi_{l_\mu}}\right] \sigma \left[\prod_{\mu'}^k \Sigma^{\chi'_{l'_{\mu'}}\dagger} U_{\mu'}^\dagger\right]. \tag{D22}$$

The absolute norm error, given the nonunitary classical noise has norm $\|\Sigma\| = \epsilon$ and the operators are unitaries with
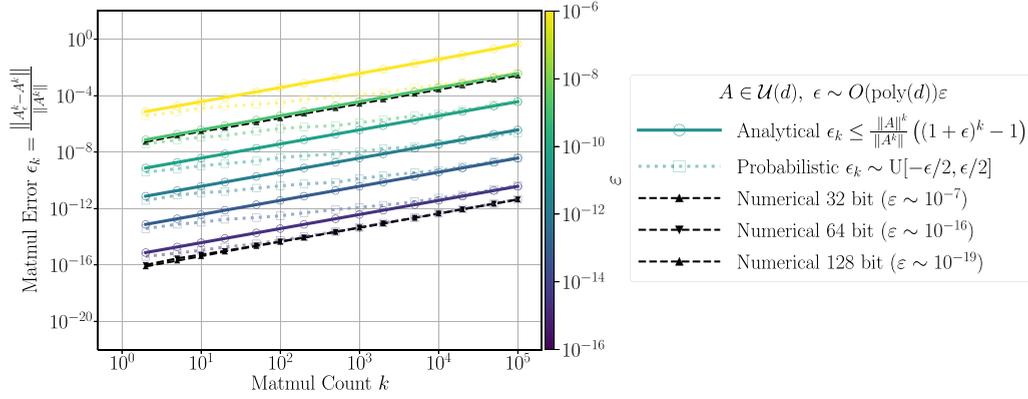
FIG. 10. Matrix multiplication floating point error scaling with number of multiplications k, with analytical (circles/solid lines), numerical probabilistic (squares/dotted lines), and numerical data-type (triangles/dashed lines) models scaling for $d = 2^2$-dimensional random unitaries A. Simulated floating point error is shown to increase polynomially with the number of floating point operations. Exact floating point data-types are in exact agreement with the analytical upper bounds, with the exception of simulated 128-bit data-type, whose error appears to be bounded by the 64-bit precision, likely attributed to intermediate backend calculations being cast to 64-bit precision.

$\|U_\mu\| = \text{poly}(d)$, is therefore bounded by

$$\|\rho_\epsilon - \rho\| \leqslant \sum_{l+l'>0} \sum_{\chi_l, \chi'_{l'}} \left\| \left[\prod_\mu^k U_\mu \Sigma^{\chi_{l\mu}} \right] \sigma \left[\prod_{\mu'}^k \Sigma^{\chi'_{l'\mu'}{}^\dagger} U_{\mu'}^\dagger \right] \right\| \tag{D23}$$

$$\leqslant \sum_{l+l'>0} \sum_{\chi_l, \chi'_{l'}} \left[\prod_\mu^k \|U_\mu \Sigma^{\chi_{l\mu}}\| \right] \|\sigma\| \left[\prod_{\mu'}^k \|\Sigma^{\chi'_{l'\mu'}{}^\dagger} U_{\mu'}^\dagger\| \right] \tag{D24}$$

$$= \sum_{l+l'>0} \sum_{\chi_l, \chi'_{l'}} \left[\prod_\mu^k \|\Sigma^{\chi_{l\mu}}\| \right] \|\sigma\| \left[\prod_{\mu'}^k \|\Sigma^{\chi'_{l'\mu'}{}^\dagger}\| \right] \tag{D25}$$

$$\leqslant \sum_{l+l'>0} \sum_{\chi_l, \chi'_{l'}} \|\sigma\| \prod_{\mu, \mu'}^k \|\Sigma\|^{\chi_{l\mu}+\chi'_{l'\mu'}} \tag{D26}$$

$$= \sum_{l+l'>0}^k \binom{k}{l}\binom{k}{l'} \|\sigma\| \epsilon^{l+l'} \tag{D27}$$

$$= \|\sigma\| \left[\left(\sum_l^k \binom{k}{l} \epsilon^l \right)^2 - 1 \right] \tag{D28}$$

$$= \|\sigma\|((1+\epsilon)^{2k} - 1). \tag{D29}$$

Given the norm of pure input and noiseless output states is $\|\rho\| = \|\sigma\| = 1$, the relative error for pure states is

$$\frac{\|\rho_\epsilon - \rho\|}{\|\rho\|} \leqslant |1 - (1+\epsilon)^{2k}|. \tag{D30}$$

We may then calculate functions of classical noisy states, such as the infidelity with a (pure) state ρ, as

$$\mathcal{L}_\epsilon = 1 - \text{tr}(\rho\rho_\epsilon) = \mathcal{L} - \text{tr}(\rho\delta_\epsilon). \tag{D31}$$

Using the Cauchy-Schwartz inequality, we may bound the bias of the classical noisy infidelities as

$$|\mathcal{L}_\epsilon - \mathcal{L}| = |\text{tr}(\rho\delta_\epsilon)| \leqslant \min_{\frac{1}{p}+\frac{1}{q}=1} \|\rho\|_p \|\delta_\epsilon\|_q$$

$$\leqslant \min_{\frac{1}{p}+\frac{1}{q}=1} \|\rho\|_p \|\sigma\|_q |1 - (1+\epsilon)^{2k}|, \tag{D32}$$

and this noisy linear objective deviation, for pure states, scales as

$$|\mathcal{L}_\epsilon - \mathcal{L}| \leqslant |1 - (1+\epsilon)^{2k}|. \tag{D33}$$

An example simulation of classical floating point error for k successive matrix multiplications, by artificially including noise with different scales ε, is shown in Fig. 10. To investigate this classical error, we compare analytical results (exact upper bounds on the difference between noisy and noiseless matrix multiplications), probabilistic results (numerically adding zero-mean uniformly random errors $U[-\varepsilon/2, \varepsilon/2]$ to results of floating point operations), and numerical results (numerically performing matrix multiplications without artificial noise for standard data types: single point 32-bit precision with $\varepsilon \approx 10^{-7}$, double point 64-bit precision with $\varepsilon \approx 10^{-16}$, and quadruple point 128-bit precision with $\varepsilon \approx 10^{-19}$–$10^{-24}$, depending on whether the precision is only simulated to $\sim$80-bit for double precision architectures). We observe good agreement across all models, and we note the drift in the probabilistic results at a large number of matrix multiplications k to an effective ε an order of magnitude smaller than simulated. This suggests the built-in floating point operation algorithms may be caching results or eliminating some sources of error. We repeat such experiments across $S = 10$ samples, and we also observe negligible error bars within plot markers. This classical error also shows similar trends to the quantum noise case of noise-induced convergent and divergent regimes. Finally, we note that it is difficult to precisely and consistently define the machine precision ε, or the number of decimal points of accuracy of a simulation, even for a recognized floating point data-type. Apart from the often nonlinear dependence of the true machine precision value on the number of operations and dimensionality of the problem, there are additionally software effects. These effects include how basic floating point operations may be vectorized or fused together. Similarly, there may be hardware effects, such as how the exact memory layout of different computing architectures can affect the way in which operations are performed. These studies, in particular the exact agreement between analytical and 32- and 64-bit data-type numerical results, however, confirm the appropriateness of our derived models, and they offer valuable insight into upper bound estimates for a range of machine precision values.

The simulation package developed for this work [75], as discussed in Section F, can perform single $\varepsilon \sim O(10^{-7})$, double $\varepsilon \sim O(10^{-16})$, or (simulated) quadratic $\varepsilon \sim O(10^{-19})$ floating point arithmetic, and each data type is useful in understanding the sources of floating point error. However, quadruple floating point arithmetic is unable to currently be compiled efficiently for large systems, in addition to the overhead of the simulated precision. Automatic differentiation computations also empirically have less than double precision accuracy, typically $\varepsilon \sim O(10^{-12})$. Therefore, current large-scale simulations are infeasible with this data type.

3. Quantum error scaling

Let us define a unitary evolution as a product of k, d-dimensional unitaries $U = \prod_\mu^k U_\mu$ that transforms an initial state as $\sigma \to \rho = U \sigma U^\dagger$. Suppose the evolution is subject to quantum error $U \to U_\gamma$ and $\sigma \to \rho_\gamma$. This noisy evolution with quantum error can be represented by the normalized Kraus operators $\{\sqrt{1-\gamma}I, \sqrt{\gamma}\Sigma\}$, where $(1 - \gamma)I + \gamma \Sigma^\dagger \Sigma = I$, meaning the operation is trace-preserving.

The composite adjoint action of unitaries with $K = NM$ matrix multiplications with quantum error is therefore

$$\rho_\gamma = U_\gamma \sigma U_\gamma^\dagger = (1-\gamma)^K \rho + \delta_\gamma, \tag{D34}$$

where the perturbation is

$$\delta_\gamma = \sum_{l>0}^{K} \sum_{\chi_l} \gamma^l (1-\gamma)^{K-l} \left[\prod_\mu^K U_\mu \Sigma^{\chi_{l\mu}} \right] \sigma \left[\prod_{\mu'}^K \Sigma^{\chi'_{l'\mu'}{}^\dagger} U_{\mu'}^\dagger \right]. \tag{D35}$$

The absolute norm error, given the unitary quantum noise has norm $\|\Sigma\| = \text{poly}(d)$ and the operators are unitaries with $\|U_\mu\| = \text{poly}(d)$, is therefore bounded by

$$\|\rho_\gamma - \rho\| \leqslant [(1-\gamma)^K - 1]\|U\sigma U^\dagger\| + \sum_{l>0}^{K} \sum_{\chi_l} \gamma^l (1-\gamma)^{K-l} \left\| \left[\prod_\mu^K U_\mu \Sigma^{\chi_{l\mu}} \right] \sigma \left[\prod_{\mu'}^K \Sigma^{\chi'_{l'\mu'}{}^\dagger} U_{\mu'}^\dagger \right] \right\| \tag{D36}$$

$$= [(1-\gamma)^K - 1]\|\sigma\| + \sum_{l>0}^{K} \sum_{\chi_l} \gamma^l (1-\gamma)^{K-l} \|\sigma\| \tag{D37}$$

$$= [(1-\gamma)^K - 1]\|\sigma\| + \sum_{l>0}^{K} \binom{K}{l} \gamma^l (1-\gamma)^{K-l} \|\sigma\| \tag{D38}$$

$$= 2\|\sigma\|[(1-\gamma)^K - 1]. \tag{D39}$$

Given the norm of pure input and noiseless output states is $\|\rho\| = \|\sigma\| = 1$, the relative error for pure states is

$$\frac{\|\rho_\epsilon - \rho\|}{\|\rho\|} \leqslant 2|1 - (1-\gamma)^K|. \tag{D40}$$

We may then calculate functions of quantum noisy states, such as the infidelity with a (pure) state ρ as

$$\mathcal{L}_\gamma = 1 - \text{tr}(\rho \rho_\gamma) = \mathcal{L} - \text{tr}(\rho \delta_\gamma). \tag{D41}$$

TABLE II. Constrained parameters for NMR *Ansatz* [10]. All parameters are chosen to be experimentally relevant. Some parameters, however, in particular the time step of 100 μs (unitary compilation) and 75 μs (state preparation), are chosen to be on the maximum end of what is experimentally feasible, to allow for reasonably efficient simulation.

Parameter	Description	Value
N	Number of qubits	1–4
M	Number of time steps	5–5000
τ	Trotterization time step	75–100 μs
T	Evolution time	375 μs–500 ms
Q	Spatial Trotterization order	2
P	Number of parameters	$MN(N+5)/2$
$J_{i<j}$	Constant longitudinal coupling	$\pi/2$ {72.4, −130.0, 50.0, 210.0, 20.0, −190.0, −30.0, 60.0, 90.0, −60.0} Hz
h_i	Constant longitudinal field	$\pi/2$ {−10.0, 0, −1.0, 29.0, −20.0} kHz
$\bar{\theta}$	Variable transverse field	$\pi/2$ MHz
γ	Noise scale	10^{-14}–10^{-1}
$\tilde{\theta}$	Constraints	Constrained and shared transverse fields Zero-field Dirichlet boundary conditions in time $\theta_i^{x,y(m)} = \theta^{x,y(m)}$, $|\theta^{x,y(m)}| \leqslant \bar{\theta}$, $\theta^{x,y(0,M)} = 0$

Using the Cauchy-Schwartz inequality, we may bound the bias of the quantum noisy infidelities

$$|\mathcal{L}_\gamma - \mathcal{L}| = |\text{tr}(\rho \delta_\gamma)| \leqslant \min_{\frac{1}{p}+\frac{1}{q}=1} \|\rho\|_p \|\delta_\gamma\|_q$$

$$\leqslant \min_{\frac{1}{p}+\frac{1}{q}=1} 2\|\rho\|_p \|\sigma\|_q |1 - (1-\gamma)^K|, \qquad (D42)$$

and this noisy linear objective deviation, for pure states, scales as

$$|\mathcal{L}_\gamma - \mathcal{L}| \leqslant 2|1 - (1-\gamma)^K|. \qquad (D43)$$

4. Classical versus quantum error scaling

We may relate the classical and quantum error scales by identifying the dimension-dependent classical scale

$$\epsilon = \epsilon(d) \qquad (D44)$$

and the number of matrix multiplication operations for M poly(N) local unitaries being

$$k = O(K) = O(\text{poly}(N)M). \qquad (D45)$$

The error scaling of classical and quantum noisily evolved states may then be summarized as

$$\|\rho_\epsilon - \rho\| \leqslant |1 - (1+\epsilon)^{2k}|, \qquad (D46)$$

$$\|\rho_\gamma - \rho\| \leqslant 2|1 - (1-\gamma)^K|, \qquad (D47)$$

and the linear functions of the perturbed states are also perturbed with identical scaling,

$$|\mathcal{L}_\epsilon - \mathcal{L}| \leqslant |1 - (1+\epsilon)^{2k}|, \qquad (D48)$$

$$|\mathcal{L}_\gamma - \mathcal{L}| \leqslant 2|1 - (1-\gamma)^K|. \qquad (D49)$$

A subtle point to keep in mind is that we are comparing infidelities at fixed variable parameters values. If the parameters have been optimized in a noisy setting, then the associated noiseless infidelity $\mathcal{L}_{\theta_\gamma}$, evaluated with these parameters, likely differs from the true optimal noiseless infidelity $\mathcal{L}_\theta$ with parameters optimized in a noiseless setting. If we happen to

be in an overparameterized regime, where the phenomenon of lazy training occurs [76], parameters may change negligibly from their initial values. In this regime, the noiseless infidelities, with noisy and noiseless trained parameters, possibly approximately coincide,

$$\theta_\gamma^* \approx \theta^*. \qquad (D50)$$

In this setting, functions of the parameters such as the infidelities may also approximately coincide,

$$\mathcal{L}_{\theta_\gamma^*} \approx \mathcal{L}_{\theta^*}, \qquad (D51)$$

and they can be used in place of each other. Bounds such as those above comparing noisy and noiseless fidelities may be relevant, however future studies should investigate the noise-induced bias in optimization.

APPENDIX E: NUCLEAR MAGNETIC RESONANCE ANSATZ

In these Appendixes, we include details about the nuclear magnetic resonance (NMR) *Ansatz* studied in this work. We describe the model Hamiltonian, and we document all parameter scales

$$H_\theta^{(t)} = \sum_i \theta_i^{x(t)} X_i + \sum_i \theta_i^{y(t)} Y_i$$

$$+ \sum_i h_i Z_i + \sum_{i<j} J_{ij} Z_i Z_j. \qquad (E1)$$

Here we have control over the variable time-dependent transverse X and Y fields, with constant time-independent longitudinal Z and ZZ fields. In these units, the parameters in Table II are used for this *Ansatz*. We note that the nonlocal coupling scale is $J \ll h, \theta$, and there are time step sizes such that $\tau J \ll 1$. To generate a finite angle ($\pi/4$ for ZZ) rotation necessary to implement a single entangling gate, depths of order $M \sim 1/\tau J \sim O(10^4)$ are necessary. Furthermore, the number of entangling gates necessary to compile Haar random unitaries is generally exponential $O(D^N)$ in the number of qubits [48]. Therefore, these parameter scales, even for

TABLE III. Comparison of 1-qubit gate times T_{U1}, 2-qubit gate times T_{U2}, decoherence times T_γ, experimentally achieved number of qubits N_U, experimentally achieved depths of 2-qubit gates M_U, ratio of 2-qubit to 1-qubit gate times $M_{U12} = T_{U2}/T_{U1}$, and the ratio of decoherence to 2-qubit gate times $M_{\gamma U2} = T_\gamma/T_{U2}$ for various experimental quantum computing implementations. The ratio of 2- to 1-qubit gate times can be interpreted as the effective depth M_{U12} necessary to implement 2-qubit gates. The ratio of decoherence times T_γ to 2-qubit gate times can be interpreted as the effective maximum depth $M_{\gamma U2}$ before decoherence. For NMR, 1-qubit gate times $T_{U1} = O(\tau)$ are approximately the pulse time steps, 2-qubit gate times $T_{U2} = O(1/J)$ are approximately inversely related to the 2-qubit coupling, and decoherence times $T_\gamma = O(T_2)$ are approximately the dephasing times. Tabulated values are experimental apparatus and algorithm specific [45], taken from a range of values in recent literature.

Implementation	T_{U1} (µs)	T_{U2} (µs)	T_γ (ms)	N_U	M_U	$M_{U12} = T_{U2}/T_{U1}$	$M_{\gamma U2} = T_\gamma/T_{U2}$
NMR [10,17,46]	75	5000	1000 ms	12	1000	70	2×10^2
Trapped ions [1,2,46,77,78]	1	50	10000 ms	12	64	50	1×10^6
Superconducting [3,4,46]	0.02	0.1	1 ms	127	60	5	1×10^4
Neutral atoms [6–8]	2	0.5	1000 ms	280	516	0.25	2×10^6

relatively small system sizes of $N \leqslant 4$, indicate that in general an even greater total number, possibly superexponential $M > O(D^N)$, of physical gates are necessary in practice for these tasks of interest.

We also compare gate times, decoherence times, experimentally tested depths, and effective depths based on these times for various NISQ implementations in Table III. We note that each implementation has better and worse properties. NMR in this work has the largest 2-qubit gate times $T_{U2} = O(1/J)$, which translates to NMR having the largest effective depth $T_{U2}/T_{U1} \sim O(100)$ required for each 2-qubit gate, and having the smallest effective maximum depth $T_\gamma/T_{U2} \sim O(10^2)$ before coherence. We emphasize, however, that tabulated values taken from a range of values in recent literature are very experimental apparatus- and algorithm-specific [45].

APPENDIX F: CLASSICAL SIMULATION AND OPTIMIZATION

In these Appendixes, we discuss details of the classical simulation and optimization of the quantum systems studied in this work. We explain our hyperparameter choices, optimization routines, and we list all optimization settings.

The quantum systems in this work are simulated and optimized classically using a compiled, automatic-differentiation library using the Python JAX backend, developed as a general differentiable exponentially deep circuit simulator [75]. This library is optimized for noisy density matrix simulations with few qubits $N < 6$, and state-of-the-art large depth circuits with $k \sim O(10^5)$ gates per circuit instance.

For example, in this work's NMR *Ansatz,* circuits consist of $Q = 2$ order spatially Trotterized evolution, consisting of fully connected two-body gates. Local noise on all qubits after each layer requires up to $M \sim 5 \times 10^3$ circuit layers, and the circuits are simulated roughly 10^9 times throughout the optimization routines and loops over *Ansatz* settings. All optimizer hyperparameters are shown in Table IV.

When performing gradient-based optimization, the gradients of these objectives can be expressed analytically with parameter shift rules, however for the $N \leqslant 4$ system sizes considered in this work, we use automatic differentiation for efficiency. An exception is when computing the Hessian

and quantum Fisher information, where computing analytical gradients is most memory-efficient. For small system sizes, the unitary and quantum noise operations and infidelities, specifically the traces over the full space, can also be calculated exactly. Sampling methods, which introduces forms of shot-noise, and other quantum algorithms, such as the Hilbert-Schmidt test, and other forms of process tomography, are not necessary to be implemented. The effects of estimating such infidelities with sampling and approximate operators for larger system sizes are important studies to be conducted in future works.

For computing statistics such as the mean and variance across S independent optimizations, we sample any inputs from specific distributions. We generally sample the initial states σ, target unitaries U, and target states ρ according to the Haar measure to avoid any biases in targeting a specific subspace [36]. Expectation values of parametrized functions of the initial and target states $\mathcal{F}_\theta(\sigma, \rho)$, such as infidelities, may then be computed as

$$\langle \mathcal{F}_\theta \rangle^{(S)} = \frac{1}{S} \sum_{\substack{s \\ \tilde{\theta}^{(s)} \sim \text{Uniform} \\ \sigma^{(s)} \sim \text{Haar}, \, \rho^{(s)} \sim \text{Haar}}}^{S} \mathcal{F}_{\theta^{(s)}}(\sigma^{(s)}, \rho^{(s)}). \quad \text{(F1)}$$

For artificially simulating floating point error, we add random matrices to each successive distinct matrix multiplication in a calculation. Plotted quantities are also calculated at the optimal parametrization, which is not necessarily at the last optimization iteration. We also uniformly randomly sample initial parameters $\theta = \tilde{\theta}$, and then we smooth them over the M time steps with cubic interpolation to be experimentally implementable.

In this work, we define the parameters as $\theta = \theta(\phi)$, functions of explicit variables ϕ that are explicitly optimized. In the unconstrained case, all parameters for each qubit are independent and are not constrained in magnitude. However in the constrained case, we impose that the fields are constrained and coupled to act uniformly across all sites for each operator. Therefore, $\theta_i^{x,y} = \theta^{x,y}$ for each qubit $i \in [N]$. We also bound all transverse field magnitudes $|\theta_i^{x,y}| \leqslant \bar{\theta}$. We finally impose Dirichlet boundary conditions in time, such that the initial and final fields are approximately zero, $\theta^{x,y(0)} = \theta^{x,y(M)} = 0$.

To perform the classical simulation and optimization of the parametrized channels, we perform first-order gradient based

TABLE IV. Optimization hyperparameters. All settings are selected from manual parameter searches, and are only heuristically shown to reasonably guarantee optimization convergence across all system settings.

Hyperparameter	Description	Value
Optimizer	Optimizer routine	Conjugate gradient
Line search	Line search routine	Wolf conditions
Conjugate search	Conjugate search routine	Hestenes-Stiefel
$p_{\bar{\theta}}$	Initial parameters distribution	Uniform/Cubic Smoothing
$p_{U,\rho}$	Objectives distribution	Haar
p_{σ}	Initial state distribution	Haar
S	Number of sample objectives	50
L	Maximum number of optimization iterations	500
$L_<$	Minimum number of optimization iterations	50
L_{α}	Maximum number of line search iterations per iteration	2500
$\epsilon_{\mathcal{L}}$	Minimum objective stopping condition	10^{-16}
$\epsilon_{\Delta\mathcal{L}}$	Minimum absolute difference in objective per iteration stopping condition	0
$\epsilon_{\delta\mathcal{L}}$	Maximum relative increase in objective per iteration stopping condition	10^{-3}
$\epsilon_{\partial\mathcal{L}}$	Minimum gradient norm stopping condition	0
$\epsilon_{\Delta\partial\mathcal{L}}$	Minimum absolute difference in gradient norm per iteration stopping condition	0
$\epsilon_{\delta\partial\mathcal{L}}$	Maximum relative increase in gradient norm per iteration stopping condition	∞
α	Initial parameter learning rate	10^{-4}
β	Initial search learning rate	10^{-4}
$\alpha_{\lessgtr}$	Bounds on parameter learning rate before reset to α	$[0, \infty]$
$\beta_{\lessgtr}$	Bounds on search learning rate before reset to β	$[10^{-10}, 10^{10}]$
c_1	Wolf objective parameter	10^{-5}–10^{-4}
c_2	Wolf gradient parameter	10^{-1}–9×10^{-1}

optimization routines to minimize the objectives $\mathcal{L}_{\theta\gamma}$ with respect to the variable parameters θ. We may then compute gradients of objectives $\zeta = \partial\mathcal{L}_{\theta}$.

Due to the high dimensionality of the problem, more effective variants of classical gradient descent must be performed. Here we choose a variant of the first-order conjugate gradient scheme, where the search direction ξ is updated iteratively [50] at iteration $0 \leqslant l \leqslant L$:

$$\theta^{(l+1)} = \theta^{(l)} + \alpha^{(l)}\xi^{(l)}, \tag{F2}$$

$$\xi^{(l+1)} = -\zeta^{(l+1)} + \beta^{(l)}\xi^{(l)}, \tag{F3}$$

given initial conditions of $\theta^{(0)}$ and $\xi^{(0)} = -\zeta^{(0)}$.

The learning rates α, β must be chosen to satisfy the Wolfe convergence conditions [50], which guarantee the parameters and search directions are updated such that the objective is monotonically decreasing.

For the parameter learning rate α, a line search is conducted that involves at most L_{α} objective calls per iteration l, and it ensures the objective maximally decreases. For the search learning rate β, rates that obey

$$|\beta| \leqslant \bar{\beta} = \frac{\zeta^{(l+1)} \cdot \zeta^{(l+1)}}{\zeta^{(l)} \cdot \zeta^{(l)}} \tag{F4}$$

ensure convergence [50]. Various forms for this parameter include the standard Fletcher-Reeves rate $\beta_{FR}^{(l)} = \bar{\beta}^{(l)}$. However, for the range of problems in this work, we find that it and many definitions lead to the optimizer immediately getting stuck in local minima. We find the Hestenes-Stiefel rate, however, to be quite effective,

$$\beta_{HS}^{(l)} = \frac{\zeta^{(l+1)} \cdot (\zeta^{(l+1)} - \zeta^{(l)})}{\xi^{(l)} \cdot (\zeta^{(l+1)} - \zeta^{(l)})}. \tag{F5}$$

The hyperparameters are chosen after heuristic manual searches that indicated stability and adequately fast convergence for noisy and noiseless state preparation and unitary compilation tasks. Additional heuristic stopping conditions are also implemented to avoid unnecessary iterations.

[1] C. D. Bruzewicz, J. Chiaverini, R. McConnell, and J. M. Sage, Trapped-ion quantum computing: Progress and challenges, Appl. Phys. Rev. **6**, 021314 (2019).

[2] P. J. Low, B. M. White, A. A. Cox, M. L. Day, and C. Senko, Practical trapped-ion protocols for universal qudit-based quantum computing, Phys. Rev. Res. **2**, 033128 (2020).

[3] Y. Kim, A. Eddins, S. Anand, K. X. Wei, E. van den Berg, S. Rosenblatt, H. Nayfeh, Y. Wu, M. Zaletel, K. Temme, and A. Kandala, Evidence for the utility of quantum computing before fault tolerance, Nature (London) **618**, 500 (2023).

[4] M. Werninghaus, D. J. Egger, F. Roy, S. Machnes, F. K. Wilhelm, and S. Filipp, Leakage reduction in fast superconducting qubit gates via optimal control, npj Quantum Inf. **7**, 14 (2021).

[5] G. Wendin, Quantum information processing with superconducting circuits: a review, Rep. Prog. Phys. **80**, 106001 (2017).

[6] D. Bluvstein, S. J. Evered, A. A. Geim, S. H. Li, H. Zhou, T. Manovitz, S. Ebadi, M. Cain, M. Kalinowski, D. Hangleiter, J. P. B. Ataides, N. Maskara, I. Cong, X. Gao, P. S. Rodriguez, T. Karolyshyn, G. Semeghini, M. J. Gullans, M. Greiner, V. Vuletić, and M. D. Lukin, Logical quantum processor based on reconfigurable atom arrays, Nature (London) **626**, 58 (2024).

[7] D. Bluvstein, H. Levine, G. Semeghini, T. T. Wang, S. Ebadi, M. Kalinowski, A. Keesling, N. Maskara, H. Pichler, M. Greiner, V. Vuletić, and M. D. Lukin, A quantum processor based on coherent transport of entangled atom arrays, Nature (London) **604**, 451 (2022).

[8] K. Wintersperger, F. Dommert, T. Ehmer, A. Hoursanov, J. Klepsch, W. Mauerer, G. Reuber, T. Strohm, M. Yin, and S. Luber, Neutral atom quantum computing hardware: Performance and end-user perspective, EPJ Quantum Technol. **10**, 32 (2023).

[9] H. Levine, A. Keesling, A. Omran, H. Bernien, S. Schwartz, A. S. Zibrov, M. Endres, M. Greiner, V. Vuletić, and M. D. Lukin, High-fidelity control and entanglement of Rydberg-atom qubits, Phys. Rev. Lett. **121**, 123603 (2018).

[10] J. P. S. Peterson, R. S. Sarthour, and R. Laflamme, Enhancing quantum control by improving shaped-pulse generation, Phys. Rev. Appl. **13**, 054060 (2020).

[11] K. R. K. Rao, H. Katiyar, T. S. Mahesh, A. Sen(De), U. Sen, and A. Kumar, Multipartite quantum correlations reveal frustration in a quantum Ising spin system, Phys. Rev. A **88**, 022312 (2013).

[12] J. Preskill, Quantum computing in the NISQ era and beyond, Quantum **2**, 79 (2018).

[13] X. Ge, R. B. Wu, and H. Rabitz, The optimization landscape of hybrid quantum-classical algorithms: From quantum control to NISQ applications, Annu. Rev. Control **54**, 314 (2022).

[14] S. Wang, P. Czarnik, A. Arrasmith, M. Cerezo, L. Cincio, and P. J. Coles, Can error mitigation improve trainability of noisy variational quantum algorithms? Quantum **8**, 1287 (2024).

[15] K. Sharma, S. Khatri, M. Cerezo, and P. J. Coles, Noise resilience of variational quantum compiling, New J. Phys. **22**, 043006 (2020).

[16] M. Dalgaard and F. Motzoi, Fast, high precision dynamics in quantum optimal control theory, J. Phys. B **55**, 085501 (2022).

[17] J. P. Peterson, H. Katiyar, and R. Laflamme, Fast simulation of magnetic field gradients for optimization of pulse sequences, arXiv:2006.10133.

[18] G. Feng, F. H. Cho, H. Katiyar, J. Li, D. Lu, J. Baugh, and R. Laflamme, Gradient-based closed-loop quantum optimal control in a solid-state two-qubit system, Phys. Rev. A **98**, 052341 (2018).

[19] A. B. Magann, C. Arenz, M. D. Grace, T. S. Ho, R. L. Kosut, J. R. McClean, H. A. Rabitz, and M. Sarovar, From pulses to circuits and back again: A quantum optimal control perspective on variational quantum algorithms, PRX Quantum **2**, 010101 (2021).

[20] 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).

[21] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, Variational quantum algorithms, Nat. Rev. Phys. **3**, 625 (2021).

[22] J. Liu, F. Tacchino, J. R. Glick, L. Jiang, and A. Mezzacapo, Representation learning via quantum neural tangent kernels, PRX Quantum **3**, 030323 (2022).

[23] M. Schuld, I. Sinayskiy, and F. Petruccione, An introduction to quantum machine learning, Contemp. Phys. **56**, 172 (2015).

[24] Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli, Statistical mechanics of deep learning, Annu. Rev. Condens. Matter Phys. **11**, 501 (2020).

[25] L. Chizat, E. Oyallon, and F. Bach, On lazy training in differentiable programming, Adv. Neural Inf. Process. Syst. **32** (2018).

[26] B. T. Kiani, S. Lloyd, and R. Maity, Learning unitaries by gradient descent, arXiv:2001.11897.

[27] R. Wiersema, C. Zhou, Y. De Sereville, J. F. Carrasquilla, Y. B. Kim, and H. Yuen, Exploring entanglement and optimization within the Hamiltonian variational Ansatz, PRX Quantum **1**, 020319 (2020).

[28] M. Larocca, N. Ju, D. García-Martín, P. J. Coles, and M. Cerezo, Theory of overparametrization in quantum neural networks, Nat. Comput. Sci. **3**, 542 (2021).

[29] M. Benedetti, M. Fiorentini, and M. Lubasch, Hardware-efficient variational quantum algorithms for time evolution, Phys. Rev. Res. **3**, 033083 (2021).

[30] G. S. Ravi, K. Smith, J. M. Baker, T. Kannan, N. Earnest, A. Javadi-Abhari, H. Hoffmann, and F. T. Chong, Navigating the dynamic noise landscape of variational quantum algorithms with QISMET, in *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems* (Association for Computing Machinery, New York, 2023), Vol. 2, pp. 515–529.

[31] L. Botelho, A. Glos, A. Kundu, J. A. Miszczak, Ö. Salehi, and Z. Zimborás, Error mitigation for variational quantum algorithms through mid-circuit measurements, Phys. Rev. A **105**, 022441 (2022).

[32] E. Fontana, M. Cerezo, A. Arrasmith, I. Rungger, and P. J. Coles, Non-trivial symmetries in quantum landscapes and their resilience to quantum noise, Quantum **6**, 804 (2020).

[33] D. S. Franca and R. Garcia-Patron, Limitations of optimization algorithms on noisy quantum devices, Nat. Phys. **17**, 1221 (2021).

[34] S. Wang, E. Fontana, M. Cerezo, K. Sharma, A. Sone, L. Cincio, and P. J. Coles, Noise-induced barren plateaus in variational quantum algorithms, Nat. Commun. **12**, 6961 (2020).

[35] E. Fontana, N. Fitzpatrick, D. M. Ramo, R. Duncan, and I. Rungger, Evaluating the noise resilience of variational quantum algorithms, Phys. Rev. A **104**, 022403 (2021).

[36] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, Connecting ansatz expressibility to gradient magnitudes and barren plateaus, PRX Quantum **3**, 010313 (2022).

[37] R. Laflamme, E. Knill, D. G. Cory, E. M. Fortunato, T. Havel, C. Miquel, R. Martinez, C. Negrevergne, G. Ortiz, M. A. Pravia, Y. Sharf, S. Sinha, R. Somma, and L. Viola, Introduction to NMR quantum information processing, arXiv:quant-ph/0207172.

[38] D. G. Cory, R. Laflamme, E. Knill, L. Viola, T. F. Havel, N. Boulant, G. Boutis, E. Fortunato, S. Lloyd, R. Martinez, C. Negrevergne, M. Pravia, Y. Sharf, G. Teklemariam, Y. S. Weinstein, and W. H. Zurek, NMR based quantum information processing: Achievements and prospects, Fortschr. Phys. **48**, 875 (2000).

[39] X. You, S. Chakrabarti, and X. Wu, A convergence theory for over-parameterized variational quantum eigensolvers, arXiv:2205.12481.

[40] D. García-Martín, M. Larocca, and M. Cerezo, Effects of noise on the overparametrization of quantum neural networks, Phys. Rev. Res. **6**, 013295 (2024).

[41] K. W. Moore and H. Rabitz, Exploring constrained quantum control landscapes, J. Chem. Phys. **137**, 134113 (2012).

[42] Y. Song, J. Li, Y. J. Hai, Q. Guo, and X. H. Deng, Optimizing quantum control pulses with complex constraints and few variables through autodifferentiation, Phys. Rev. A **105**, 012616 (2022).

[43] K. Georgopoulos, C. Emary, and P. Zuliani, Modeling and simulating the noisy behavior of near-term quantum computers, Phys. Rev. A **104**, 062432 (2021).

[44] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke, W. K. Mok, S. Sim, L. C. Kwek, and A. Aspuru-Guzik, Noisy intermediate-scale quantum algorithms, Rev. Mod. Phys. **94**, 015004 (2022).

[45] N. M. Linke, D. Maslov, M. Roetteler, S. Debnath, C. Figgatt, K. A. Landsman, K. Wright, and C. Monroe, Experimental comparison of two quantum computing architectures, Proc. Natl. Acad. Sci. USA **114**, 3305 (2017).

[46] S. Majidy, C. Wilson, and R. Laflamme, *Building Quantum Computers: A Practical Introduction* (Cambridge University Press, 2024).

[47] D. A. Lidar, Review of decoherence-free subspaces, noiseless subsystems, and dynamical decoupling, Adv. Chem. Phys. **154**, 295 (2014).

[48] S. Ashhab, N. Yamamoto, F. Yoshihara, and K. Semba, Numerical analysis of quantum circuits for state preparation and unitary operator synthesis, Phys. Rev. A **106**, 022426 (2022).

[49] S. J. Glaser, T. Schulte-Herbrüggen, M. Sieveking O, Schedletzky, N. C. Nielsen, O. W. Sørensen, and C. Griesinger, Unitary control in quantum ensembles: Maximizing signal intensity in coherent spectroscopy, Science **280**, 421 (1998).

[50] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. (Springer, New York, 2006).

[51] T. Haug and M. S. Kim, Generalization with quantum geometry for learning unitaries, arXiv:2303.13462.

[52] D. Wierichs, J. Izaac, C. Wang, and C. Y.-Y. Lin, General parameter-shift rules for quantum gradients, Quantum **6**, 677 (2022).

[53] L. Mirsky, A trace inequality of John von Neumann, Monatsh. Math. **79**, 303 (1975).

[54] M. Duschenes, J. Carrasquilla, and R. Laflamme, Data: Characterization of overparameterization in simulation of realistic quantum systems (2024), https://zenodo.org/records/10295792.

[55] J. Kattemölle and G. Burkard, Ability of error correlations to improve the performance of variational quantum algorithms, Phys. Rev. A **107**, 042426 (2023).

[56] Z. Cai, X. Xu, and S. C. Benjamin, Mitigating coherent noise using Pauli conjugation, npj Quantum Inf. **6**, 17 (2020).

[57] P. D. Johnson, J. Romero, J. Olson, Y. Cao, and A. Aspuru-Guzik, QVECTOR: An algorithm for device-tailored quantum error correction, arXiv:1711.02249.

[58] P. Niroula, S. Gopalakrishnan, and M. J. Gullans, Error mitigation thresholds in noisy quantum circuits, arXiv:2302.04278.

[59] B. Skinner, J. Ruhman, and A. Nahum, Measurement-induced phase transitions in the dynamics of entanglement, Phys. Rev. X **9**, 031009 (2019).

[60] S. Sarkar, On the trace of the product of Pauli and Dirac matrices, Acta Phys. Acad. Sci. Hung. **30**, 351 (1971).

[61] M. Schuld and N. Killoran, Is quantum advantage the right goal for quantum machine learning? PRX Quantum **3**, 030101 (2022).

[62] A. Abbas, A. Ambainis, B. Augustino, A. Bärtschi, H. Buhrman, C. Coffrin, G. Cortiana, V. Dunjko, D. J. Egger, B. G. Elmegreen, N. Franco, F. Fratini, B. Fuller, J. Gacon, C. Gonciulea, S. Gribling, S. Gupta, S. Hadfield, R. Heese, G. Kircher *et al.*, and Quantum optimization: potential, challenges, and the path forward, arXiv:2312.02279.

[63] S. Endo, S. C. Benjamin, and Y. Li, Practical quantum error mitigation for near-future applications, Phys. Rev. X **8**, 031027 (2018).

[64] www.vectorinstitute.ai/#partners.

[65] X. Yang, X. Nie, Y. Ji, T. Xin, D. Lu, and J. Li, Improved quantum computing with the higher-order trotter decomposition, Phys. Rev. A **106**, 042401 (2022).

[66] D. J. Egger, C. Capecci, B. Pokharel, P. K. Barkoutsos, L. E. Fischer, L. Guidoni, and I. Tavernelli, Pulse variational quantum eigensolver on cross-resonance-based hardware, Phys. Rev. Res. **5**, 033159 (2023).

[67] R. Wiersema, E. Kökcü, A. F. Kemper, and B. N. Bakalov, Classification of dynamical Lie algebras for translation-invariant 2-local spin systems in one dimension, arXiv:2309.05690.

[68] J. J. Meyer, Fisher information in noisy intermediate-scale quantum applications, Quantum **5**, 539 (2021).

[69] J. Haah, R. Kothari, R. O'Donnell, and E. Tang, Query-optimal estimation of unitary channels in diamond distance, in *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE Computer Society, Los Alamitos, CA, 2023), pp. 363–390.

[70] H. Fischer, *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*, edited by J. Buchwald (Springer, New York, 2010), pp. 1–415.

[71] T. Crowder, A linearization of quantum channels, J. Geom. Phys. **92**, 157 (2015).

[72] G. Kimura, The Bloch vector for N-level systems, Phys. Lett. A **314**, 339 (2003).

[73] T. Nuradha and M. M. Wilde, Fidelity-based smooth min-relative entropy: Properties and applications, IEEE Trans. Inform. Theory **70**, 4170 (2024).

[74] J. Wilkinson, *The Algebraic Eigenvalue Problem*, 1st ed. (Oxford University Press, 1965), pp. 112–116.

[75] M. Duschenes, simulation: Jax based simulator for quantum systems (2022), https://github.com/mduschenes/simulation.

[76] N. Shirai, K. Kubo, K. Mitarai, and K. Fujii, Quantum tangent kernel, arXiv:2111.02951.

[77] C. Monroe, W. C. Campbell, L.-M. Duan, Z.-X. Gong, A. V. Gorshkov, P. W. Hess, R. Islam, K. Kim, N. M. Linke, G. Pagano, P. Richerme, C. Senko, and N. Y. Yao, Programmable quantum simulations of spin systems with trapped ions, Rev. Mod. Phys. **93**, 025001 (2021).

[78] L. Zhao, J. Goings, K. Shin, W. Kyoung, J. I. Fuks, J. K. Kevin Rhee, Y. M. Rhee, K. Wright, J. Nguyen, J. Kim, and S. Johri, Orbital-optimized pair-correlated electron simulations on trapped-ion quantum computers, npj Quantum Inf. **9**, 60 (2023).