Can Error Mitigation Improve Trainability of Noisy Variational Quantum Algorithms?

Samson Wang[1,2], Piotr Czarnik[1,3,4], Andrew Arrasmith[1,5], M. Cerezo[1,5,6], Lukasz Cincio[1,5], and Patrick J. Coles[1,5]

[1]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[2]Department of Physics, Imperial College London, London, SW7 2AZ, UK
[3]Faculty of Physics, Astronomy, and Applied Computer Science, Jagiellonian University, Kraków, Poland
[4]Mark Kac Center for Complex Systems Research, Jagiellonian University, Kraków, Poland
[5]Quantum Science Center, Oak Ridge, TN 37931, USA
[6]Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Variational Quantum Algorithms (VQAs) are often viewed as the best hope for near-term quantum advantage. However, recent studies have shown that noise can severely limit the trainability of VQAs, e.g., by exponentially flattening the cost landscape and suppressing the magnitudes of cost gradients. Error Mitigation (EM) shows promise in reducing the impact of noise on near-term devices. Thus, it is natural to ask whether EM can improve the trainability of VQAs. In this work, we first show that, for a broad class of EM strategies, exponential cost concentration cannot be resolved without committing exponential resources elsewhere. This class of strategies includes as special cases Zero-Noise Extrapolation, Virtual Distillation, Probabilistic Error Cancellation, and Clifford Data Regression. Second, we perform analytical and numerical analysis of these EM protocols, and we find that some of them (e.g., Virtual Distillation) can make it harder to resolve cost function values compared to running no EM at all. As a positive result, we do find numerical evidence that Clifford Data Regression (CDR) can aid the training process in certain settings where cost concentration is not too severe. Our results show that care should be taken in applying EM protocols as they can either worsen or not improve trainability. On the other hand, our positive results for CDR highlight the possibility of engineering error mitigation methods to improve trainability.

1 Introduction

The prospect of obtaining quantum computational advantage for practical problems, such as simulating systems in chemistry and materials science, has generated much excitement. The past few years have witnessed tremendous progress towards this end, with significant focus on algorithm development for Noisy Intermediate-Scale Quantum (NISQ) computers. In particular, Variational Quantum Algorithms (VQAs) are a leading algorithmic approach because they adapt to the constraints of NISQ devices. Specifically, VQAs minimize a cost function by training a parameterized quantum circuit via a classical-quantum feedback loop [1, 2]. The cost is computed efficiently on a quantum computer whilst the parameter optimization is carried out classically. Different implementations of this versatile framework have been proposed for a broad spectrum of problems from dynamical quantum simulation [3–13] to machine learning [14–20] and beyond [21–40].

A central challenge in the NISQ regime is to combat the effects of noise as full error correction is not possible [41]. Decoherence, gate errors, and measurement noise all conspire to limit the complexity of quantum circuits that can be implemented on NISQ devices. While VQAs themselves offer some strategy to mitigate the impact of noise [1], it is widely viewed that VQAs alone

Samson Wang: samsonwang@outlook.com

will not be enough, and additional strategies will be needed to obtain quantum advantage in the face of noise. This has spawned the field of error mitigation (EM), and many researchers believe that VQAs combined with EM techniques will be the path forward. Indeed, EM methods like Zero-Noise Extrapolation [10,42–44], Clifford Data Regression [45], Virtual Distillation [46,47], Probabilistic Error Cancellation [42,43] and others [48–55] have been demonstrated to reduce errors of observable expectation values, sometimes by orders of magnitude. Hence, there has been hope that one can simply train the VQA in the presence of noise, and then after training, one can apply an EM method to extract the correct cost value (e.g., the ground state energy in the case of the variational quantum eigensolver [21]).

However, new challenges have recently been discovered for this approach [56, 57]. It is now recognized that noise impacts the trainability of VQAs, that is, the ability of the classical optimizer to find the global cost minimum. For ansatzes (i.e., parameterized quantum circuits) with depth linear or superlinear in the number of qubits and local Pauli noise, the cost function landscape exponentially flattens, leading to an exponentially vanishing cost gradient, a phenomenon known as Noise-Induced Barren Plateaus (NIBPs) [56]. Thus, noise impedes the training process of VQAs, as in such a setting one requires an exponential number of shots per optimization step to resolve the cost landscape against finite sampling noise. As with other barren plateau effects [58, 59], this exponential scaling does not only arise for gradient-based optimizers but also impacts gradient-free methods [60] and optimizers that use higher-order derivatives [61]. NIBPs represent a serious issue for VQA scalability, and could ultimately be a roadblock for near-term quantum advantage. It is therefore crucial to investigate potential methods to mitigate them.

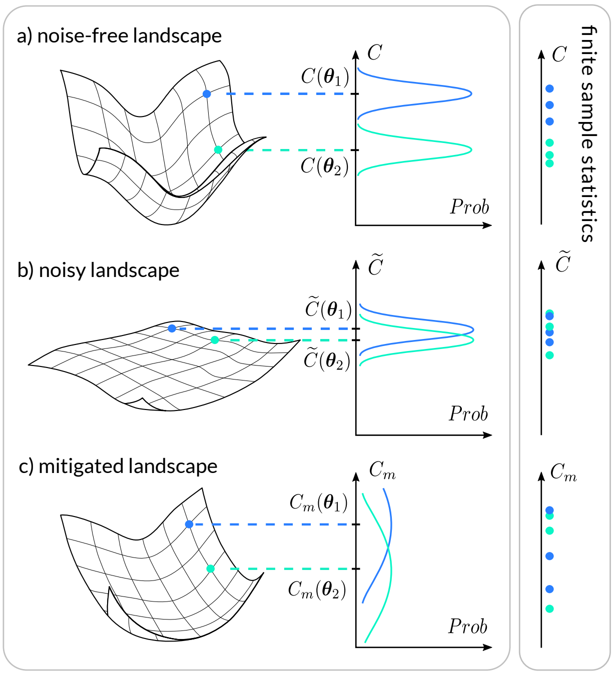
a) noise-free landscape
C
$C(\boldsymbol{\theta}_1)$
$C(\boldsymbol{\theta}_2)$
$Prob$
b) noisy landscape
$\widetilde{C}$
$\widetilde{C}(\boldsymbol{\theta}_1)$
$\widetilde{C}(\boldsymbol{\theta}_2)$
$Prob$
c) mitigated landscape
C_m
$C_m(\boldsymbol{\theta}_1)$
$C_m(\boldsymbol{\theta}_2)$
$Prob$
finite sample statistics
C
$\widetilde{C}$
C_m

Figure 1: **Error mitigation can impair the resolvability of cost function landscapes.** (a): A central primitive in training VQAs is the task of comparing two cost function values ($C(\boldsymbol{\theta}_1)$ and $C(\boldsymbol{\theta}_2)$) on the cost landscape in parameter space. Ideally (with infinite sampling), these cost values correspond to the mean values of some probability distributions (left panel). However, in an experimental setup, one only has a finite shot budget and by collecting measurement statistics one obtains an estimate of the mean values by sampling from these distributions (right panel). (b): The effect of certain types of noise models is to concentrate cost function values. This impedes trainability as any two cost function values ($\widetilde{C}(\boldsymbol{\theta}_1)$ and $\widetilde{C}(\boldsymbol{\theta}_2)$) have small separation and require many shots to accurately distinguish. (c): Error mitigation can mitigate many effects of noise and potentially recover key features of the noise-free cost function. In an ideal scenario, the separation of the mitigated cost values ($C_m(\boldsymbol{\theta}_1)$ and $C_m(\boldsymbol{\theta}_2)$) closely resembles that of the noise-free landscape. However, the caveat is that the variance of statistical outcomes can increase greatly. The effect of this is that the two cost function points can often require even more shots to resolve accurately, compared to the unmitigated case.

Given the great success of EM methods in suppressing error in observable expectation values, it is natural to ask whether EM methods could address NIBPs. More generally, one could simply ask: does it help to use error mitigation during the training process for VQAs? This question is precisely the topic of our article. We remark that error mitigation has been successfully implemented during the VQA training process for a small-scale problem [44]. However, it is an open question as to whether or not EM can resolve large-scale trainability issues associated with cost concentration. This is due to the fact that even though EM can reverse the concentration of cost values, it also increases the statistical uncertainty in the mitigated quantities, as summarized in Figure 1. If the statistical uncertainty increases too quickly, then error mitigation may make it harder to find cost minimizing directions, or reliably compare relative magnitudes of cost values, which are the key tasks in order to reliably train on a cost landscape. Centrally, it is a non-trivial question as to whether or not EM improves the *resolvability* of cost function values. To be clear, we are not solely quantifying the effectiveness of error mitigation in reconstructing the cost landscape given unbounded samples, which is a central question widely studied in the literature. Rather, viewing NIBPs as an exponential sample issue, in this work we compare the sample efficiency of extracting information needed for optimization over the cost landscape using error mitigation versus not using error mitigation at all.

In this work, we investigate the effects of error mitigation on the resolvability of the cost function landscape. First, we consider a broad class of error mitigation protocols and show that, under the class of local Pauli noise that is known to cause NIBPs, in order to reverse exponential cost concentration any such protocol needs to spend resources (e.g., shot resources or number of state copies) scaling at least exponentially in the number of qubits. This suggests that NIBPs are a serious scaling issue that cannot be simply resolved with error mitigation.

Second, we study four specific error mitigation protocols in further detail: Zero-Noise Extrapolation, Virtual Distillation, Probabilistic Error Cancellation, and strategies that implement a linear ansatz which includes Clifford Data Regression. We find that Virtual Distillation can actually decrease the resolvability of the noisy cost landscape, and impede trainability. Under more restrictive assumptions on the cost landscape, we find a similar result for Zero-Noise Extrapolation. We also show that any improvement in the resolvability after applying Probabilistic Error Cancellation under local depolarizing noise exponentially degrades with increasing number of qubits. Finally, for strategies that use a linear ansatz such as Clifford Data Regression, we show that there is no change to the resolvability of any pair of cost values if the same ansatz is used. However, we do observe numerically that Clifford Data Regression increases trainability in some settings. This last observation provides some hope that a careful choice of error mitigation method can be useful. It also suggests that researchers could design and engineer error mitigation methods to enhance VQA trainability.

The rest of the manuscript is structured as follows. Section 2 introduces the framework and notation for our work. We present our theoretical results in Section 3 and our numerical results in Section 4. Finally, our concluding discussions are presented in Section 5. The proofs for our main results are presented in the Appendix.

2 Framework

2.1 Variational Quantum Algorithms

The main goal of Variational Quantum Algorithms (VQAs) is to solve an optimization problem by minimizing a cost function that can be efficiently estimated on a quantum computer. In this work we consider settings where the cost function takes the form

$$C(\boldsymbol{\theta}) = \mathrm{Tr}\left[U(\boldsymbol{\theta})\rho_{in}U\dagger(\boldsymbol{\theta})O\right] . \qquad (1)$$

In the above, given some Hilbert space $\mathcal{H}$, we define the set of density operators $\mathcal{S}(\mathcal{H})$ and set of bounded linear operators $\mathcal{B}(\mathcal{H})$. We then denote $\rho_{in} \in \mathcal{S}(\mathcal{H})$ as the input state, $U(\boldsymbol{\theta}) \in \mathcal{B}(\mathcal{H})$ as a unitary that corresponds to a parametrized quantum circuit with trainable parameters $\boldsymbol{\theta}$, and $O \in \mathcal{B}(\mathcal{H})$ is a Hermitian operator. The Variational Quantum Eigensolver [21], variational quantum compiling [33–35,62], quantum autoencoders [63], and several other VQAs fit under the framework of Eq. (1).

A quantum computer is employed to evaluate the cost function, or gradients thereof, and part of the computational complexity of the algorithm is designated to a classical computer that leverages the power of classical optimizers to solve the problem

$$\arg\min_{\boldsymbol{\theta}} C(\boldsymbol{\theta}) . \qquad (2)$$

The optimization task defined in Eq. (2) has been shown to be NP-hard [64]. Moreover, on top of the typical difficulties associated with solving classical non-convex optimization problems, there are challenges that arise when training the parameters of a VQA due to the quantum nature of the problem itself.

As quantum mechanics is intrinsically a probabilistic theory, one has to deal with shot noise arising from finite sampling when estimating the cost function (or its gradient). This has led to the development of several quantum-aware optimizers that are frugal in the number of shots [65–67]. Additionally, it has been recently shown that certain properties of the cost function can induce so-called barren plateaus, originating due to highly expressive ansatzes [58, 68, 69], global cost functions [59], high levels of entanglement [70, 71], or the controllability of $U(\boldsymbol{\theta})$ [72]. When a cost function exhibits a barren plateau, with high probability the cost function partial derivatives are exponentially suppressed across the landscape. This means that an exponentially large number of shots are needed to navigate the flat landscape and determine a cost-minimizing direction [60, 61].

In this work we investigate the effect of noise and error mitigation techniques in solving the optimization task of Eq. (2). For this purpose we investigate the task of resolving two points on the cost function landscape, as presented in Fig. 1. This is a central primitive in the training process that is utilized at each optimization step, regardless of whether one is using gradient-based or gradient-free methods. In gradient-based methods, a common strategy is to use the parameter shift rule, which constructs partial derivatives from two cost function values [73, 74]. Gradient-free methods such as simplex-based methods also compare two or more cost function values at each optimization step [75, 76]. Thus, this task is a key step for both gradient-based and gradient-free optimizers, and it reflects the ability of the optimizer to find a cost-minimizing direction at each step of the optimization. As discussed below, under a finite shot budget this task becomes harder under cost concentration, leading to trainability issues.

2.2 Effect of noise on the training landscape

Hardware noise can impact the cost function landscape in a variety of ways such as changing the optimal cost function value, shifting the position of minima, and demoting a global minimum to a local minimum. All of the above present further challenges in the training of VQAs. In this section we briefly review some of the literature on the effect of noise on VQAs cost function landscapes. We summarize some of these effects in Fig. 2.

2.2.1 Noise resilience

Certain cost functions have been demonstrated to show optimal parameter resilience under particular noise models [34]. This is a phenomenon where the position of the global cost minimum of the cost landscape is invariant under the action of noise. This has important consequences for trainability. There are many VQAs where the goal is to obtain optimal parameters, rather than the optimal cost value, such as when solving combinatorial optimization problems with the Quantum Approximate Optimization Algorithm (QAOA) [24]. If such cost landscapes display optimal parameter resilience, this leaves open the possibility of noisy training even if the cost value of the global minimum is altered by the noise. In fact, it has recently been shown that a small amount of dephasing errors can recover layerwise training of the QAOA [77]. However, noise can also severely affect the trainability of the landscape in a number of ways, which we summarize below.

2.2.2 Noise-induced cost concentration and noise-induced barren plateaus

Here we summarize the phenomenon of noise-induced cost concentration and noise-induced barren plateaus (NIBPs), as well as introduce some notation that we will use throughout the rest of this manuscript. This was formulated in Ref. [56] for a general class of VQAs and a class of Pauli noise that includes as a special case local depolarizing noise. (See also Refs. [57, 78, 79] for other discussions of the impact of noise.) Consider a model of noise acting through a depth L circuit with n-qubit input state ρ_{in} as

$$\widetilde{\rho} = (\mathcal{N} \circ \mathcal{U}_L \circ \cdots \circ \mathcal{N} \circ \mathcal{U}_1 \circ \mathcal{N})(\rho_{in}) \qquad (3)$$

where $\{\mathcal{U}_k\}_{k=1}^{L}$ denote unitary channels that describe collections of gates that act together in a layer, and $\mathcal{N} = \bigotimes_{i=1}^{n} \mathcal{N}_i$ is an instance of local Pauli channels. In general we can consider different Pauli noise channels in each layer and our theoretical results can be simply extended to such settings, but we do not consider it here for simplicity of presentation. The action of $\mathcal{N}_j$ on a local Pauli operator $\sigma \in \{X, Y, Z\}$ can be expressed as

$$\mathcal{N}_j(\sigma) = q_{\sigma}^{(j)} \sigma \, , \qquad (4)$$

where we assume $-1 < q_X^{(j)}, q_Y^{(j)}, q_Z^{(j)} < 1$ for all qubit labels j. Here, we characterize the noise strength with a single parameter $q = \max_{j,\sigma}\{|q_\sigma^{(j)}|\} < 1$. We denote a noisy cost function as

$$\widetilde{C} = \mathrm{Tr}\left[O\widetilde{\rho}\right] , \tag{5}$$

where O is some Hermitian measurement operator (throughout the article we will use a tilde to denote noisy quantities). In Ref. [56] it was shown that

$$\left|\widetilde{C} - \frac{1}{2^n}\mathrm{Tr}[O]\right| \leq D(q,n) , \tag{6}$$

where $D(q,n) \in \mathcal{O}(q^{\alpha n})$ for some positive constant α if $L \in \Omega(n)$. More generally the quantity on the left-hand side of Eq. (6) vanishes exponentially with increasing circuit depth L for any fixed n. Thus, in the presence of the class of noise models considered, the noisy cost function exponentially concentrates on a fixed value if the depth scales linearly or superlinearly in the number of qubits.

The gradients across the cost function landscape show similar scaling [56], in that they also vanish exponentially in the number of qubits for linear depth circuits, demonstrating a phenomenon known as NIBPs. This implies that the task of accurately determining gradients or cost function differences during the training process requires an exponential number of shots due to the need to resolve quantities to an exponentially small precision.

2.2.3 Cost corruption

In general, a noise model that exhibits cost concentration and NIBPs would not simply uniformly flatten the cost landscape. Instead, we expect noise to additionally alter the cost landscape in many non-trivial ways. We refer to any additional adverse effects on the landscape as cost corruption. For example, it was shown in Ref. [80] that non-unital noise can break the degeneracy of exponentially-occurring global minima, thus proliferating local minima and impacting trainability. In addition, cost functions that do not exhibit optimal parameter resilience [34] limit the quality of noisy optimization, as the optimal parameters of $\widetilde{C}(\boldsymbol{\theta})$ do not correspond to the optimal parameters of $C(\boldsymbol{\theta})$.

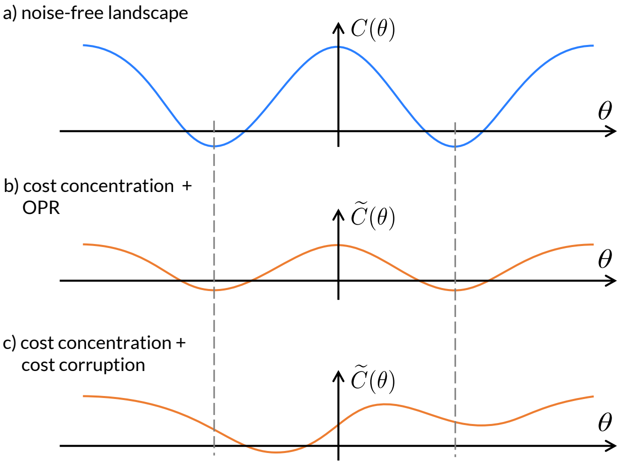

Figure 2: **Schematic of different effects due to noise on cost landscapes.** We present a 1-dimensional slice of a simplified cost landscape corresponding to a single parameter θ. a) Depending on the parameterization strategy, some ansatzes can have degenerate minima. b) Certain types of local Pauli noise can cause the cost landscape to exponentially concentrate on a fixed value. Some can problems display optimal parameter resilience (OPR), where the location of the optimal parameters are invariant under action of the certain noise models. c) Aside from cost concentration, noise can also corrupt the cost landscape by breaking the degeneracy of optimal parameters, and shifting the location of minima.

2.3 Error Mitigation Techniques

We finish the discussion of our framework with a summary of the key features of the error mitigation techniques that we study in this article. For a more detailed review, readers can refer to Refs. [2, 81].

Consider the effects of noise on the cost function in Eq. (1). We suppose the noise can be characterized by a single (scalar) parameter ε and we denote the corresponding noisy state and cost function as $\widetilde{\rho}(\boldsymbol{\theta},\varepsilon)$ and $\widetilde{C}(\boldsymbol{\theta},\varepsilon) = \mathrm{Tr}[\widetilde{\rho}(\boldsymbol{\theta},\varepsilon)O]$ respectively. The goal of error mitigation is to construct an experimental protocol which obtains a mitigated cost function estimator $C_m(\boldsymbol{\theta})$ that approximates the noise-free value $C(\boldsymbol{\theta})$. The protocol to obtain $C_m(\boldsymbol{\theta})$ generally consists of running circuits that modify the original circuit of interest by inserting additional gates, preparing multiple copies of a state, changing the measurement operator, and classical post-processing of the expectation values of these circuits. These different utilizations of resources are summarized in a schematic in Fig. 3.

Error mitigation protocols often lead to a larger variance in the statistical outcomes of each experiment, and thus more shots are required to estimate the error-mitigated cost value $C_m(\boldsymbol{\theta}, \varepsilon)$ to a desired precision compared to the unmitigated noisy value $\widetilde{C}(\boldsymbol{\theta}, \varepsilon)$. This is often quantified by the error mitigation cost, which is defined below.

Definition 1 (Error mitigation cost)**.** *We define the error mitigation cost as*

$$\gamma(\boldsymbol{\theta}, \varepsilon) = \frac{\mathrm{Var}[C_m(\boldsymbol{\theta}, \varepsilon)]}{\mathrm{Var}[\widetilde{C}(\boldsymbol{\theta}, \varepsilon)]}\,, \tag{7}$$

where $\widetilde{C}(\boldsymbol{\theta}, \varepsilon)$ denotes the noisy cost function value corresponding to vector of parameters $\boldsymbol{\theta}$ at noise level ε, and $C_m(\boldsymbol{\theta}, \varepsilon)$ denotes the corresponding error-mitigated quantity.

In certain settings we encounter in our theoretical analyses, $\gamma(\boldsymbol{\theta}, \varepsilon)$ is independent of $\boldsymbol{\theta}$. In other cases, we assume that it is parameter independent, or we seek parameter-independent bounds. Thus, from hereon in this manuscript we will generally drop the parameter dependence of $\gamma(\boldsymbol{\theta}, \varepsilon)$.

We now summarize the error mitigation techniques that we study in this article. We note that recently, unified error mitigation techniques have also been proposed that combine two or more of the protocols that we discuss in this section [82–84]. Our results are also applicable to such strategies, however, we will only review the root strategies here.

2.3.1 Zero-Noise Extrapolation

The goal of Zero-Noise Extrapolation is to run a given circuit of interest at $m+1$ increasing noise levels $\varepsilon < a_1\varepsilon < \ldots < a_m\varepsilon$, and to use information from the resulting expectation values to obtain an estimate of the zero-noise result. Here we summarize the key features of a protocol using Richardson extrapolation [10, 42], and exponential extrapolation [43].

Richardson Extrapolation. Suppose that $\widetilde{C}(\boldsymbol{\theta}_i, \varepsilon)$ admits a Taylor expansion in small noise parameter ε as

$$\widetilde{C}(\boldsymbol{\theta}_i, \varepsilon) = \widetilde{C}(\boldsymbol{\theta}_i, 0) + \sum_{k=1}^{m} p_k(\boldsymbol{\theta}_i)\varepsilon^k + \mathcal{O}(\varepsilon^{m+1})\,, \tag{8}$$

where p_k are unknown parameters and $\widetilde{C}(\boldsymbol{\theta}_i, 0) = C(\boldsymbol{\theta})$ is the zero-noise cost function. By considering the equivalent expansion of $\widetilde{C}(\boldsymbol{\theta}_i, a_1\varepsilon)$ and combining the two equations one obtains

$$\begin{aligned} C_m(\boldsymbol{\theta}_i) &= \frac{a_1\widetilde{C}(\boldsymbol{\theta}_i, \varepsilon) - \widetilde{C}(\boldsymbol{\theta}_i, a_1\varepsilon)}{a_1 - 1} && (9)\\ &= \widetilde{C}(\boldsymbol{\theta}_i, 0) + \mathcal{O}(\varepsilon^2)\,, && (10) \end{aligned}$$

which is a higher-order approximation of $\widetilde{C}(\boldsymbol{\theta}_i, 0)$ compared to simply using $\widetilde{C}(\boldsymbol{\theta}_i, \varepsilon)$. This process can be repeated iteratively m times to obtain an estimator which is accurate up to $\mathcal{O}(\varepsilon^{m+1})$ error.

Exponential extrapolation. In some cases the noisy behavior may not be well-depicted by a Taylor expansion. As an alternative one can consider an exponential model

$$\widetilde{C}(\boldsymbol{\theta}_i, \varepsilon) = r(\boldsymbol{\theta}_i, \varepsilon)^{-t(\boldsymbol{\theta}_i,\varepsilon)}\Big(\sum_{k=0}^{m} p_k(\boldsymbol{\theta}_i)\varepsilon^k + \mathcal{O}(\varepsilon^{m+1})\Big), \tag{11}$$

for some r and t which in general can be functions of ε. For instance, in Ref. [43] it is chosen that $r(\boldsymbol{\theta}_i, \varepsilon)^{-t(\boldsymbol{\theta}_i,\varepsilon)} = e^{-N_g\varepsilon}$ where N_g is the number of gates. In this case, the noise-free cost value is $p_0(\boldsymbol{\theta}_i)$. We can also construct an extrapolation strategy that is tailored towards noisy cost function values that are dominated by NIBP scaling as in Eq. (6), where we model the effects of noise as

$$\begin{aligned} \widetilde{C}(\boldsymbol{\theta}_i, q) = A + q^L\Big(B(\boldsymbol{\theta}_i) + \sum_{k=1}^{m} p_k(1-q)^k \\ + \mathcal{O}\big((1-q)^{m+1}\big)\Big)\,, \end{aligned} \tag{12}$$

where $q < 1$ is the Pauli noise parameter defined in Eq. (4) which equals zero for maximal noise. Here, A is the fixed point of the noise (corresponding to the maximally mixed state) and $A + B(\boldsymbol{\theta}_i)$ is the noise-free cost value. For these two strategies we can similarly construct $C_m(\boldsymbol{\theta}_i)$ as linear combinations of $\{\widetilde{C}(\boldsymbol{\theta}_i, a_i\varepsilon)\}_{i=0}^m$ or $\{\widetilde{C}(\boldsymbol{\theta}_i, q/a_i)\}_{i=0}^m$ to achieve $\mathcal{O}(\varepsilon^{m+1})$ approximations of the zero-noise cost value. We detail these constructions in Section B.1.1 of the Appendix.

2.3.2 Virtual Distillation

Virtual Distillation, also known as Error Suppression by Derangement, was proposed concurrently in Refs. [47] and [46]. In this article we consider the two error mitigation protocols in Ref. [47] (denoted "A" and "B") to respectively prepare

$$C_m^{(A)}(\boldsymbol{\theta}_i) = \mathrm{Tr}[\widetilde{\rho}_i^M O]/\mathrm{Tr}[\widetilde{\rho}_i^M]\,, \tag{13}$$

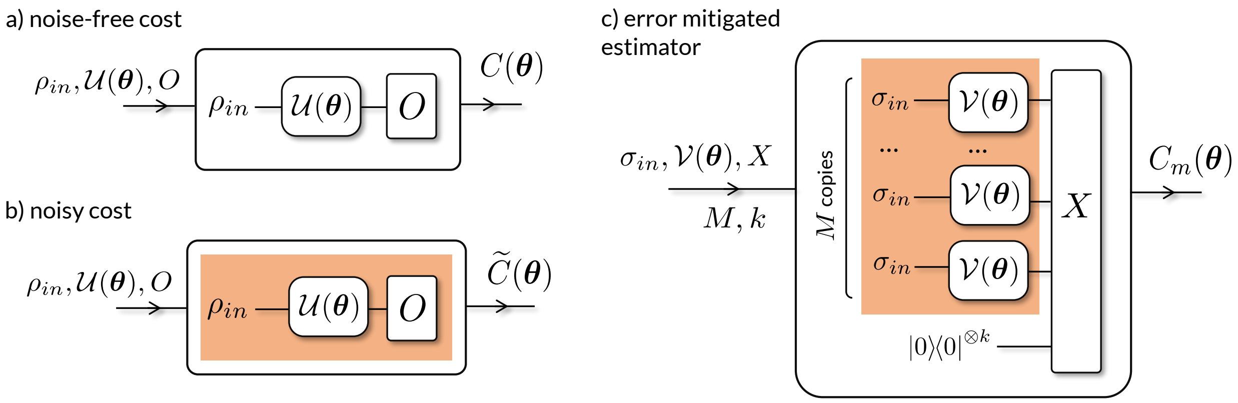
a) noise-free cost
$\rho_{in}, \mathcal{U}(\boldsymbol{\theta}), O$
ρ_{in}
$\mathcal{U}(\boldsymbol{\theta})$
O
$C(\boldsymbol{\theta})$
b) noisy cost
$\rho_{in}, \mathcal{U}(\boldsymbol{\theta}), O$
ρ_{in}
$\mathcal{U}(\boldsymbol{\theta})$
O
$\widetilde{C}(\boldsymbol{\theta})$
c) error mitigated estimator
$\sigma_{in}, \mathcal{V}(\boldsymbol{\theta}), X$
M, k
M copies
σ_{in}
$\mathcal{V}(\boldsymbol{\theta})$
X
$|0\rangle\langle 0|^{\otimes k}$
$C_m(\boldsymbol{\theta})$

Figure 3: **Schematic of resource use in error mitigation.** Noise is indicated by the shaded orange region. (a) Cost function values are obtained by taking input state ρ_{in}, applying parameterized gates which we denote as a unitary channel $\mathcal{U}(\boldsymbol{\theta})$, and measuring the resulting state $\mathcal{U}(\boldsymbol{\theta})(\rho_{in})$ with observable O. (b) Noise can corrupt the gates in the circuit, as well as the state preparation and measurement processes. (c) Error mitigation aims to obtain a good approximation to the noise-free cost $C(\boldsymbol{\theta})$ by employing a number of strategies such as: modifying the gates implemented $\mathcal{U}(\boldsymbol{\theta}) \to \mathcal{V}(\boldsymbol{\theta})$ or the input state $\rho_{in} \to \sigma_{in}$, utilizing multiple copies of the quantum circuit, modifying the measurement operator $O \to X$, and utilizing clean ancillary qubits at the end of the circuit. Many such circuits with different hyperparameters can be run, with their expectation values combined in a post-processing step, in order to construct the final error mitigated cost value $C_m(\boldsymbol{\theta})$. Note that here we have only indicated noise occuring in the initial part of the circuit–this reflects the assumptions of analyses in prior works [46, 47]. As we investigate the limitations of such error mitigation schemes, we keep these assumptions as a "best case" analysis. One feature that distinguishes the approaches to error mitigation studied here from error correction is that error correction allows global access to the larger Hilbert space from the start of the circuit, whereas the error mitigation techniques studied here only allow the possibility for global operations at the end of the circuit.

and

$$C_m^{(B)}(\boldsymbol{\theta}_i) = \mathrm{Tr}[\widetilde{\rho}_i^{M} O]/\lambda_i^M \,, \tag{14}$$

where λ_i is the dominant eigenvalue of $\widetilde{\rho}_i \equiv \widetilde{\rho}(\boldsymbol{\theta}_i)$. The operator $\widetilde{\rho}_i^M$ can be obtained by preparing M copies of $\widetilde{\rho}_i$ in a tensor product state $\widetilde{\rho}_i^{\otimes M}$ and applying a cyclic shift operator. We note that protocol B presumes access to the dominant eigenvalue beforehand, which could potentially be computed via the techniques of Ref. [37].

2.3.3 Probabilistic Error Cancellation

Probabilisitic Error Cancellation utilizes many modified circuit runs in order to construct a quasiprobability representation of the noise-free cost function [42, 43]. We assume that the effect of the noise can be described by a quantum channel $\mathcal{N}$ that occurs after a gate that we denote with unitary channel $\mathcal{U}$. Here we make the simplifying assumption that this is the only gate in the circuit, and we treat the general case in Section B.1.2 of the Appendix, as well as provide a more detailed exposition. The goal of this protocol is to simulate the inverse map $\mathcal{N}^{-1}$. Note that, in general, this will not always correspond to a CPTP map. Despite this fact, if one has a basis of (noisy) quantum channels $\{\mathcal{B}_\alpha\}_\alpha$, corresponding to experimentally available channels, one can expand the inverse map in this basis as $\mathcal{N}^{-1} = \sum_\alpha q_\alpha \mathcal{B}_\alpha$, for some set of $q_\alpha \in \mathbb{R}$. By defining a probability distribution $p_\alpha = |q_\alpha|/G_\mathcal{N}$ where $G_\mathcal{N} = \sum_\alpha |q_\alpha|$, the noise-free expectation value can then be written as a quasiprobability distribution

$$C_{\mathcal{U}(\rho)} = G_\mathcal{N} \sum_\alpha \mathrm{sgn}(q_\alpha)\, p_\alpha \,\mathrm{Tr}\left[\mathcal{B}_\alpha \mathcal{N}\mathcal{U}(\rho_{in}) O\right] , \tag{15}$$

where ρ_{in} is the input state, O is the measurement operator, and $\mathrm{sgn}(q_\alpha)$ denotes the sign of q_α. The idea is that if one has access to the set of CPTP maps $\{\mathcal{B}_\alpha\}_\alpha$ in the noisy native hardware gate set, then one can obtain an estimate of the noise-free cost $C_{\mathcal{U}(\rho)}$ as follows: (1) With probability p_α, prepare the circuit of interest with additional gate $\mathcal{B}_\alpha$ in order to obtain the expectation value $\mathrm{Tr}\left[\mathcal{B}_\alpha \mathcal{N}\mathcal{U}(\rho_{in}) O\right]$. (2) Multiply the result by $\mathrm{sgn}(q_\alpha) G_\mathcal{N}$. (3) Repeat process many times and sum results.

2.3.4 Clifford Data Regression (CDR) and linear ansatz methods

The main idea of linear ansatz methods is to assume that we can approximately reverse the effects of noise with an affine map, and thus we construct a linear ansatz of the form

$$C_m(\boldsymbol{\theta}, \boldsymbol{a}) = a_1(\boldsymbol{\theta})\widetilde{C}(\boldsymbol{\theta}) + a_2(\boldsymbol{\theta}) \,, \qquad (16)$$

where $\boldsymbol{a}(\boldsymbol{\theta}) = (a_1(\boldsymbol{\theta}), a_2(\boldsymbol{\theta}))$ is a vector of parameters to be determined. In general we expect $\boldsymbol{a}$ to be highly dependent on $\boldsymbol{\theta}$. In Ref. [45], the authors use data regression to learn the optimal parameters $\boldsymbol{a}^*(\boldsymbol{\theta})$ with training data comprising of pairs of noise-free and corresponding noisy cost function values $\mathcal{T}_{\boldsymbol{\theta}} = \{(C_j, \widetilde{C}_j)\}_j$, where the circuits are predominantly constructed from Clifford gates. The noise-free cost values can be simulated efficiently on a classical computer whilst the noisy cost values can be evaluated directly on the quantum computer. This strategy is known as Clifford Data Regression.

Other methods have been proposed to learn the optimal parameters $\boldsymbol{a}^*(\boldsymbol{\theta})$. In Ref. [85] the authors further develop the idea of training-based error mitigation by considering alternative training data comprising of fermionic linear optics circuits. One can also model the noise as global depolarizing noise. Under this assumption, $\boldsymbol{a}^*(\boldsymbol{\theta})$ has an exact solution in terms of a single noise parameter. Subsequently, various techniques can be used to estimate the noise parameter [86–90]. Finally, we note that an alternative learning-based method has been proposed in which Clifford data is used to learn the optimal quasi-probability distribution [91]. In this case, our results on probabilistic error cancellation are directly applicable.

2.3.5 Previous results on sampling overhead

It is well known that error mitigation techniques require a larger shot budget than estimating unmitigated expectation values, due to the amplification of statistical variance. Indeed, this has been discussed as part of the original proposal of many error mitigation schemes [42, 46, 47, 91, 92]. For probabilistic error cancellation, in Ref. [93] Xiong et al. investigate the sampling overhead of probabilistic error cancellation in further detail for various noise channels. We stress that all of the aforementioned analysis quantifies the sampling overhead in recovering individual expectation values to constant precision. We note that, to the best of our knowledge, prior to our work the effects of error mitigation in resolving trainability issues due to noise have not been studied.

3 Theoretical Results

We present two sets of theoretical results. First, in Section 3.1 we show that a broad class of error mitigation techniques cannot undo the exponential resource requirement that exponential cost concentration presents. This has implications for both the trainability of noisy VQAs, as well as the accurate estimation of noise-free cost function values in general. Second, in Section 3.2, we work predominantly in the non-asymptotic regime (in terms of scaling in n) and investigate to what extent different error mitigation strategies can improve the resolvability of the noisy cost landscape, assuming that some cost concentration has occurred. For these purposes we introduce a class of quantities which quantify the improvement of the resolvability of the cost function landscape after error mitigation, which we call the relative resolvability (see Defs. 2-4). Using these quantities we study Zero-Noise Extrapolation (Sec. 3.2.2), Virtual Distillation (Sec. 3.2.3), Probabilistic Error Cancellation (Sec. 3.2.4) and linear ansatz methods which include Clifford Data Regression (Sec. 3.2.5). In the settings that we consider, we find that in many cases error mitigation impedes the optimizer's ability to find good optimization steps, and is worse than performing no error mitigation.

3.1 Asymptotic scaling results (exponential estimator concentration)

In this section we show that full mitigation of exponential cost concentration is not possible for a general class of error mitigation strategies. Specifically, we show that one cannot remove the exponential scaling that local Pauli noise incurs without investing exponential resources elsewhere in the mitigation protocol.

We start by remarking that, as summarized in Fig. 3, all of the strategies presented in Sec. 2.3 consist of preparing linear combinations of expectation values of the form

$$E_{\sigma,X,M,k} = \mathrm{Tr}\left[X\left(\sigma^{\otimes M} \otimes |0\rangle\langle 0|^{\otimes k}\right)\right] \,, \qquad (17)$$

for some n-qubit quantum state $\sigma \in S(\mathcal{H})$ that in general can be prepared by a different circuit to that of the state of interest, for $|0\rangle\langle 0| \in S(\mathcal{H}')$ and for some $X \in B(\mathcal{H}^{\otimes M} \otimes \mathcal{H}'^{\otimes k})$. That is, one can prepare multiple copies of a state, prepare different quantum circuits, and apply general measurement operators. In order to generalize the setting further, we also allow the possibility to utilize multiple clean ancillary qubits at the end of the circuit. By considering linear combinations of such quantities, we also account for the ability to post-processing measurement results classically with a linear map, such as is the case with Probabilistic Error Cancellation. In the following theorem we show how quantities of the form (17) concentrate under local Pauli noise of the form (3).

Theorem 1. *Consider an error mitigation strategy that, as a step in its protocol, estimates $E_{\sigma,X,M,k}$ as defined in Eq. (17). Suppose that σ is prepared with a depth L_σ circuit and experiences local Pauli noise according to Eq. (3). Under these conditions, $E_{\sigma,X,M,k}$ exponentially concentrates with increasing circuit depth on a state-independent fixed point as*

$$\left| E_{\sigma,X,M,k} - \mathrm{Tr}\left[X \left(\frac{\mathbb{1}^{\otimes M}}{2^{Mn}} \otimes |0\rangle\langle 0|^{\otimes k} \right) \right] \right| \tag{18}$$

$$\leq G_{\sigma,X,M}(n)\,, \tag{19}$$

where $\mathbb{1} \in B(\mathcal{H})$ is the n-qubit identity operator and

$$G_{\sigma,X,M}(n) = \sqrt{\ln 4}\, \|X\|_\infty M n^{1/2} q^{c(L_\sigma+1)}\,, \tag{20}$$

with noise parameter $q \in [0,1)$ and constant $c = 1/(2\ln 2) \approx 0.72$.

Theorem 1 shows that quantities of the form (17) exponentially concentrate in the depth of the circuit. As we summarize in the schematic in Fig. 3, such quantities generalize expectation values that are prepared by many different error mitigation protocols. We provide a proof of the theorem in Appendix D. We now explicitly demonstrate how Theorem 1 affects the mitigated cost values that these protocols output.

Corollary 1 (Exponential estimator concentration)**.** *Consider an error mitigation protocol that approximates the noise-free cost value $C(\boldsymbol{\theta})$ by estimating the quantity*

$$C_m(\boldsymbol{\theta}) = \sum_{(\sigma(\boldsymbol{\theta}),X,M,k)\in T} a_{X,M,k}\, E_{\sigma(\boldsymbol{\theta}),X,M,k}\,, \tag{21}$$

where each $E_{\sigma,X,M,k}$ takes the form (17) and each $a_{X,M,k} \in \mathbb{C}$. We denote M_{max} and a_{max} as the maximum values of M and $a_{X,M,k}$ respectively accessible from a set T defined by the given protocol. Assuming $\|X\|_\infty \in \mathcal{O}(\mathrm{poly}(n))$, there exists a fixed point F independent of $\boldsymbol{\theta}$ such that

$$|C_m(\boldsymbol{\theta}) - F| \in \mathcal{O}(2^{-\beta n} a_{max} |T| M_{max})\,, \tag{22}$$

for some constant $\beta \geq 1$ if the circuit depths satisfy

$$L_{\sigma(\boldsymbol{\theta})} \in \Omega(n)\,, \tag{23}$$

for all $\sigma(\boldsymbol{\theta})$ in the construction (21). That is, if the depths of the circuits scale linearly or superlinearly in n then one requires at least exponential resources to distinguish C_m from its fixed point, for instance by requiring an exponential number of shots, or by requiring an exponential number of state copies M_{max}.

We note that the assumption $\|X\|_\infty \in \mathcal{O}(\mathrm{poly}(n))$ is satisfied in most settings, and in particular is satisfied for all error mitigation protocols discussed in Sec. 2.3. For instance, in the case of Virtual Distillation, X corresponds to a cyclic shift operator followed by a Pauli observable, and thus $\|X\|_\infty \in \mathcal{O}(1)$. Corollary 1 implies that under conditions that generate a NIBP, in order to distinguish any two cost values with constant probability, one requires resource consumption (in shots or number of state copies) that scales exponentially in the number of qubits. Thus, if one views NIBPs as an exponential resource (shots) issue, Corollary 1 shows that the class of error mitigation schemes considered cannot circumvent this issue. In Appendix D we present a more detailed statement that explains how such resources may be consumed.

Whilst the use of clean ancillary qubits as part of an error mitigation protocol, utilized as in Equation (17) and Fig. 3, has not been widely studied, Corollary 1 rules out the possibility that such resources used at the end of the circuit would offer advantage in countering the exponential scaling effects due to cost concentration. Indeed, upon inspecting (19), the ancilla appear explicitly in the form of the fixed point. This highlights a key difference between many error mitigation strategies and error correction, as error correction utilizes resources (such as a larger Hilbert space) in the middle of the computation, whilst the error mitigation protocols considered here are based on processing states obtained at the end of a noisy computation. Our result leaves open the possibility that novel error mitigation protocols that move beyond the framework of (21) and Fig. 3 can have hope of countering the exponential scaling of exponential cost concentration and NIBPs.

3.2 Non-asymptotic protocol-specific results

In this section we present predominantly non-asymptotic results for Zero-Noise Extrapolation (Sec. 3.2.2), Virtual Distillation (Sec. 3.2.3), Probabilistic Error Cancellation (Sec. 3.2.4), and methods which use a linear ansatz such as Clifford Data Regression (Sec. 3.2.5). For each protocol, we investigate the effect of error mitigation on the resolvability of the cost landscape, for different classes of noisy states. To this end, we first define a class of resolvability measures which quantify how many shots it takes to resolve the cost landscape after applying error mitigation, compared to no mitigation at all. We provide proofs of these theoretical statements, as well as some extensions thereof, in Appendix E.

3.2.1 Definitions

Definition 2 (Relative resolvability for two points)**.** *Consider two locations in parameter space $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and their corresponding points on the cost landscape. Denote the number of shots to resolve these two points from each other to precision proportional to their cost difference with and without error mitigation as N_{EM} and N_{noisy} respectively. We define the relative resolvability for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ at error level ε as*

$$\chi(\boldsymbol{\theta}_{1,2},\varepsilon) = \frac{N_{noisy}(\boldsymbol{\theta}_{1,2},\varepsilon)}{N_{EM}(\boldsymbol{\theta}_{1,2},\varepsilon)} \tag{24}$$

$$= \frac{1}{\gamma(\varepsilon)}\left(\frac{\Delta C_m(\boldsymbol{\theta}_{1,2},\varepsilon)}{\Delta\widetilde{C}(\boldsymbol{\theta}_{1,2},\varepsilon)}\right)^2, \tag{25}$$

where we have used the shorthand notation for functional dependence on $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2$ as $f(\boldsymbol{\theta}_{1,2}) = f(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$, γ is the error mitigation cost as defined in Definition 1, and where we denote

$$\Delta\widetilde{C}(\boldsymbol{\theta}_{1,2},\varepsilon) = \widetilde{C}(\boldsymbol{\theta}_1,\varepsilon) - \widetilde{C}(\boldsymbol{\theta}_2,\varepsilon)\,, \tag{26}$$

$$\Delta C_m(\boldsymbol{\theta}_{1,2},\varepsilon) = C_m(\boldsymbol{\theta}_1,\varepsilon) - C_m(\boldsymbol{\theta}_2,\varepsilon)\,. \tag{27}$$

Our definition of relative resolvability is centered around quantifying the sample overhead of the operational task of distinguishing two states (points in parameter space) via their corresponding cost values. The relative resolvability compares this task for the noisy setting with the error-mitigated setting. In order to obtain Eq. (25) from Eq. (24), we consider the usual formula for the standard error of the sample mean, that is, $\sqrt{\mathrm{Var}[C_m(\boldsymbol{\theta},\varepsilon)]/N_{EM}}$ for mitigated cost values and $\sqrt{\mathrm{Var}[\widetilde{C}(\boldsymbol{\theta},\varepsilon)]/N_{noisy}}$ for non-mitigated cost values. Specifically, we suppose that successful resolution of the two points corresponds to achieving a small enough sample mean error proportional to the difference in exact cost value corresponding to those two points. Said differently, we ask an error mitigated optimizer and a non-mitigated optimizer to resolve a length scale proportional to the separation between two cost function values on the mitigated cost landscape and non-mitigated cost landscape, respectively. The proportionality constant can be thought of as being chosen arbitrarily, as in Eq. (24) we consider a ratio of shots (thus whatever proportionality constant is chosen cancels out).

We see that if $\chi(\boldsymbol{\theta}_{1,2},\varepsilon) > 1$, then $N_{\mathrm{EM}}(\boldsymbol{\theta}_{1,2},\varepsilon) < N_{\mathrm{noisy}}(\boldsymbol{\theta}_{1,2},\varepsilon)$. Thus, error mitigation has successfully increased the resolvability of the cost values corresponding to the cost values at $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Note that this criterion is a necessary but not sufficient condition for error mitigation to reverse the effects of cost concentration on the cost landscape. Namely, it does not require the mitigated landscape to accurately reflect the noise-free landscape, and it does not account for other trainability issues such as proliferation of minima. If $\chi(\boldsymbol{\theta}_{1,2},\varepsilon) < 1$ then this implies that error mitigation has exacerbated the resolvability issues associated with cost concentration and NIBPs, and it has been counterproductive in fixing these trainability issues.

For a general cost function, the relative resolvability of cost function points after mitigation may vary significantly across the landscape, or be different for different choices of ansatzes and noise models. Specifically, we seek a more representative length scale to resolve rather than the cost difference between two arbitrary cost values. This motivates us to seek averaged measures of resolvability. We consider two types of averaging: first, an ("operationally-motivated") average over cost function points across a given cost landscape; second, a ("physically motivated") average over a set of noisy states.

Definition 3 (Average relative resolvability across cost landscape)**.** *Denote* $\boldsymbol{\theta}_* = \text{argmin}_{\boldsymbol{\theta}}\, \widetilde{C}(\boldsymbol{\theta},\varepsilon)$ *the vector of parameters that corresponds to the global noisy cost minimum at noise parameter* ε*. We then define the averaged relative resolvability as*

$$\overline{\chi}(\varepsilon) = \frac{1}{\gamma(\varepsilon)}\left(\frac{\langle \Delta C_m(\boldsymbol{\theta}_{i,*},\varepsilon)\rangle_i}{\langle \Delta\widetilde{C}(\boldsymbol{\theta}_{i,*},\varepsilon)\rangle_i}\right)^2 , \tag{28}$$

where $\langle\cdot\rangle_i$ *denotes the mean over all parameter vectors* $\boldsymbol{\theta}_i$ *accessible with the given ansatz of consideration, and where we denote*

$$\Delta C_m(\boldsymbol{\theta}_{i,*},\varepsilon) = C_m(\boldsymbol{\theta}_i,\varepsilon) - C_m(\boldsymbol{\theta}_*,\varepsilon)\,, \tag{29}$$

$$\Delta\widetilde{C}(\boldsymbol{\theta}_{i,*},\varepsilon) = \widetilde{C}(\boldsymbol{\theta}_i,\varepsilon) - \widetilde{C}(\boldsymbol{\theta}_*,\varepsilon)\,. \tag{30}$$

Definition 3 provides a quantity which evaluates the average performance of a given error mitigation protocol across some given cost landscape. Averaging across a given cost landscape gives a result that is particular to the choice of ansatz, measurement operator and noise model. Alternatively, we can view cost concentration as ultimately physically originating from a concentration of states to the maximally mixed state. In order to evaluate the performance of error mitigation in aiding the resolution of states, we consider a physically-motivated average over the basis of a noisy state. Specifically, we consider an average over noisy states that have the same spectrum. This choice of class of noisy states is also motivated by the fact that a central mechanism of cost concentration and NIBPs under unital Pauli noise is the loss of purity. When we consider the basis-averaged relative resolvability for Virtual Distillation in Section 3.2.3, it will turn out to be bounded by a function of the purity for such states.

Definition 4 (Basis-averaged relative resolvability)**.** *Consider a normalized spectrum* $\boldsymbol{\lambda} \in \mathbb{R}_+^{2^n}$ *which corresponds to the eigenspectrum of some noisy state. We define the 2-design-averaged relative resolvability as*

$$\overline{\overline{\chi}}_{\boldsymbol{\lambda}} = \frac{1}{\gamma(\boldsymbol{\lambda})}\frac{\langle (C_m(\rho_{\boldsymbol{\lambda}},U_i) - \text{Tr}[O]/2^n)^2\rangle_{U_i}}{\langle (\widetilde{C}(\rho_{\boldsymbol{\lambda}},U_i) - \text{Tr}[O]/2^n)^2\rangle_{U_i}}\,, \tag{31}$$

where $\langle\cdot\rangle_{U_i}$ *denotes an average over unitaries* U_i *drawn from a unitary 2-design,* $\rho_{\boldsymbol{\lambda}}$ *is an arbitrarily chosen reference state with spectrum* $\boldsymbol{\lambda}$*, and where we denote*

$$\widetilde{C}(\rho_{\boldsymbol{\lambda}},U_i) = \text{Tr}[U_i\rho_{\boldsymbol{\lambda}}U_i^\dagger O]\,, \tag{32}$$

$$C_m(\rho_{\boldsymbol{\lambda}},U_i) = \text{Tr}[\mathcal{M}(U_i\rho_{\boldsymbol{\lambda}}U_i^\dagger)O]\,, \tag{33}$$

where $\mathcal{M} : S(\mathcal{H}) \mapsto B(\mathcal{H})$ *is a map that describes the action of the error mitigation protocol.*

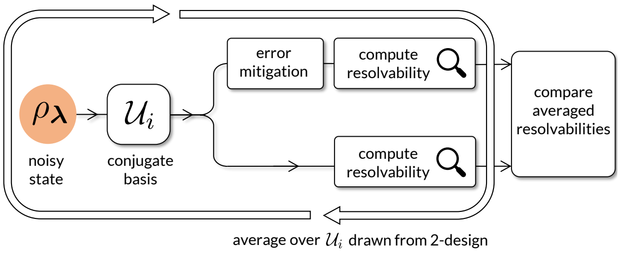
$\rho_{\boldsymbol{\lambda}}$
noisy state
$\mathcal{U}_i$
conjugate basis
error mitigation
compute resolvability
compute resolvability
compare averaged resolvabilities
average over $\mathcal{U}_i$ drawn from 2-design

Figure 4: **Schematic for basis-averaged relative resolvability.** In Definition 4 we consider a broader averaged resolvability measure where the average is taken over a class of noisy states, rather than a cost landscape generated by a particular ansatz. This is constructed from the following game: (1) Prepare a reference noisy state $\rho_{\boldsymbol{\lambda}}$ with spectrum $\boldsymbol{\lambda}$ and conjugate by unitary U_i drawn from a 2-design. (2) Pass the resulting state through the considered error mitigation protocol, and evaluate the resolvability from the fixed point of the noise. (3) Do the same, without error mitigation. (4) Average over the 2-design and compare the averaged resolvabilities.

In Fig. 4 we present a schematic of our basis-averaged relative resolvability.

In the results that follow we do not constrain ourselves to investigating specific models of Pauli noise. Instead, we simply suppose that there exists a noise model that causes some concentration of the cost function onto some state-independent fixed point. Any further assumptions on the effects of the noise model are explicitly specified in the statement of the particular result in question. As a precursor to looking at more generic models of noisy states, we will also find it useful to investigate the performance of error mitigation strategies in the presence of global depolarizing noise of the form

$$\rho \xrightarrow{\mathcal{D}} \widetilde{\rho} = (1-p)\rho + p\frac{\mathbb{1}}{2^n}\,, \tag{34}$$

where $\mathcal{D}$ is the global depolarizing channel and p is the depolarizing probability. Our justification for studying this noise model is twofold. First, global depolarizing noise provides a clean model of cost concentration with no other cost corrupting effects of the noise. Therefore, if a given error mitigation strategy is to mitigate the effects of cost concentration and NIBPs, we expect it to be able to perform well on this noise model. Second, the structure of many error mitigation strategies is directly motivated by the model of global depolarizing noise [45,85–90]. Indeed, many such strategies have been shown to achieve good or perfect performance with this noise model in mitigating noisy cost function values [45–47,86]. However, we stress that trainability may simultaneously get worse, which is what we will now investigate.

3.2.2 Zero-Noise Extrapolation

In this section we present our results on Zero-Noise Extrapolation. Throughout this section, in order to estimate the sample cost of error mitigation we will make the simplifying assumption that

$$\mathrm{Var}[\widetilde{C}(\boldsymbol{\theta}, a\varepsilon)] \geq \mathrm{Var}[\widetilde{C}(\boldsymbol{\theta}, \varepsilon)]\,, \tag{35}$$

for all $\boldsymbol{\theta}$ and $a \geq 1$ that is, the statistical fluctuations in measurement outcomes at the boosted noise level are no smaller than that at the base noise level. We note that similar assumptions are made in the literature for Zero-Noise Extrapolation and Quasi-Probability Methods [43, 81]. In Appendix E.1 we provide intuition as to why we expect this assumption to be true for large a for noise models whose fixed point is the maximally mixed state.

First, we consider the simple model of global depolarizing noise.

Proposition 1 (Relative resolvability of Zero-Noise Extrapolation with global depolarizing noise, 2 noise levels)**.** *Consider a circuit with L instances of global depolarizing noise of the form Eq. (34). Consider a Richardson extrapolation strategy based on Eq. (8), an exponential extrapolation strategy based on Eq. (11) and a NIBP extrapolation strategy based on Eq. (12). We presume access to an augmented noisy circuit where the error probability is exactly increased by factor $a_1 > 1$ as $p \to a_1 p$. Then, we have*

$$\chi_{depol} \leq \frac{\left(c - \frac{(1-a_1 p)^L}{(1-p)^L}\right)^2}{c^2+1}\,, \tag{36}$$

where χ_{depol} is the relative resolvability (see Definition 2) for global depolarizing noise, and where

$$c = \begin{cases} a_1 & \text{for Richardson extrapolation,} \\ \frac{a_1 r(\varepsilon)^{t(\varepsilon)}}{r(a_1\varepsilon)^{t(a_1\varepsilon)}} & \text{for exponential extrapolation,} \\ a_1^{-(L+1)} & \text{for NIBP extrapolation.} \end{cases} \tag{37}$$

Thus, $\chi_{depol} \leq 1$ for all of the above extrapolation strategies with access to 2 noise levels.

We see that for all the above techniques, Zero-Noise Extrapolation with access to 2 noise levels decreases the resolvability of the cost function under global depolarizing noise. Further, if one attempts to directly reverse the exponential scaling of NIBPs that global depolarizing noise incurs, one obtains an exponentially worse relative resolvability. We now consider how resolvability behaves under Zero-Noise Extrapolation on average across the cost landscape, given a generic noise model.

Proposition 2 (Average relative resolvability of Zero-Noise Extrapolation, 2 noise levels)**.** *Consider a Richardson extrapolation strategy based on Eq. (8), an exponential extrapolation strategy based on Eq. (11) and a NIBP extrapolation strategy based on Eq. (12). We presume perfect access to an augmented noisy circuit where the noise rate is increased by factor $a_1 > 1$. We denote $\boldsymbol{\theta}_{\varepsilon*}$ as the parameter corresponding to the global cost minimum at base noise parameter ε. Further denote $\frac{\langle \Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon*}, a_1\varepsilon)\rangle_i}{\langle \Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon*}, \varepsilon)\rangle_i} = z$. Any such noise model has an average relative resolvability*

$$\overline{\chi} \leq \frac{(z-c)^2}{c^2+1}\,, \tag{38}$$

where c is defined in Eq. (37). Thus, under the assumption that $z \leq 1$ and $\langle \Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon*}, a_1\varepsilon)\rangle_i \geq 0$, $\overline{\chi} \leq 1$ for all of the above extrapolation strategies with access to 2 noise levels.

Proposition 2 shows that under mild assumptions of the effect of the noise on the cost landscape, Zero-Noise Extrapolation with access to 2 noise levels impairs the resolvability of the cost landscape. These assumptions have physical meaning: $z \leq 1$ implies that on average the cost concentrates when the noise parameter is boosted, whilst $\langle \Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon*}, a_1\varepsilon)\rangle \geq 0$ implies that the landscape is not heavily corrupted after boosting the noise parameter so that the global minimum at the base noise level remains below the average cost value. We also see that in the presence of exponential cost concentration and NIBPs, the relative resolvability is exponentially small if one attempts to directly reverse the exponential scaling of NIBPs.

In the Appendix, we study a modification of the averaged resolvability in Definition 4 and find that this is bounded by a function of the purity of the noisy states, such that the resolvability decreases if purity decreases with increasing noise level. This result, along with the proofs of the above propositions, can be found in Appendix E.1. Finally, we remark that in the above results we consider a scenario where the Richardson, exponential or NIBP extrapolation strategies utilize expectation values from only two noise levels. In Appendix E.1.3 we show that similar results may be obtained for Richardson extrapolation with access to 3 distinct noise levels.

3.2.3 Virtual Distillation

Here we present our results on Virtual Distillation. Similar to our results for Zero-Noise Extrapolation, throughout this section we make the assumption that the statistical uncertainty of the measurement outcomes of the circuit that prepares $\mathrm{Tr}[\widetilde{\rho}_i^M O]$ are no smaller than that of $\mathrm{Tr}[\widetilde{\rho}_i O]$ for any choice of parameters $\boldsymbol{\theta}_i$. In Appendix E.2 we provide some intuition for this assumption.

In the following proposition we start again with the simple model of global depolarizing noise.

Proposition 3 (Relative resolvability of Virtual Distillation with global depolarizing noise)**.** *Consider global depolarizing noise of the form in Eq.* (34) *acting on some n-qubit pure state ρ with error probability p. We consider the two error mitigation protocols of Ref. [47] (denoted "A" and "B") to respectively prepare* (13) *and* (14)*, using M copies of a quantum state. The relative resolvabilities to resolve any two arbitrary cost function points satisfy*

$$\chi_{depol}^{(A)} \leq \chi_{depol}^{(B)} = \Gamma(n, M, p)\,, \tag{39}$$

for all $n \geq 1$, $M \geq 2$, $p \in [0,1]$, and where

$$\Gamma(n, M, p) \leq 1\,, \tag{40}$$

is a monotonically decreasing function in M (with asymptotically exponential decay) for $n \geq 1$, $M \geq 2$. Within this region the bound is saturated as $\Gamma(1,2,p) = 1$ for all p.

Proposition 3 shows that Virtual Distillation decreases the resolvability of cost landscapes suffering from global depolarizing noise. Moreover, as the number of state copies M increases, the effect worsens. We find similar results in the following proposition when considering averaged resolvabilities over a class of noisy states.

Proposition 4 (Average relative resolvability of Virtual Distillation)**.** *Consider an error mitigation protocol that prepares estimator $C_m(\boldsymbol{\theta}_i) = \mathrm{Tr}[\widetilde{\rho}_i^M O]/\mathrm{Tr}[\widetilde{\rho}_i^M]$ from some noisy parameterized quantum state $\widetilde{\rho}_i \equiv \widetilde{\rho}(\boldsymbol{\theta}_i)$. Consider the average relative resolvability $\overline{\overline{\chi}}_{\boldsymbol{\lambda}}$ for noisy states of some spectrum $\boldsymbol{\lambda}$ with purity P as defined in Definition 3. We have*

$$\overline{\overline{\chi}}_{\boldsymbol{\lambda}} \leq G(n, M, P) \leq 1\,, \tag{41}$$

where $G(n, M, P)$ is a monotonically decreasing function in M (with asymptotically exponential decay) for all $n \geq 1$, $M \geq 2$. Within this region the bound is saturated as $G(1, 2, P) = 1$ for all P, and as $G(n, M, 1) = 1$ for all $n \geq 1, M \geq 2$.

We present the explicit forms of $\Gamma(n, M, p)$ and $G(n, M, P)$ as well as a proof of the above propositions in Appendix E.2. In Appendix E.2.3 we also show that outside of the highly mixed regime, the bound in Proposition 4 decreases with decreasing noisy state purity. This indicates that within such settings, the greater the loss of purity due to noise, the worse the impact on resolvability is after error mitigation with Virtual Distillation.

3.2.4 Probabilistic Error Cancellation

Here we present our results for Probabilistic Error Cancellation. We utilize the optimal quasiprobability decompositions studied in Ref. [94], and the proofs can be found in Section E.3 of the Appendix.

Proposition 5 (Relative resolvability of Probabilistic Error Cancellation under global depolarizing noise)**.** *Consider a quasi-probability method that corrects global depolarizing noise of the form* (34)*. For any pair of states corresponding to points on the cost function landscape, the optimal quasiprobability scheme gives*

$$\chi_{depol} = \frac{2^{2n}}{2^{2n} - p(2-p)} \geq 1\,, \tag{42}$$

for all $n \geq 1$, $p \in [0,1]$, which is achieved with access to noisy Pauli gates.

Proposition 5 shows that for the special case of global depolarizing noise, Probabilistic Error Cancellation actually improves the resolvability of the noisy cost landscape. However, this improvement is generally small and is decreasing quickly with the number of qubits n. For instance, for $n = 1$, χ_{depol} has maximum value 4/3 (achieved in the limit of maximum depolarization probability). For $n = 2$, χ_{depol} has maximum value ≈ 1.07. In the limit of large n, χ_{depol} tends to 1.

We extend this study to local depolarizing noise in Appendix E.3. We find that for a single instance of local depolarizing noise, Probabilistic Error Correction can either improve resolvability or worsen it, depending on the strength of concentration of the cost. In addition, we show in the following proposition that if one wishes to mitigate all the noisy gates in the circuit and one has NIBP scaling, the improvement due to Probabilistic Error Cancellation degrades exponentially, and ultimately for large problem sizes this impairs resolvability.

Proposition 6 (Scaling of Probabilistic Error Cancellation with local depolarizing noise)**.** *Consider local depolarizing noise with depolarizing probability p acting in L layers through a depth L circuit as in Eq. (3). Suppose that the effect of this noise is to cause cost concentration*

$$\langle \Delta \widetilde{C}(\boldsymbol{\theta}_{i,*}) \rangle_i = A q^L \langle \Delta C(\boldsymbol{\theta}_{i,*}) \rangle_i \,, \tag{43}$$

for some constant A and noise parameter $q \in [0,1)$. The optimal quasiprobability method to mitigate the depolarizing noise in the circuit yields

$$\overline{\chi} = \frac{1}{A^2 q^{2L}} \left(Q(p)\right)^{nL} \,, \tag{44}$$

where $Q(p) = 1 - \frac{3p(2-p)}{4-p(2-p)} \in [0,1)$ for $p \in (0,1]$.

We note that it is known that noisy cost differences under local depolarizing noise are known to be at best as large as $\langle \Delta \widetilde{C}(\boldsymbol{\theta}_{i,*}) \rangle_i \propto (1-p)^L \langle \Delta C(\boldsymbol{\theta}_{i,*}) \rangle_i$ [56]. Thus, Eq. (43) gives the best possible scaling of noisy cost differences allowed by (6). Proposition 6 shows that if this exponential scaling is no worse than Eq. (43), then under local depolarizing noise the relative resolvability has unfavourable scaling with respect to system size, for any depth circuit.

3.2.5 Linear ansatz methods

In Proposition 7 we consider a scenario where the same linear ansatz (16) is applied to two points on the noisy cost landscape. For Clifford Data Regression this is a reasonable assumption in scenarios where one is comparing two points that are close in parameter space, for instance, when a simplex-based optimizer is exploring a small local region. However, we remark this is not always true in general settings.

Proposition 7 (Linear ansatz methods)**.** *Consider any error mitigation strategy that mitigates noisy cost function value $\widetilde{C}(\boldsymbol{\theta})$ by constructing an estimator $C_m(\boldsymbol{\theta})$ of the form (16). For any two noisy cost function points to which the same ansatz is applied, we have*

$$\chi = 1 \,, \tag{45}$$

for any noise process.

Corollary 2 (Linear ansatz methods under global depolarizing noise)**.** *Under global depolarizing noise, the optimal linear ansatz gives $\chi = 1$ for any pair of cost function points.*

Corollary 2 comes simply by noting that the optimal choice of linear ansatz under global depolarizing noise corrects the noise exactly and is state independent [45]. The above results imply that in some settings CDR has a neutral effect on the resolvability of the cost function landscape. This opens up the possibility that in practical settings CDR can improve the trainability of cost landscapes, if it can remedy other cost corrupting effects due to noise outside of cost concentration. This motivates our numerical studies of CDR, which we present in the following section.

4 Numerical Results

As discussed in Sec. 3.2, in many settings, current state-of-the-art error mitigation methods do not mitigate the effects of cost concentration. Nevertheless, as discussed in Sec. 2.2.3, trainability of VQAs is also affected by other cost-corrupting effects. We expect that error mitigation can reverse some of the effects due to cost corruption that affect the trainability of VQAs when the effects of cost concentration are not too severe. In this section, we numerically investigate the effects of

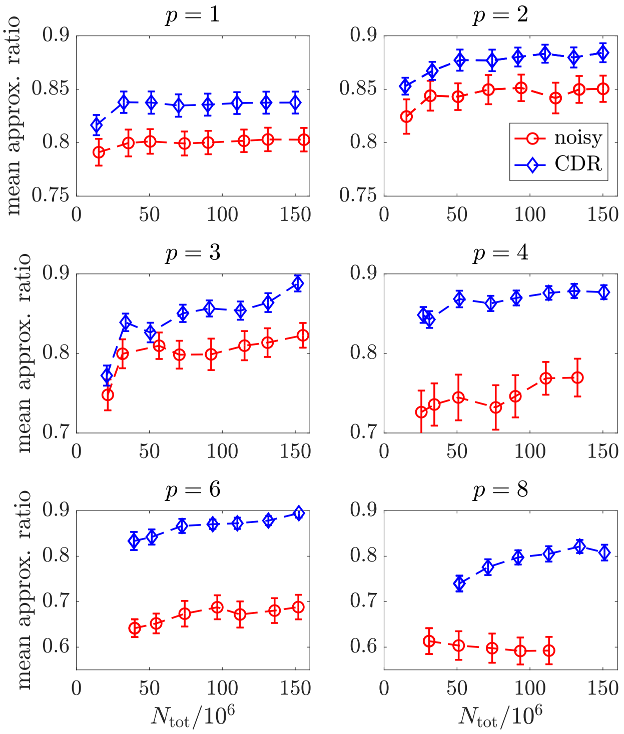

Figure 5: **Comparing CDR-mitigated and noisy optimization for 5-qubit Max-Cut QAOA.** We plot the approximation ratio of solutions for noisy (red circles) and CDR-mitigated (blue diamonds) optimization of Max-Cut QAOA for 5 qubits. Different panels show results for different numbers of QAOA rounds p plotted versus total number of shots $N_{\rm tot}$ spent on the optimization of a MaxCut problem. Here, we compute the approximation ratios using exact $H_{\rm MaxCut}$ energies to benchmark quality of the noisy and CDR-mitigated optimization. The approximation ratio is defined as the ratio of a given solution's energy to the true ground state energy. A higher approximation ratio indicates better solution quality. For each p we average the approximation ratio over 36 MaxCut graphs chosen randomly from Erdös-Rényi ensemble. The error bars show a standard deviation of the mean computed as a standard deviation of the ratio for a graph sample divided by a square root of the number of graphs. For all p and $N_{\rm tot}$ values we see an advantage of the CDR-mitigated optimization over noisy optimization.

error mitigation on trainability in such a setting to provide possible evidence towards beneficial effects of error mitigation. To this end, we focus on CDR as in some settings it does not worsen the effects of cost concentration, as shown in Sec. 3.2.5. While we use this result as a guiding heuristic for the choice of the error mitigation method, we note that a direct comparison of the performance of various methods would be necessary to establish the optimal method for a particular optimization task.

We perform our numerical experiments by simulating the Quantum Approximate Optimization Algorithm (QAOA) [24] for 5-qubit and 8-qubit MaxCut problems. For $n = 5$ we use a realistic noise model of an IBM quantum computer [95], which has been obtained by gate set tomography of IBM's Ourense quantum device. In the case of $n = 8$ we modify the noise model taking a convex combination of the IBM's Ourense and noiseless process matrices to reduce the noise strength by a factor of 5 with respect to the real device. This noise reduction was necessary to ensure trainability of the ansatz as for our problem larger n implies more layers of native gates in the optimized circuits and consequently larger cost concentration. Furthermore, we assume here linear connectivity of the simulated quantum computer.

A MaxCut problem is defined for a graph $G = (V, E)$ of nodes V and edges E. The problem is to find a bipartition of the nodes into two sets which maximizes the number of edges connecting the sets. This problem can be reformulated as finding the ground state of a Hamiltonian

$$H_{\rm MaxCut} = -\frac{1}{2}\sum_{ij\in E}(\mathbb{1} - Z_i Z_j), \qquad (46)$$

where Z_i, Z_j are Pauli Z matrices. Here we consider graphs with $n = 5$ ($n = 8$) vertices, and with 36 (30) randomly generated instances, respectively. The instances are obtained according to the Erdös-Rényi model [96], where for each pair of vertices in the graph there is a connecting edge with probability 0.5.

To approximate the ground state of $H_{\rm MaxCut}$ we simulate the QAOA for number of rounds ranging from $p = 1$ to 8. The QAOA ansatz applied to the input state is given as

$$\prod_{j=p,p-1\ldots,1} e^{i\beta_j H_M} e^{i\gamma_j H_{\rm MaxCut}} (|+\rangle)^{\otimes n}, \qquad (47)$$

where $H_M = \sum_j X_j$, X_j are Pauli X matrices, we denote $|+\rangle = (|0\rangle + |1\rangle)/\sqrt{2}$, and β_j, γ_j are variational parameters. We minimize the cost function $\langle H_{\rm MaxCut}\rangle$ using the Nelder-Mead algorithm [75] and choose the initial values of β_j, γ_j randomly. We perform the optimization with shot budgets ranging from $N_{\rm tot} = 10^7$ to 1.5×10^8 for $n = 5$ and from $N_{\rm tot} = 10^7$ to 7×10^7 for $n = 8$. We define $N_{\rm tot}$ as total number of shots spent on the optimization. We detail the optimization procedure in Appendix F. In our numerics, the values of $N_{\rm tot}$ are chosen to enable an optimization runtime which is feasible with current quantum computers.[1] To quantify the quality of the noisy or CDR-mitigated optimization we compute approximation ratios of the solutions using the exact expectation value of $\langle H_{\mathrm{MaxCut}}\rangle$. The approximation ratio is defined here as the ratio of a given solution's energy to the true ground state energy.

We gather our numerical results for CDR at $n = 5$ in Fig. 5. In the figure we plot the approximation ratio averaged over 36 randomly chosen graphs versus N_{tot}. We compare the quality of the solutions of noisy (unmitigated) and CDR-mitigated optimization and find that CDR-mitigated optimization outperforms noisy optimization for all considered p and N_{tot} values. We observe that the solutions for $p = 2$ outperform those for $p = 1$ for both CDR-mitigated and noisy optimization. The quality of $p > 2$ solutions decline with increasing p for noisy optimization, while it remains approximately the same for $p = 2$ to 6 for CDR-mitigated optimization. With CDR-mitigated optimization we see a decrease in quality of solution for the largest considered $p = 8$.

In Fig. 6 we gather CDR results for 30 instances of Erdös-Renyi graphs at $n = 8$ and $p = 1 - 4$ plotting again the approximation ratio averaged over instances versus N_{tot}. Similar to the case of $n = 5$ we typically see an advantage of the CDR mitigated optimization over noisy optimization, although the improvement in approximation ratio is smaller than observed for $n = 5$. This result underscores the need for more detailed investigation of the properties of an optimization problem which make it favorable for error-mitigated optimization. We leave such an investigation for future work. For the deepest $p = 3, 4$ ansatze we see that performance of both the noisy and CDR-mitigated optimization is degraded in comparison to the shallower $p = 2$ case.

The numerical results presented here are obtained for circuits shallow enough to be trainable while using the CDR-mitigated cost function. Therefore they are outside of the NIBP scaling regime. As discussed in Section 2.2, even outside the NIBP regime noise may adversely impact trainability by corrupting the cost function landscape, which error mitigation has a chance to remedy. Our results give hope that CDR-mitigated optimization may overall offer a trainability advantage for problems with such cost function landscape corruption.

As discussed in Section 3.2 optimizing an error mitigated cost function is not guaranteed to outperform its noisy optimization even outside the NIBP regime. Indeed, in Appendix F.2 we find that for the $p = 2, 4$ MaxCut graphs and moderate N_{tot} used here, VD-mitigated optimization does not outperform noisy optimization. In Appendix F.3 we reach a similar conclusion for optimization mitigated with Zero-Noise Extrapolation when increasing noise strength digitally by widely-used CNOT identity insertions [98].

We note that this conclusion may be problem-dependent. In particular, for a sufficiently large N_{tot}, an idealized error mitigation method that perfectly corrects the expectation values in the limit of an infinite shot number should improve optimization quality as the cost corruption effects will become dominant. For a realistic case, error mitigation has a bias that may depend on a problem choice, noise, and even error mitigation method's implementation details. For example, it has been found that Zero-Noise Extrapolation's bias depends on the method of increasing the noise strength [99]. Therefore, we expect all those factors to be relevant when the shot number is large enough that the cost corruption limits VQA's performance. Consequently, caution is necessary when judging the power of a particular error mitigation approach compared to others in removing the optimization landscape corruption based on a few test cases.

5 Discussion

Noise can exponentially degrade the trainability of linear (or superlinear) depth Variational Quantum Algorithms (VQAs) by flattening the cost landscape, thus requiring an exponential precision in system size (and therefore exponential shot budget) to resolve its features [56, 57]. This limits the scope for achieving possible quantum

[1] To approximately estimate the time required to run the optimization we assume a delay time of 250 μs in between shots which is a default setting of current IBM quantum computers [97]. Furthermore, we assume that the circuit compilation and execution time is negligible. This assumption is justified for sufficiently shallow circuits and sufficiently large numbers of shots per circuit [97]. Under such assumptions for shot budgets used in our 5-qubit MaxCut QAOA numerics we obtain times ranging from 1 to 14 hours for $n = 5$ and 14 to 140 hours for $n = 8$.

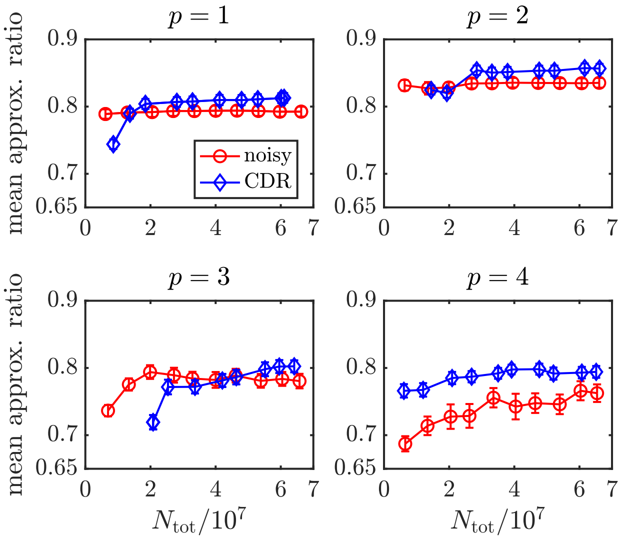

Figure 6: **Comparing CDR-mitigated and noisy optimization for 8-qubit Max-Cut QAOA.** Similar to Fig. 5, here we plot the approximation ratio of solutions for noisy and CDR-mitigated optimization averaged over 30 randomly chosen graphs from Erdös-Rényi ensemble. The error bars are computed as in Fig. 5.

advantage with VQAs. At present there are no known strategies to avoid this exponential scaling completely aside from pursuing algorithms with sublinear circuit depth, and current strategies to mitigate this effect consist only of reducing hardware noise rates. Thus, it is a pressing challenge to search for possible solutions to this problem. Error mitigation strategies emerge as a natural candidate to tackle this problem under near-term constraints.

In this work we investigate the effects of error mitigation on the trainability of noisy cost function landscapes in two regimes. We note that despite the fact it is known that error mitigation increases shot budgets, it has been a priori unclear whether error mitigation strategies in general can have a positive or negative contribution towards this problem, as it relies on a careful balance of how effectively errors are mitigated, compared to how quickly statistical uncertainty is amplified. First, we work in the asymptotic regime (in terms of scaling with system size) and find that if a VQA is suffering from exponential cost concentration, requiring an exponential number of shots to accurately resolve cost values, then a broad class of error mitigation strategies (including as special cases Zero-Noise Extrapolation, Virtual Distillation, Probabilistic Error Cancellation, Clifford Data Regression) cannot remove this exponential scaling. Within the considered paradigm, this exponential scaling implies that at least an exponential number of resources needs to be spent in order to extract accurate information from the cost landscape in order to find a cost-minimizing optimization direction. In Corollary 1 we identify circuit samples (or shots) as well as number of copies of a quantum state as two such resources.

Second, we move out of the asymptotic regime and investigate whether or not particular error mitigation protocols can improve the resolvability of noisy cost landscapes. Should such a landscape be burdened with exponential cost concentration, this would correspond to an improvement in the coefficient in the exponential scaling. Our results indicate that some error mitigation protocols can worsen the resolvability, and ultimately the trainability, of cost landscapes in certain settings. In particular, in Propositions 3 and 4 we show analytically that Virtual Distillation impairs resolvability with worsening resolvability as the number of state copies increases. We obtain similar results for Zero-Noise Extrapolation in Propositions 1 and 2 under some assumptions of the cost landscape. Numerical analysis of a particular MaxCut problem for a moderate shot number indicates that trainability is impaired for Virtual Distillation and for some Zero-Noise Extrapolation strategies.

For the considered problem, Clifford Data Regression (CDR) distinguishes itself from the other error mitigation techniques considered in this article, as in contrast to the other protocols it does not necessarily increase the statistical uncertainty of cost values more than it reverses their concentration (Probabilistic Error Cancellation also improves the resolvability under a global depolarizing noise assumption, but the scaling quickly deteriorates under a local depolarizing noise assumption). This is reflected in the fact that as it only uses a linear ansatz, CDR has neutral impact on resolvability (Proposition 7). However, it is also known that CDR can remedy the effects of more complex noise models. This suggests that CDR could resolve trainability issues arising due to corruptions of the cost function outside of cost concentration, whilst having a neutral effect on cost concentration itself, and thus overall improve trainability. In the numerical example studied, presented in Fig. 5, we observe this to be the case. This points to deeper future work studying the mechanisms that allow error mitigation to improve the trainability of noisy cost landscapes. Such work should consider a wide range of problems and state-of-the-art hardware implementations as the bias of error mitigation methods (which may limit reversing cost function corruption) has been shown to depend on the problem choice and details of the methods' implementation. [99, 100]. We note that here we have also disregarded the burden of training data for CDR, which is an important consideration.

Finally, we identify that the broad class of error mitigation protocols we study in our asymptotic analysis all only consist of post-processing expectation values of noisy circuits, as summarized in Fig. 3. This gives intuition as to why they cannot escape the exponential scaling of noise-induced barren plateaus (NIBPs). However, the theory of error correction indicates that with sufficient resources NIBPs can indeed be avoided. This gives hope that there can exist novel error mitigation strategies that move beyond the framework of the protocols considered in this article and thereby avoid the exponential impairment to trainability that NIBPs present.

6 Code avaialability

Further implementation details are available from the authors upon request.

7 Acknowledgements

This work was supported by the Quantum Science Center (QSC), a National Quantum Information Science Research Center of the U.S. Department of Energy (DOE). SW was supported by the U.S. DOE, Office of Science, Office of Advanced Scientific Computing Research, under the Quantum Computing Application Teams program. SW was also partially supported by the Samsung GRP grant. Piotr C. and AA were supported by the Laboratory Directed Research and Development (LDRD) program of LANL under project numbers 20190659PRD4 (Piotr C.) and 20210116DR (AA and Piotr. C). MC acknowledge support from the Center for Nonlinear Studies at Los Alamos National Laboratory. MC and LC were also initially supported by the LDRD program of LANL under project number 20190065DR. PJC also acknowledges initial support from the LANL ASC Beyond Moore's Law project. This research used resources provided by the Los Alamos National Laboratory Institutional Computing Program, which is supported by the U.S. Department of Energy National Nuclear Security Administration under Contract No. 89233218CNA000001.

Note added. Concurrently, a related work, Ref. [101] appeared in the literature. They establish a similar result to Theorem 1, bounding the sample overhead to mitigate local depolarizing noise in an L-layered circuit. More broadly, the authors establish fundamental bounds on the sampling overhead of error mitigation schemes. In contrast, the rest of our results focuses on the ability of error mitigation to aid optimization of variational quantum algorithms, which depends on a trade-off between sample overheads and mitigation ability.

References

[1] Jarrod R McClean, Jonathan Romero, Ryan Babbush, and Alán Aspuru-Guzik. "The theory of variational hybrid quantum-classical algorithms". New Journal of Physics **18**, 023023 (2016).

[2] M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. "Variational quantum algorithms". Nature Reviews Physics **3**, 625–644 (2021).

[3] Sam McArdle, Tyson Jones, Suguru Endo, Ying Li, Simon C Benjamin, and Xiao Yuan. "Variational ansatz-based quantum simulation of imaginary time evolution". npj Quantum Information **5**, 1–6 (2019).

[4] Harper R Grimsley, Sophia E Economou, Edwin Barnes, and Nicholas J Mayhall. "An adaptive variational algorithm for exact molecular simulations on a quantum computer". Nature Communications **10**, 1–9 (2019).

[5] Cristina Cirstoiu, Zoe Holmes, Joseph Iosue, Lukasz Cincio, Patrick J. Coles, and Andrew Sornborger. "Variational fast forwarding for quantum simulation beyond the

coherence time". npj Quantum Information **6**, 1–10 (2020).

[6] Benjamin Commeau, M. Cerezo, Zoë Holmes, Lukasz Cincio, Patrick J. Coles, and Andrew Sornborger. "Variational hamiltonian diagonalization for dynamical quantum simulation". arXiv preprint arXiv:2009.02559 (2020).

[7] Joe Gibbs, Kaitlin Gili, Zoë Holmes, Benjamin Commeau, Andrew Arrasmith, Lukasz Cincio, Patrick J. Coles, and Andrew Sornborger. "Long-time simulations with high fidelity on quantum hardware". arXiv preprint arXiv:2102.04313 (2021).

[8] Yong-Xin Yao, Niladri Gomes, Feng Zhang, Thomas Iadecola, Cai-Zhuang Wang, Kai-Ming Ho, and Peter P Orth. "Adaptive variational quantum dynamics simulations". arXiv preprint arXiv:2011.00622 (2020).

[9] Suguru Endo, Jinzhao Sun, Ying Li, Simon C Benjamin, and Xiao Yuan. "Variational quantum simulation of general processes". Physical Review Letters **125**, 010501 (2020).

[10] Y. Li and S. C. Benjamin. "Efficient variational quantum simulator incorporating active error minimization". Phys. Rev. X **7**, 021050 (2017).

[11] Jonathan Wei Zhong Lau, Kishor Bharti, Tobias Haug, and Leong Chuan Kwek. "Quantum assisted simulation of time dependent hamiltonians". arXiv preprint arXiv:2101.07677 (2021).

[12] Kentaro Heya, Ken M Nakanishi, Kosuke Mitarai, and Keisuke Fujii. "Subspace variational quantum simulator". arXiv preprint arXiv:1904.08566 (2019).

[13] Xiao Yuan, Suguru Endo, Qi Zhao, Ying Li, and Simon C Benjamin. "Theory of variational quantum simulation". Quantum **3**, 191 (2019).

[14] Maria Schuld, Alex Bocharov, Krysta M Svore, and Nathan Wiebe. "Circuit-centric quantum classifiers". Physical Review A **101**, 032308 (2020).

[15] Guillaume Verdon, Michael Broughton, and Jacob Biamonte. "A quantum algorithm to train neural networks using low-depth circuits". arXiv preprint arXiv:1712.05304 (2017).

[16] Jonathan Romero and Alán Aspuru-Guzik. "Variational quantum generators: Generative adversarial quantum machine learning for continuous distributions". Advanced Quantum Technologies **4**, 2000003 (2021).

[17] Edward Farhi and Hartmut Neven. "Classification with quantum neural networks on near term processors". arXiv preprint arXiv:1802.06002 (2018).

[18] Kerstin Beer, Dmytro Bondarenko, Terry Farrelly, Tobias J. Osborne, Robert Salzmann, Daniel Scheiermann, and Ramona Wolf. "Training deep quantum neural networks". Nature Communications **11**, 808 (2020).

[19] Iris Cong, Soonwon Choi, and Mikhail D Lukin. "Quantum convolutional neural networks". Nature Physics **15**, 1273–1278 (2019).

[20] Edward Grant, Marcello Benedetti, Shuxiang Cao, Andrew Hallam, Joshua Lockhart, Vid Stojevic, Andrew G Green, and Simone Severini. "Hierarchical quantum classifiers". npj Quantum Information **4**, 1–8 (2018).

[21] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J Love, Alán Aspuru-Guzik, and Jeremy L O'brien. "A variational eigenvalue solver on a photonic quantum processor". Nature Communications **5**, 1–7 (2014).

[22] Bela Bauer, Dave Wecker, Andrew J Millis, Matthew B Hastings, and Matthias Troyer. "Hybrid quantum-classical approach to correlated materials". Physical Review X **6**, 031045 (2016).

[23] Tyson Jones, Suguru Endo, Sam McArdle, Xiao Yuan, and Simon C Benjamin. "Variational quantum algorithms for discovering hamiltonian spectra". Physical Review A **99**, 062304 (2019).

[24] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. "A quantum approximate optimization algorithm". arXiv preprint arXiv:1411.4028 (2014).

[25] Zhihui Wang, S. Hadfield, Z. Jiang, and E. G. Rieffel. "Quantum approximate optimization algorithm for MaxCut: A fermionic view". Physical Review A **97**, 022304 (2018).

[26] Gavin E Crooks. "Performance of the quantum approximate optimization algorithm on the maximum cut problem". arXiv preprint arXiv:1811.08419 (2018).

[27] Stuart Hadfield, Zhihui Wang, Bryan O'Gorman, Eleanor G Rieffel, Davide Venturelli, and Rupak Biswas. "From the quantum approximate optimization algorithm to a quantum alternating operator ansatz". Algorithms **12**, 34 (2019).

[28] Carlos Bravo-Prieto, Ryan LaRose, M. Cerezo, Yigit Subasi, Lukasz Cincio, and Patrick Coles. "Variational quantum linear solver". arXiv preprint arXiv:1909.05820 (2019).

[29] Xiaosi Xu, Jinzhao Sun, Suguru Endo, Ying Li, Simon C Benjamin, and Xiao Yuan. "Variational algorithms for linear algebra". Science Bulletin **66**, 2181–2188 (2021).

[30] Bálint Koczor, Suguru Endo, Tyson Jones, Yuichiro Matsuzaki, and Simon C Benjamin. "Variational-state quantum metrology". New Journal of Physics (2020).

[31] Johannes Jakob Meyer, Johannes Borregaard, and Jens Eisert. "A variational toolbox for quantum multi-parameter estimation". NPJ Quantum Information **7**, 1–5 (2021).

[32] Eric Anschuetz, Jonathan Olson, Alán Aspuru-Guzik, and Yudong Cao. "Variational quantum factoring". Quantum Technology and Optimization Problems (2019).

[33] Sumeet Khatri, Ryan LaRose, Alexander Poremba, Lukasz Cincio, Andrew T Sornborger, and Patrick J Coles. "Quantum-assisted quantum compiling". Quantum **3**, 140 (2019).

[34] Kunal Sharma, Sumeet Khatri, M. Cerezo, and Patrick J Coles. "Noise resilience of variational quantum compiling". New Journal of Physics **22**, 043006 (2020).

[35] Tyson Jones and Simon C Benjamin. "Quantum compilation and circuit optimisation via energy dissipation". arXiv preprint arXiv:1811.03147 (2018).

[36] Andrew Arrasmith, Lukasz Cincio, Andrew T Sornborger, Wojciech H Zurek, and Patrick J Coles. "Variational consistent histories as a hybrid algorithm for quantum foundations". Nature Communications **10**, 1–7 (2019).

[37] M. Cerezo, Kunal Sharma, Andrew Arrasmith, and Patrick J Coles. "Variational quantum state eigensolver". arXiv preprint arXiv:2004.01372 (2020).

[38] Ryan LaRose, Arkin Tikku, Étude O'Neel-Judy, Lukasz Cincio, and Patrick J Coles. "Variational quantum state diagonalization". npj Quantum Information **5**, 1–10 (2019).

[39] Guillaume Verdon, Jacob Marks, Sasha Nanda, Stefan Leichenauer, and Jack Hidary. "Quantum Hamiltonian-based models and the variational quantum thermalizer algorithm". arXiv preprint arXiv:1910.02071 (2019).

[40] Peter D Johnson, Jonathan Romero, Jonathan Olson, Yudong Cao, and Alán Aspuru-Guzik. "Qvector: an algorithm for device-tailored quantum error correction". arXiv preprint arXiv:1711.02249 (2017).

[41] John Preskill. "Quantum computing in the NISQ era and beyond". Quantum **2**, 79 (2018).

[42] Kristan Temme, Sergey Bravyi, and Jay M. Gambetta. "Error mitigation for short-depth quantum circuits". Phys. Rev. Lett. **119**, 180509 (2017).

[43] Suguru Endo, Simon C Benjamin, and Ying Li. "Practical quantum error mitigation for near-future applications". Physical Review X **8**, 031027 (2018).

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

[45] Piotr Czarnik, Andrew Arrasmith, Patrick J. Coles, and Lukasz Cincio. "Error mitigation with Clifford quantum-circuit data". Quantum **5**, 592 (2021).

[46] William J Huggins, Sam McArdle, Thomas E O'Brien, Joonho Lee, Nicholas C Rubin, Sergio Boixo, K Birgitta Whaley, Ryan Babbush, and Jarrod R McClean. "Virtual distillation for quantum error mitigation". Physical Review X **11**, 041036 (2021).

[47] Bálint Koczor. "Exponential error suppression for near-term quantum devices". Physical Review X **11**, 031057 (2021).

[48] Jarrod R McClean, Mollie E Kimchi-Schwartz, Jonathan Carter, and Wibe A De Jong. "Hybrid quantum-classical hierarchy for mitigation of decoherence and determination of excited states". Physical Review A **95**, 042308 (2017).

[49] Thomas E. O'Brien, Stefano Polla, Nicholas C. Rubin, William J. Huggins, Sam McArdle, Sergio Boixo, Jarrod R. McClean, and Ryan Babbush. "Error mitigation via verified phase estimation". PRX Quantum **2**, 020317 (2021).

[50] Sam McArdle, Xiao Yuan, and Simon Benjamin. "Error-mitigated digital quantum simulation". Phys. Rev. Lett. **122**, 180501 (2019).

[51] Xavi Bonet-Monroig, Ramiro Sagastizabal, M Singh, and TE O'Brien. "Low-cost error mitigation by symmetry verification". Physical Review A **98**, 062339 (2018).

[52] William J Huggins, Jarrod R McClean, Nicholas C Rubin, Zhang Jiang, Nathan Wiebe, K Birgitta Whaley, and Ryan Babbush. "Efficient and noise resilient measurements for quantum chemistry on near-term quantum computers". npj Quantum Information **7**, 1–9 (2021).

[53] George S Barron and Christopher J Wood. "Measurement error mitigation for variational quantum algorithms". arXiv preprint arXiv:2010.08520 (2020).

[54] Alistair W. R. Smith, Kiran E. Khosla, Chris N. Self, and M. S. Kim. "Qubit readout error mitigation with bit-flip averaging". Science Advances**7** (2021).

[55] Daiqin Su, Robert Israel, Kunal Sharma, Haoyu Qi, Ish Dhand, and Kamil Brádler. "Error mitigation on a near-term quantum photonic device". Quantum **5**, 452 (2021).

[56] Samson Wang, Enrico Fontana, M. Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J Coles. "Noise-induced barren plateaus in variational quantum algorithms". Nature Communications **12**, 1–11 (2021).

[57] Daniel Stilck França and Raul Garcia-Patron. "Limitations of optimization algorithms on noisy quantum devices". Nature Physics **17**, 1221–1227 (2021).

[58] Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. "Barren plateaus in quantum neural network training landscapes". Nature Communications **9**, 1–6 (2018).

[59] M. Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J Coles. "Cost function dependent barren plateaus in shallow parametrized quantum circuits". Nature Communications **12**, 1–12 (2021).

[60] Andrew Arrasmith, M. Cerezo, Piotr Czarnik, Lukasz Cincio, and Patrick J Coles. "Effect of barren plateaus on gradient-free optimization". Quantum **5**, 558 (2021).

[61] M. Cerezo and Patrick J Coles. "Higher order derivatives of quantum neural networks with barren plateaus". Quantum Science and Technology **6**, 035006 (2021).

[62] Kentaro Heya, Yasunari Suzuki, Yasunobu Nakamura, and Keisuke Fujii. "Variational quantum gate optimization". arXiv preprint arXiv:1810.12745 (2018).

[63] Jonathan Romero, Jonathan P Olson, and Alan Aspuru-Guzik. "Quantum autoencoders for efficient compression of quantum data". Quantum Science and Technology **2**, 045001 (2017).

[64] Lennart Bittel and Martin Kliesch. "Training variational quantum algorithms is np-hard". Phys. Rev. Lett. **127**, 120502 (2021).

[65] Jonas M Kübler, Andrew Arrasmith, Lukasz Cincio, and Patrick J Coles. "An adaptive optimizer for measurement-frugal

variational algorithms". Quantum **4**, 263 (2020).

[66] Andrew Arrasmith, Lukasz Cincio, Rolando D Somma, and Patrick J Coles. "Operator sampling for shot-frugal optimization in variational algorithms". arXiv preprint arXiv:2004.06252 (2020).

[67] Andi Gu, Angus Lowe, Pavel A Dub, Patrick J. Coles, and Andrew Arrasmith. "Adaptive shot allocation for fast convergence in variational quantum algorithms". arXiv preprint arXiv:2108.10434 (2021).

[68] Zoë Holmes, Kunal Sharma, M. Cerezo, and Patrick J Coles. "Connecting ansatz expressibility to gradient magnitudes and barren plateaus". PRX Quantum **3**, 010313 (2022).

[69] Zoë Holmes, Andrew Arrasmith, Bin Yan, Patrick J. Coles, Andreas Albrecht, and Andrew T Sornborger. "Barren plateaus preclude learning scramblers". Physical Review Letters **126**, 190501 (2021).

[70] Carlos Ortiz Marrero, Mária Kieferová, and Nathan Wiebe. "Entanglement-induced barren plateaus". PRX Quantum **2**, 040316 (2021).

[71] Taylor L Patti, Khadijeh Najafi, Xun Gao, and Susanne F Yelin. "Entanglement devised barren plateau mitigation". Physical Review Research **3**, 033090 (2021).

[72] Martin Larocca, Piotr Czarnik, Kunal Sharma, Gopikrishnan Muraleedharan, Patrick J. Coles, and M. Cerezo. "Diagnosing barren plateaus with tools from quantum optimal control". arXiv preprint arXiv:2105.14377 (2021).

[73] Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. "Quantum circuit learning". Physical Review A **98**, 032309 (2018).

[74] Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. "Evaluating analytic gradients on quantum hardware". Physical Review A **99**, 032331 (2019).

[75] John A Nelder and Roger Mead. "A simplex method for function minimization". The computer journal **7**, 308–313 (1965).

[76] M. J. D. Powell. "A direct search optimization method that models the objective and constraint functions by linear interpolation". Advances in Optimization and Numerical Analysis (1994).

[77] E. Campos, D. Rabinovich, V. Akshay, and J. Biamonte. "Training saturation in layerwise quantum approximate optimization". Physical Review A**104** (2021).

[78] Cheng Xue, Zhao-Yun Chen, Yu-Chun Wu, and Guo-Ping Guo. "Effects of quantum noise on quantum approximate optimization algorithm". Chinese Physics Letters **38**, 030302 (2021).

[79] Jeffrey Marshall, Filip Wudarski, Stuart Hadfield, and Tad Hogg. "Characterizing local noise in qaoa circuits". IOP SciNotes **1**, 025208 (2020). url: `https://iopscience.iop.org/article/10.1088/2633-1357/abb0d7`.

[80] Enrico Fontana, M. Cerezo, Andrew Arrasmith, Ivan Rungger, and Patrick J. Coles. "Non-trivial symmetries in quantum landscapes and their resilience to quantum noise". Quantum **6**, 804 (2022).

[81] Suguru Endo, Zhenyu Cai, Simon C Benjamin, and Xiao Yuan. "Hybrid quantum-classical algorithms and quantum error mitigation". Journal of the Physical Society of Japan **90**, 032001 (2021).

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

[83] Andrea Mari, Nathan Shammah, and William J Zeng. "Extending quantum probabilistic error cancellation by noise scaling". Physical Review A **104**, 052607 (2021).

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

[85] Ashley Montanaro and Stasja Stanisic. "Error mitigation by training with

fermionic linear optics". arXiv preprint arXiv:2102.02120 (2021).

[86] Joseph Vovrosh, Kiran E Khosla, Sean Greenaway, Christopher Self, Myungshik S Kim, and Johannes Knolle. "Simple mitigation of global depolarizing errors in quantum simulations". Physical Review E **104**, 035309 (2021).

[87] Eliott Rosenberg, Paul Ginsparg, and Peter L McMahon. "Experimental error mitigation using linear rescaling for variational quantum eigensolving with up to 20 qubits". Quantum Science and Technology **7**, 015024 (2022).

[88] Andre He, Benjamin Nachman, Wibe A. de Jong, and Christian W. Bauer. "Zero-noise extrapolation for quantum-gate error mitigation with identity insertions". Physical Review A **102**, 012426 (2020).

[89] Andrew Shaw. "Classical-quantum noise mitigation for nisq hardware". arXiv preprint arXiv:2105.08701 (2021).

[90] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Andreas Bengtsson, Sergio Boixo, Michael Broughton, Bob B Buckley, et al. "Observation of separated dynamics of charge and spin in the fermi-hubbard model". arXiv preprint arXiv:2010.07965 (2020).

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

[92] Piotr Czarnik, Andrew Arrasmith, Lukasz Cincio, and Patrick J Coles. "Qubit-efficient exponential suppression of errors". arXiv preprint arXiv:2102.06056 (2021).

[93] Yifeng Xiong, Daryus Chandra, Soon Xin Ng, and Lajos Hanzo. "Sampling overhead analysis of quantum error mitigation: Uncoded vs. coded systems". IEEE Access **8**, 228967–228991 (2020).

[94] Ryuji Takagi. "Optimal resource cost for error mitigation". Phys. Rev. Res. **3**, 033178 (2021).

[95] Lukasz Cincio, Kenneth Rudinger, Mohan Sarovar, and Patrick J. Coles. "Machine learning of noise-resilient quantum circuits". PRX Quantum **2**, 010324 (2021).

[96] P Erdös and A Rényi. "On random graphs I". Publicationes Mathematicae Debrecen **6**, 18 (1959). url: `http://snap.stanford.edu/class/cs224w-readings/erdos59random.pdf`.

[97] Andrew Wack, Hanhee Paik, Ali Javadi-Abhari, Petar Jurcevic, Ismael Faro, Jay M. Gambetta, and Blake R. Johnson. "Quality, speed, and scale: three key attributes to measure the performance of near-term quantum computers". arXiv preprint arXiv:2110.14108 (2021).

[98] Tudor Giurgica-Tiron, Yousef Hindy, Ryan LaRose, Andrea Mari, and William J Zeng. "Digital zero noise extrapolation for quantum error mitigation". 2020 IEEE International Conference on Quantum Computing and Engineering (QCE) (2020).

[99] Youngseok Kim, Christopher J. Wood, Theodore J. Yoder, Seth T. Merkel, Jay M. Gambetta, Kristan Temme, and Abhinav Kandala. "Scalable error mitigation for noisy quantum circuits produces competitive expectation values". arXiv preprint arXiv:2108.09197 (2021).

[100] Cristina Cirstoiu, Silas Dilkes, Daniel Mills, Seyon Sivarajah, and Ross Duncan. "Volumetric benchmarking of error mitigation with Qermit". arXiv preprint arXiv:2204.09725 (2022).

[101] Ryuji Takagi, Suguru Endo, Shintaro Minagawa, and Mile Gu. "Fundamental limits of quantum error mitigation". npj Quantum Information **8**, 114 (2022).

[102] Avram Sidi. "Practical extrapolation methods: Theory and applications". Volume 10. Cambridge University Press. (2003).

[103] Masanori Ohya and Dénes Petz. "Quantum entropy and its use". Springer Science & Business Media. (2004).

[104] Christoph Hirche, Cambyse Rouzé, and Daniel Stilck França. "On contraction coefficients, partial orders and approximation of capacities for quantum channels". Quantum **6**, 862 (2022).

[105] Jeffrey C. Lagarias, James A. Reeds, Margaret H. Wright, and Paul E. Wright. “Convergence properties of the nelder–mead simplex method in low dimensions”. SIAM Journal on Optimization **9**, 112–147 (1998).

[106] Abhijith J., Adetokunbo Adedoyin, John Ambrosiano, Petr Anisimov, William Casper, Gopinath Chennupati, Carleton Coffrin, Hristo Djidjev, David Gunter, Satish Karra, Nathan Lemons, Shizeng Lin, Alexander Malyzhenkov, David Mascarenas, Susan Mniszewski, Balu Nadiga, Daniel O’malley, Diane Oyen, Scott Pakin, Lakshman Prasad, Randy Roberts, Phillip Romero, Nandakishore Santhi, Nikolai Sinitsyn, Pieter J. Swart, James G. Wendelberger, Boram Yoon, Richard Zamora, Wei Zhu, Stephan Eidenbenz, Andreas Bärtschi, Patrick J. Coles, Marc Vuffray, and Andrey Y. Lokhov. “Quantum algorithm implementations for beginners”. ACM Transactions on Quantum Computing (2022).

[107] Bálint Koczor. “The dominant eigenvector of a noisy quantum state”. New Journal of Physics **23**, 123047 (2021).

A Road map of appendices

In Appendix B we present some notation and definitions that we need in order to prove our main results, as well as provide further details on the error mitigation protocols studied in this article. In Appendix C we derive some useful lemmas that are required for our proofs. In Appendix D we present the proof for our asymptotic results on the exponential concentration of estimators. In Appendix E we present our protocol-specific results on the change in resolvability of the cost landscape under error mitigation. Finally, in Appendix F we discuss details of our numerical implementations for Clifford Data Regression, Virtual Distillation, and Zero-Noise Extrapolation.

B Preliminaries

B.1 Further details on error mitigation techniques

In this section we expand on our discussion in Section 2.3 and provide further details on the Zero-Noise Extrapolation and Probabilistic Error Cancellation protocols.

B.1.1 Zero-Noise Extrapolation (ZNE)

For convenience in this section we recall the key points of Zero-Noise Extrapolation as summarized in Section (3.2.2). We also detail the explicit forms of the estimators that can be constructed for exponential extrapolation and an extrapolation strategy tailored towards NIBP effects, which will be required in order to prove our results.

Richardson Extrapolation. We suppose that $\widetilde{C}(\boldsymbol{\theta}_i,\varepsilon)$ admits a Taylor expansion in small noise parameter ε as

$$\widetilde{C}(\boldsymbol{\theta}_i,\varepsilon)=\widetilde{C}(\boldsymbol{\theta}_i,0)+\sum_{k=1}^{m}p_k(\boldsymbol{\theta}_i)\varepsilon^k+\mathcal{O}(\varepsilon^{m+1})\,, \tag{48}$$

where p_k are unknown parameters and $\widetilde{C}(\boldsymbol{\theta}_i,0)$ is the zero-noise cost function. By considering the equivalent expansion of $\widetilde{C}(\boldsymbol{\theta}_i,a_1\varepsilon)$ and combining the two equations we can obtain

$$C_m^{(2)}(\boldsymbol{\theta}_i)=\frac{a_1\widetilde{C}(\boldsymbol{\theta}_i,\varepsilon)-\widetilde{C}(\boldsymbol{\theta}_i,a_1\varepsilon)}{a_1-1}=\widetilde{C}(\boldsymbol{\theta}_i,0)+\mathcal{O}(\varepsilon^2)\,, \tag{49}$$

which is a higher-order approximation of $\widetilde{C}(\boldsymbol{\theta}_i,0)$ compared to simply using $\widetilde{C}(\boldsymbol{\theta}_i,\varepsilon)$. This process can be repeated iteratively m times to obtain an estimator which is accurate up to $\mathcal{O}(\varepsilon^{m+1})$ error. It can be shown that the general form for the estimator that uses k noise levels can be written as

$$C_m^{(k)}(\boldsymbol{\theta}_i) = \sum_{j=0}^{k} \beta_j \widetilde{C}(\boldsymbol{\theta}_i, a_j \varepsilon)\,, \tag{50}$$

where the coefficients β_j satisfy the linear system of equations $\sum_{j=0}^{k} \beta_j = 1$ and $\sum_{l=0}^{k} \beta_l a_l^t = 0$ for all $t \in \{1, ..., k\}$ [102]. For 3 noise levels, (50) explicitly gives

$$C_m^{(3)}(\boldsymbol{\theta}_i) = \frac{a_1 a_2 (a_2 - a_1)\widetilde{C}(\boldsymbol{\theta}_i, \varepsilon) - a_2(a_2 - 1)\widetilde{C}(\boldsymbol{\theta}_i, a_1\varepsilon) + a_1(a_1 - 1)\widetilde{C}(\boldsymbol{\theta}_i, a_2\varepsilon)}{(a_1 - 1)(a_2 - 1)(a_2 - a_1)} = \widetilde{C}(\boldsymbol{\theta}_i, 0) + \mathcal{O}(\varepsilon^3)\,. \tag{51}$$

Exponential extrapolation. We can also consider an exponential model

$$\widetilde{C}(\boldsymbol{\theta}_i, \varepsilon) = r(\boldsymbol{\theta}_i, \varepsilon)^{-t(\boldsymbol{\theta}_i, \varepsilon)} \left(\sum_{k=0}^{m} p_k(\boldsymbol{\theta}_i)\varepsilon^k + \mathcal{O}(\varepsilon^{m+1}) \right)\,, \tag{52}$$

for some r and t, which in general can be functions of ε. Following in similar steps to Richardson extrapolation we can consider the same expansion at an augmented noise level $a_1\varepsilon$. This enables us to construct the estimator

$$\begin{aligned} C_m(\boldsymbol{\theta}_i) = \frac{1}{a_1 - 1}\Big(& a_1 r(\boldsymbol{\theta}_i, \varepsilon)^{t(\boldsymbol{\theta}_i, \varepsilon)} \widetilde{C}(\boldsymbol{\theta}_i, \varepsilon) \\ & - r(\boldsymbol{\theta}_i, a_1\varepsilon)^{t(\boldsymbol{\theta}_i, a_1\varepsilon)} \widetilde{C}(\boldsymbol{\theta}_i, a_1\varepsilon) \Big)\,, \end{aligned} \tag{53}$$

which approximates $\widetilde{C}(\boldsymbol{\theta}_i, 0)$ to a higher order in ε compared to $\widetilde{C}(\boldsymbol{\theta}_i, \varepsilon)$.

NIBP extrapolation. We can construct an alternative Zero-Noise Extrapolation strategy that is tailored towards noisy cost function values that are dominated by NIBP scaling as in Eq. (6). We model the effects of noise as

$$\widetilde{C}(\boldsymbol{\theta}_i, q) = A + q^L \left(B(\boldsymbol{\theta}_i) + \sum_{k=1}^{m} p_k (1 - q)^k + \mathcal{O}\left((1 - q)^{m+1} \right) \right)\,, \tag{54}$$

for all noisy cost function points $\widetilde{C}(\boldsymbol{\theta}_i, q)$, where $A = \widetilde{C}(\boldsymbol{\theta}_i, q = 0)$ is the fixed point of the noise (corresponding to the maximally mixed state) and $A + B(\boldsymbol{\theta}_i)$ is the noise-free cost value. (Note that for NIBPs we cannot consider lower-order polynomials of q, as else the NIBP condition would be broken for small q.) We construct estimators for any given parameter $\boldsymbol{\theta}_i$, as

$$C_m(\boldsymbol{\theta}_i) = \frac{a^{L+1} q^{-L} (\widetilde{C}(\boldsymbol{\theta}_i, q/a) - \widetilde{C}(\boldsymbol{\theta}_i, q = 0)) - q^{-L}(\widetilde{C}(\boldsymbol{\theta}_i, q) - \widetilde{C}(\boldsymbol{\theta}_i, q = 0))}{a - 1} + K\,, \tag{55}$$

where $K = A - \sum_k p_k$ is an additive constant. As we are only interested in cost function differences for our results, this will cancel out.

B.1.2 Probabilistic Error Cancellation

General idea. Probabilisitic Error Cancellation utilizes many modified circuit runs in order to construct a quasiprobability representation of the noise-free cost function [42, 43]. We assume that the effect of the noise can be described by a quantum channel $\mathcal{N}$ that occurs after a gate that we denote with unitary channel $\mathcal{U}$. For now we assume this is the only instance of noise in the circuit, however, we will later generalize to many instances of noise. The goal of Probabilistic Error Cancellation is to simulate the inverse map $\mathcal{N}^{-1}$. Note in general this will not always correspond to a CPTP map. Despite this, if we have a basis of (noisy) quantum channels $\{\mathcal{B}_\alpha\}_\alpha$, corresponding to experimentally available channels, we can expand the inverse map in this basis as

$$\mathcal{N}^{-1} = \sum_\alpha q_\alpha \mathcal{B}_\alpha\,, \tag{56}$$

for some set of $q_\alpha \in \mathbb{R}$. Then, the channel that describes the noiseless gate can be written as

$$\mathcal{U} = \mathcal{N}^{-1} \circ \mathcal{N} \circ \mathcal{U} \tag{57}$$

$$= \sum_\alpha q_\alpha \mathcal{K}_\alpha \,, \tag{58}$$

where we have defined $\mathcal{K}_\alpha = \mathcal{B}_\alpha \circ \mathcal{N} \circ \mathcal{U}$. Denote the input state to the gate as ρ_{in} and a measurement operator as O. The expectation value can be written

$$C_{\mathcal{U}(\rho)} = \mathrm{Tr}\left[\mathcal{U}(\rho_{in})O\right] = \sum_\alpha q_\alpha \widetilde{C}_{\mathcal{K}_\alpha(\rho)} \,, \tag{59}$$

where for simplicity we first assume that $\mathcal{U}$ is the only gate in the circuit, and $\widetilde{C}_{\mathcal{K}_\alpha(\rho)} \equiv \mathrm{Tr}\left[\mathcal{K}_\alpha(\rho)O\right]$. Finally, we can explicitly define a probability distribution $p_\alpha = |q_\alpha|/G_{\mathcal{N}}$ where $G_{\mathcal{N}} = \sum_\alpha |q_\alpha|$. This gives us an alternative way to write (58) and (59) as

$$\mathcal{U} = G_{\mathcal{N}} \sum_\alpha \mathrm{sgn}(q_\alpha)\, p_\alpha \, \mathcal{K}_\alpha \,, \tag{60}$$

$$C_{\mathcal{U}(\rho)} = G_{\mathcal{N}} \sum_\alpha \mathrm{sgn}(q_\alpha)\, p_\alpha \, \widetilde{C}_{\mathcal{K}_\alpha(\rho)} \,, \tag{61}$$

where $\mathrm{sgn}(q_\alpha)$ denotes the sign of q_α. We call this the quasiprobability representation of the gate $\mathcal{U}$. The idea is that if we have access to the set of CPTP maps $\{\mathcal{B}_\alpha\}_\alpha$ in our noisy native hardware gate set, then we can obtain an estimate of the noiseless expectation value $C_{\mathcal{U}(\rho)}$ as follows: (1) With probability p_α, prepare the circuit corresponding to $\mathcal{K}_\alpha(\rho)$ and obtain $\widetilde{C}_{\mathcal{K}_\alpha(\rho)}$. (2) Multiply the measurement result by $\mathrm{sgn}(q_\alpha)G_{\mathcal{N}}$. (3) Repeat process many times and sum results.

Correcting many gates. So far we have only considered a circuit with a single gate $\mathcal{U}$. We can generalize (60) to a general circuit $\prod_k^{N_g} \mathcal{U}_k$ with N_g gates with the quasiprobability representation

$$\prod_k \mathcal{U}_k = G_{\mathcal{N}}^{tot} \sum_{\boldsymbol{i}} \mathrm{sgn}(q_{\boldsymbol{i}}) p_{\boldsymbol{i}} \mathcal{K}_{\boldsymbol{i}} \,, \tag{62}$$

where $G_{\mathcal{N}}^{tot} = \prod_k G_k$, $\boldsymbol{i} = (i_1, ..., i_{N_g})$, $q_{\boldsymbol{i}} = \prod_k q_{i_k}$, $p_{\boldsymbol{i}} = \prod_k p_{i_k}$, $\mathcal{K}_{\boldsymbol{i}} = \prod_k \mathcal{K}_{i_k}$. Thus, a similar procedure can be carried out in order to mitigate the noise on each individual gate in the circuit.

C Useful Lemmas

C.1 Noise Induced Cost Concentration

Lemma 1. *Consider a parameterized noisy cost function $\widetilde{C}(\boldsymbol{\theta}) = \mathrm{Tr}\left[\widetilde{\rho}(\boldsymbol{\theta})O\right]$, where $\widetilde{\rho}(\boldsymbol{\theta})$ is an n-qubit noisy state given by Eq. (3) and $\boldsymbol{\theta} \in \Theta$ is drawn from some set of accessible parameters Θ. Suppose the cost is suffering from exponential cost concentration according to Ref. [56], that is*

$$\left|\widetilde{C}(\boldsymbol{\theta}) - \frac{\mathrm{Tr}[O]}{2^n}\right| \leq q^L B(n) \left\|\rho_{in} - \frac{\mathbb{1}}{2^n}\right\|_1 \,, \tag{63}$$

for all $\boldsymbol{\theta} \in \Theta$, where ρ_{in} is the input state in Eq. (3), $0 \leq q < 1$ is some noise parameter, $B(n) \in \mathrm{poly}(n)$, L is the number of layers of gates, and $\|\cdot\|_1$ denotes the Schatten 1-norm (trace norm). Then, $\exists A(n) \in \mathrm{poly}(n)$ such that

$$\left|\widetilde{C}(\boldsymbol{\theta}_1, q) - \widetilde{C}(\boldsymbol{\theta}_2, q)\right| \leq A(n) q^L \,, \tag{64}$$

for any two sets of parameters $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$.

Proof. Starting from Eq. (63) can simply write

$$\left|\widetilde{C}(\boldsymbol{\theta}_1, q) - \widetilde{C}(\boldsymbol{\theta}_2, q)\right| \leq 4B(n)q^L\,, \tag{65}$$

for any two sets of parameters $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, where we have used the triangle inequality in $1D$ and the fact that the trace distance has a maximum value of 1. □

For the next lemma we will consider the n-qubit channel

$$\mathcal{W} = \mathcal{U}_k \circ \mathcal{N} \circ \cdots \circ \mathcal{N} \circ \mathcal{U}_2 \circ \mathcal{N} \circ \mathcal{U}_1 \circ \mathcal{N}\,, \tag{66}$$

where $\{\mathcal{U}_k\}_{k=1}^L$ denote unitary channels that describe collections of gates that act together in a layer, and $\mathcal{N} = \bigotimes_{i=1}^n \mathcal{N}_i$ is an instance of local Pauli channels, such that action of $\mathcal{N}_j$ on a local Pauli operator $\sigma \in \{X, Y, Z\}$ can be expressed as

$$\mathcal{N}_j(\sigma) = q_\sigma^{(j)}\sigma\,, \tag{67}$$

where we assume $-1 < q_X^{(j)}, q_Y^{(j)}, q_Z^{(j)} < 1$ for all qubit labels j. We characterize the noise strength with a single parameter $q = \max_{j,\sigma}\{|q_\sigma^{(j)}|\} < 1$.

Lemma 2. *Consider $\mathcal{W}$ as defined in Eq. (66) acting on some input state ρ_{in}. Then we have*

$$\left\|\mathcal{W}(\rho_{in}) - \frac{\mathbb{1}}{2^n}\right\|_1 \leq q^{ck} n^{1/2}\sqrt{2\ln 2}\,, \tag{68}$$

where $c = 1/(2\ln 2)$ is a constant.

Proof. We have

$$\begin{aligned}\left\|\mathcal{W}(\rho_{in}) - \frac{\mathbb{1}}{2^n}\right\|_1 &\leq \sqrt{2\ln 2\, D\Big(\mathcal{W}(\rho_{in})\Big\|\frac{\mathbb{1}}{2^n}\Big)} &(69)\\ &\leq \sqrt{2\ln 2\, q^{2ck} D\Big(\rho_{in}\Big\|\frac{\mathbb{1}}{2^n}\Big)} &(70)\\ &\leq q^{ck} n^{1/2}\sqrt{2\ln 2}\,, &(71)\end{aligned}$$

where $D(\cdot\|\cdot)$ denotes the relative entropy. The first inequality is Pinsker's [103], and the second inequality comes from a direct application of Supplementary Lemma 6 in Ref. [56] (adapted from Corollary 5.6 of Ref. [104]). □

C.2 Averages over unitary 2-designs

Lemma 3. *Consider the cost function value $C_\sigma(U) = \mathrm{Tr}[U\sigma U^\dagger O]$ where U is a $d \times d$ unitary matrix and $\sigma \in S(\mathcal{H})$ is some quantum state. Consider expectation values over $U_i \in Y$ where $Y \subset \mathcal{U}(d)$ is a unitary 2-design and $\mathcal{U}(d)$ is the unitary group of degree d. Denote such expectation values as $\langle\cdot\rangle_U$. Then, we have*

$$\begin{aligned}\langle C_\sigma\rangle_U &= \frac{1}{d}\mathrm{Tr}[O]\,, &(72)\\ \langle C_\rho C_\sigma\rangle_U &= \frac{\mathrm{Tr}[O^2]\,(d\mathrm{Tr}[\rho\sigma] - 1) - \mathrm{Tr}[O]^2\,(\mathrm{Tr}[\rho\sigma] - d)}{d(d^2-1)}\,, &(73)\end{aligned}$$

for any two operators $\sigma, \rho \in B(\mathcal{H})$ which satisfy $\mathrm{Tr}[\rho] = \mathrm{Tr}[\sigma] = 1$, $\dim(\mathcal{H}) = d$. This implies

$$\mathrm{Var}[C_\sigma] = \frac{\left(\mathrm{Tr}[O^2] - \frac{1}{d}\mathrm{Tr}[O]^2\right)\left(\mathrm{Tr}[\sigma^2] - \frac{1}{d}\right)}{d^2-1}\,. \tag{74}$$

Proof. We use the following standard expressions for integration with respect to the Haar measure over the unitary group of degree d:

$$\int_{\mathcal{U}(d)} d\mu(W) w_{i,j} w^*_{p,k} = \frac{\delta_{i,p}\delta_{j,k}}{d}\,, \tag{75}$$

$$\begin{aligned}\int_{\mathcal{U}(d)} d\mu(W) w_{i_1,j_1} w_{i_2,j_2} w^*_{i'_1,j'_1} w^*_{i'_2,j'_2} = &\frac{1}{d^2-1}\left(\delta_{i_1,i'_1}\delta_{i_2,i'_2}\delta_{j_1,j'_1}\delta_{j_2,j'_2} + \delta_{i_1,i'_2}\delta_{i_2,i'_1}\delta_{j_1,j'_2}\delta_{j_2,j'_1}\right) \\ &- \frac{1}{d\,(d^2-1)}\left(\delta_{i_1,i'_1}\delta_{i_2,i'_2}\delta_{j_1,j'_2}\delta_{j_2,j'_1} + \delta_{i_1,i'_2}\delta_{i_2,i'_1}\delta_{j_1,j'_1}\delta_{j_2,j'_2}\right)\,,\end{aligned} \tag{76}$$

where $w_{i,j}$ are the matrix elements of the unitary $W \in \mathcal{U}(d)$. Then, the expectation values over 2-designs can be evaluated as

$$\langle C_\sigma\rangle_U = \frac{1}{d}\mathrm{Tr}[O]\,, \tag{77}$$

and

$$\begin{aligned}\langle C_\rho C_\sigma\rangle_U = &\frac{1}{d^2-1}\mathrm{Tr}[O]^2 + \frac{1}{d^2-1}\mathrm{Tr}[\rho\sigma]\mathrm{Tr}[O^2] \\ &- \frac{1}{d(d^2-1)}\mathrm{Tr}[\rho\sigma]\mathrm{Tr}[O]^2 - \frac{1}{d(d^2-1)}\mathrm{Tr}[O^2]\,,\end{aligned} \tag{78}$$

where we have used $\mathrm{Tr}[\rho] = \mathrm{Tr}[\sigma] = 1$. The final statement comes by noting that

$$\mathrm{Var}[C_\sigma] = \langle C_\sigma^2\rangle_U - \langle C_\sigma\rangle_U^2 \tag{79}$$

$$= \frac{\mathrm{Tr}[O^2]\left(d\mathrm{Tr}[\sigma^2]-1\right) - \mathrm{Tr}[O]^2\left(\mathrm{Tr}[\sigma^2]-d\right)}{d(d^2-1)} - \frac{1}{d^2}\mathrm{Tr}[O]^2 \tag{80}$$

$$= \frac{\mathrm{Tr}[O^2]\left(d\mathrm{Tr}[\sigma^2]-1\right)}{d(d^2-1)} - \frac{\mathrm{Tr}[O]^2\left(d\mathrm{Tr}[\sigma^2]-1\right)}{d^2(d^2-1)}\,, \tag{81}$$

which can be factorized to give the desired result. □

D Exponential estimator concentration

We present a proof of Theorem 1, and restate the result here for convenience.

Theorem 1. *Consider an error mitigation protocol that prepares the quantity*

$$E_{\sigma,X,M,k} = \mathrm{Tr}\left[X\left(\sigma^{\otimes M}\otimes|0\rangle\langle 0|^{\otimes k}\right)\right]\,, \tag{82}$$

for some quantum state $\sigma \in S(\mathcal{H})$, for $|0\rangle\langle 0| \in S(\mathcal{H}')$ and for some $X \in B(\mathcal{H}^{\otimes M}\otimes\mathcal{H}'^{\otimes k})$. We suppose σ is prepared with a depth L_σ circuit and experiences noise according to Eq. (3). Under these conditions, $E_{\sigma,X,M,k}$ exponentially concentrates on a state-independent fixed point in the depth of the circuit as

$$\left|E_{\sigma,X,M,k} - \mathrm{Tr}\left[X\left(\frac{\mathbb{1}^{\otimes M}}{2^{Mn}}\otimes|0\rangle\langle 0|^{\otimes k}\right)\right]\right| \leq G_{\sigma,X,M}(n)\,, \tag{83}$$

where $\mathbb{1} \in S(\mathcal{H})$ is the n-qubit identity operator and

$$G_{\sigma,X,M}(n) = \sqrt{\ln 4}\left\|X\right\|_\infty M n^{1/2} q^{c(L_\sigma+1)}\,, \tag{84}$$

with noise parameter $q \in [0,1)$ and constant $c = 1/(2\ln 2)$.

Proof. Consider

$$\left|\mathrm{Tr}\Big[\left(\sigma^{\otimes M}\otimes|0\rangle\langle 0|^{\otimes k}\right)X\Big]-\mathrm{Tr}\left[\left(\frac{\mathbb{1}}{2^n}\otimes\sigma^{\otimes M-1}\otimes|0\rangle\langle 0|^{\otimes k}\right)X\right]\right|=\left|\mathrm{Tr}\left[\left((\sigma-\frac{\mathbb{1}}{2^n})\otimes\sigma^{\otimes M-1}\otimes|0\rangle\langle 0|^{\otimes k}\right)X\right]\right| \tag{85}$$

$$\leq\left\|(\sigma-\frac{\mathbb{1}}{2^n})\otimes\sigma^{\otimes M-1}\otimes|0\rangle\langle 0|^{\otimes k}\right\|_1\|X\|_\infty \tag{86}$$

$$=\left\|\sigma-\frac{\mathbb{1}}{2^n}\right\|_1\mathrm{Tr}[\sigma]^{M-1}\mathrm{Tr}[|0\rangle\langle 0|]^k\|X\|_\infty \tag{87}$$

$$\leq q^{c(L_\sigma+1)}n^{1/2}\sqrt{2\ln 2}\,\|X\|_\infty\,. \tag{88}$$

The first inequality is due to the matrix Hölder's inequality, and the second inequality follows from Lemma 2. Similarly, we have $M-1$ further such equations, which we display with the original equation:

$$\mathrm{Tr}\Big[\left(\sigma^{\otimes M}\otimes|0\rangle\langle 0|^{\otimes k}\right)X\Big]-\mathrm{Tr}\left[\left(\frac{\mathbb{1}}{2^n}\otimes\sigma^{\otimes M-1}\otimes|0\rangle\langle 0|^{\otimes k}\right)X\right]\leq q^{c(L_\sigma+1)}n^{1/2}\sqrt{2\ln 2}\,\|X\|_\infty\,, \tag{89}$$

$$\mathrm{Tr}\left[\left(\frac{\mathbb{1}}{2^n}\otimes\sigma^{\otimes M-1}\otimes|0\rangle\langle 0|^{\otimes k}\right)X\right]-\mathrm{Tr}\left[\left(\frac{\mathbb{1}}{2^n}\otimes\frac{\mathbb{1}}{2^n}\otimes\sigma^{\otimes M-2}\otimes|0\rangle\langle 0|^{\otimes k}\right)X\right]\leq q^{c(L_\sigma+1)}n^{1/2}\sqrt{2\ln 2}\,\|X\|_\infty\,, \tag{90}$$

$$\ldots$$

$$\mathrm{Tr}\left[\left(\left(\frac{\mathbb{1}}{2^n}\right)^{\otimes M-1}\otimes\sigma\otimes|0\rangle\langle 0|^{\otimes k}\right)X\right]-\mathrm{Tr}\left[\left(\left(\frac{\mathbb{1}}{2^n}\right)^{\otimes M}\otimes|0\rangle\langle 0|^{\otimes k}\right)X\right]\leq q^{c(L_\sigma+1)}n^{1/2}\sqrt{2\ln 2}\,\|X\|_\infty\,. \tag{91}$$

The summation of these equations gives

$$\mathrm{Tr}\Big[X\left(\sigma^{\otimes M}\otimes|0\rangle\langle 0|^{\otimes k}\right)\Big]-\mathrm{Tr}\left[X\left(\frac{\mathbb{1}^{\otimes M}}{2^{Mn}}\otimes|0\rangle\langle 0|^{\otimes k}\right)\right]\leq Mq^{c(L_\sigma+1)}n^{1/2}\sqrt{2\ln 2}\,\|X\|_\infty\,, \tag{92}$$

which gives the desired bound. □

We now present a more detailed version of Corollary 1 in the main text, which explains how one can spend an exponential number of resources in different ways in order to resolve concentrated cost values.

Corollary 1 (Exponential estimator concentration)**.** *Consider an error mitigation protocol that approximates the noise-free cost value $C(\boldsymbol{\theta})$ by estimating the quantity*

$$C_m(\boldsymbol{\theta})=\sum_{(\sigma(\boldsymbol{\theta}),X,M,k)\in T}a_{X,M,k}E_{\sigma(\boldsymbol{\theta}),X,M,k}\,, \tag{93}$$

where each $E_{\sigma,X,M,k}$ takes the form (82). We denote M_{max} and a_{max} as the maximum values of M and $a_{X,M,k}$ respectively accessible from a set T defined by the given protocol. Assuming $\|X\|_\infty\in\mathcal{O}(\mathrm{poly}(n))$, there exists a fixed point F independent of $\boldsymbol{\theta}$ such that

$$|C_m(\boldsymbol{\theta})-F|\in\mathcal{O}(2^{-\beta n}a_{max}|T|M_{max})\,, \tag{94}$$

for some constant $\beta\geq 1$ if the circuit depths satisfy

$$L_{\sigma(\boldsymbol{\theta})}\in\Omega(n)\,, \tag{95}$$

for all $\sigma(\boldsymbol{\theta})$ in the construction (93). That is, if the depth of the circuits scale linearly or greater then one requires at least exponential resources to distinguish C_m from its fixed point, for instance in one of the following ways:

- $a_{max}|T|M_{max} \in \mathcal{O}(\text{poly}(n))$ *and an exponentially large number of shots are used to distinguish two quantities with exponentially small separation*

- $a_{max}|T| \in \Omega(2^{\beta' n})$ *for some constant* $\beta' \geq 1$ *and an exponentially large number of shots are required to distinguish two quantities with exponentially large statistical uncertainty, as measurement outcomes are multiplied by* $a_{max}|T|$.

- $M_{max} \in \Omega(2^{\beta'' n})$ *for some constant* $\beta'' \geq 1$ *and an exponentially large number of copies of some quantum state* σ *are required.*

Proof. Explicitly applying the results of Theorem (1) to the construction of $C_m(\boldsymbol{\theta})$ in (93) we have

$$\left| C_m(\boldsymbol{\theta}) - \sum_{(\sigma(\boldsymbol{\theta}),X,M,k)\in T} a_{X,M,k} \operatorname{Tr}\left[X \left(\frac{\mathbb{1}^{\otimes M}}{2^{Mn}} \otimes |0\rangle\langle 0|^{\otimes k} \right) \right] \right| \leq \sum_{(\sigma(\boldsymbol{\theta}),X,M,k)\in T} a_{X,M,k}\, G_{\sigma(\boldsymbol{\theta}),X,M}(n) \tag{96}$$

$$\in \mathcal{O}\left(\sum_{(\sigma(\boldsymbol{\theta}),X,M,k)\in T} a_{X,M,k} \|X\|_\infty M n^{1/2} q^{c(L_{\sigma(\boldsymbol{\theta})}+1)} \right), \tag{97}$$

where in the second line we have used (84). If $L_{\sigma(\boldsymbol{\theta})} \in \Omega(n)$ then $q^{c(L_{\sigma(\boldsymbol{\theta})}+1)} \in \mathcal{O}(2^{-\beta(\boldsymbol{\theta})n})$ for some $\beta(\boldsymbol{\theta}) \geq 1$. Thus, we can write

$$\left| C_m(\boldsymbol{\theta}) - \sum_{(\sigma(\boldsymbol{\theta}),X,M,k)\in T} a_{X,M,k} \operatorname{Tr}\left[X \left(\frac{\mathbb{1}^{\otimes M}}{2^{Mn}} \otimes |0\rangle\langle 0|^{\otimes k} \right) \right] \right| \in \mathcal{O}(2^{-\beta n} a_{max}|T|M_{max}), \tag{98}$$

as required, where we can denote $\beta = \min_{\boldsymbol{\theta}} \beta(\boldsymbol{\theta})$ and the fixed point as F, noting that F is indeed parameter independent. From here, we can inspect the three presented cases:

- If $a_{max}|T|M_{max} \in \mathcal{O}(\text{poly}(n))$ then C_m has exponentially small separation from F.

- There exists choice $\beta' \geq 1$ such that if $a_{max}|T| \in \Omega(2^{\beta' n})$ such that C_m is not exponentially concentrated on F, however, C_m now has an exponentially large statistical uncertainty, as measurement outcomes are multiplied by coefficients of order $a_{max}|T|$.

- There exists choice of $\beta'' \geq 1$ such that $M_{max} \in \Omega(2^{\beta'' n})$ and C_m is not exponentially concentrated on F.

□

E Protocol-specific results

E.1 Zero-Noise Extrapolation

In this section we present our results for Zero-Noise Extrapolation. As discussed in Section 3.2.2 of the main text, we will consider a Richardson extrapolation strategy based on Eq. (48), an exponential extrapolation strategy based on Eq. (52) and a NIBP extrapolation strategy based on Eq. (55). As we deal with two types of noise parameters, throughout this section we will adopt the unifying notation

$$\widetilde{C}(\boldsymbol{\theta}, a) = \begin{cases} \widetilde{C}(\boldsymbol{\theta}, a\varepsilon) & \text{for Richardson/exponential extrapolation} \\ \widetilde{C}(\boldsymbol{\theta}, q/a) & \text{for NIBP extrapolation}, \end{cases} \tag{99}$$

for all $a \geq 1$. Thus, $\widetilde{C}(\boldsymbol{\theta}, a)$ denotes the noisy cost value at the boosted noise level, and $\widetilde{C}(\boldsymbol{\theta}, 1)$ denotes the noisy cost value at the base noise level.

As stated in the main text, in order to estimate the sample cost of error mitigation we will make the key assumption that

$$\mathrm{Var}[\widetilde{C}(\boldsymbol{\theta}, a)] \geq \mathrm{Var}[\widetilde{C}(\boldsymbol{\theta}, 1)] \,, \tag{100}$$

for all $\boldsymbol{\theta}$ and $a \geq 1$ that is, the statistical fluctuations in measurement outcomes at the boosted noise level are no smaller than that at the base noise level. Indeed, for noise models with a maximally mixed fixed point, we expect that high noise rates will tend to lead to larger variances. For example, in the simple scenario of a local Pauli measurement, the variance of measurement outcomes takes the form $\frac{(1-p_0)(p_0)}{N}$, where p_0 is the probability of obtaining a "0" outcome and N is the number of shots. This variance is maximized for $p_0 = \frac{1}{2}$.

E.1.1 Relative resolvability under global depolarizing noise

Proposition 1 (Relative resolvability of Zero-Noise Extrapolation with global depolarizing noise, 2 noise levels)**.** *Consider a circuit with L instances of global depolarizing noise of the form* (34)*. Consider a Richardson extrapolation strategy based on Eq.* (48)*, an exponential extrapolation strategy based on Eq.* (52) *and a NIBP extrapolation strategy based on Eq.* (54)*. We presume access to an augmented noisy circuit where the error probability is exactly increased by factor $a_1 > 1$ as $p \to a_1 p$. Then we have*

$$\chi_{depol} \leq \frac{\left(c - \frac{(1-a_1 p)^L}{(1-p)^L}\right)^2}{c^2+1} \,, \tag{101}$$

where χ_{depol} is the relative resolvability (see Definition 2) for global depolarizing noise, and where

$$c = \begin{cases} a_1 & \text{for Richardson extrapolation,} \\ \frac{a_1 r(\varepsilon)^{t(\varepsilon)}}{r(a_1\varepsilon)^{t(a_1\varepsilon)}} & \text{for exponential extrapolation,} \\ a_1^{-(L+1)} & \text{for NIBP extrapolation.} \end{cases} \tag{102}$$

Thus, $\chi_{depol} \leq 1$ for all of the above extrapolation strategies with access to 2 noise levels.

Proof. Upon inspecting Eqs. (49), (53) and (55), one can verify that the Richardson, exponential and NIBP extrapolation strategies all take the form

$$C_m(\boldsymbol{\theta}) = \frac{A \cdot \widetilde{C}(\boldsymbol{\theta}, 1) - B \cdot \widetilde{C}(\boldsymbol{\theta}, a)}{D} + E \,, \tag{103}$$

where $A, B \geq 0$ (note that for NIBP extrapolation E contains the state-independent cost value that represents the fixed point of the noise) and where we have adopted the notation defined in (99). We note that under L instances of global depolarizing noise (of the form (34)) with error probability p, noisy cost differences are given by

$$\Delta\widetilde{C}(a) = (1-ap)^L \Delta C \,, \tag{104}$$

for any pair of cost function points, where ΔC is the corresponding noise-free cost difference.

The error-mitigated cost function difference $\Delta C_m = C_m(\boldsymbol{\theta}_1) - C_m(\boldsymbol{\theta}_2)$ between two arbitrary points is given by

$$\Delta C_m = \frac{A \cdot \Delta\widetilde{C}(1) - B \cdot \Delta\widetilde{C}(a)}{D} \tag{105}$$

$$= \frac{A \cdot (1-p)^L \Delta C - B \cdot (1-ap)^L \Delta C}{D} \,. \tag{106}$$

Inspecting (105), we see that the error mitigation cost can be bounded simply as

$$\gamma = \frac{A^2 + B^2 \frac{\mathrm{Var}[\widetilde{C}(\boldsymbol{\theta},a)]}{\mathrm{Var}[\widetilde{C}(\boldsymbol{\theta},1)]}}{D^2} \tag{107}$$

$$\geq \frac{A^2+B^2}{D^2} \,, \tag{108}$$

where the inequality comes from our core assumption (100). Inserting γ, ΔC_m and $\Delta\widetilde{C}(1)$ into Definition 2, we have

$$\chi_{depol} = \frac{1}{\gamma}\left(\frac{\Delta C_m}{\Delta\widetilde{C}(1)}\right)^2 \leq \frac{\left(A(1-p)^L - B(1-ap)^L\right)^2}{(A^2+B^2)(1-p)^{2L}} \tag{109}$$

$$= \frac{\left(c - \frac{(1-ap)^L}{(1-p)^L}\right)^2}{c^2+1}\,, \tag{110}$$

where we have denoted $c = A/B$. By inspecting the specific values of A and B for the Richardson, exponential and NIBP extrapolation strategies respectively, we obtain the results for each strategy. □

E.1.2 Average relative resolvability

Proposition 2 (Average relative resolvability of Zero-Noise Extrapolation, 2 noise levels)**.** *Consider a Richardson extrapolation strategy based on Eq. (48), an exponential extrapolation strategy based on Eq. (52) and a NIBP extrapolation strategy based on Eq. (54). We presume perfect access to an augmented noisy circuit where the noise rate is increased by factor $a_1 > 1$. We denote $\boldsymbol{\theta}_{\varepsilon*}$ as the parameter corresponding to the global cost minimum at base noise parameter ε. Further denote $\frac{\langle\Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon*},a_1\varepsilon)\rangle_i}{\langle\Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon*},\varepsilon)\rangle_i} = z$. Any such noise model has an average relative resolvability*

$$\overline{\chi} \leq \frac{(z-c)^2}{c^2+1}\,, \tag{111}$$

where

$$c = \begin{cases} a_1 & \textit{for Richardson extrapolation,} \\ \frac{a_1 r(\varepsilon)^{t(\varepsilon)}}{r(a_1\varepsilon)^{t(a_1\varepsilon)}} & \textit{for exponential extrapolation,} \\ a_1^{-(L+1)} & \textit{for NIBP extrapolation}. \end{cases} \tag{112}$$

Thus, under the assumption that $z \leq 1$ and $\langle\Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon*}, a_1\varepsilon)\rangle_i \geq 0$, $\overline{\chi} \leq 1$ for all of the above extrapolation strategies with access to 2 noise levels.

Proof. As in the previous proof, we can inspect Eqs. (49), (53) and (55), and see that the Richardson, exponential and NIBP extrapolation strategies all take the form

$$C_m(\boldsymbol{\theta}) = \frac{A\cdot\widetilde{C}(\boldsymbol{\theta},1) - B\cdot\widetilde{C}(\boldsymbol{\theta},a)}{D} + E \tag{113}$$

where $A, B, D \geq 0$ (note that E contains the state-independent cost value that represents the fixed point of the noise) and we have adopted the notation of (99). The average mitigated cost differences (averaged over accessible parameters $\{\boldsymbol{\theta}_i\}_i$) can be written

$$\langle\Delta C_m(\boldsymbol{\theta}_{i,\varepsilon*})\rangle_i = \frac{A\cdot\langle\Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon*},1)\rangle_i - B\cdot\langle\Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon*},a)\rangle_i}{D}\,. \tag{114}$$

Thus, we have

$$\frac{\langle\Delta C_m(\boldsymbol{\theta}_{i,\varepsilon*})\rangle_i}{\langle\Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon*})\rangle_i} = \frac{A - B\frac{\langle\Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon*},a)\rangle_i}{\langle\Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon*},1)\rangle_i}}{D} \tag{115}$$

$$= \frac{A - Bz}{D}\,. \tag{116}$$

Finally, by noting once again that the error mitigation cost is simply bounded as $\gamma \geq \frac{A^2+B^2}{D^2}$ due to (100), we have

$$\chi \leq \frac{(A - Bz)^2}{A^2 + B^2}\,, \tag{117}$$

where we can obtain the desired form by defining $c = A/B$. Finally, the specific values of c for each extrapolation strategy can be read off by inspecting Eqs. (49), (53) and (55). □

We now introduce a modification of the basis-averaged relative resolvability in Definition 4 that we will use to prove an additional result for Zero-Noise Extrapolation. Here instead of averaging over the basis of an output state, we average over the basis of the measurement observable. This can be thought of as a more natural quantity to consider for comparing Zero-Noise Extrapolation to non-mitigated optimization as the protocol calls for processing of multiple noisy states.

Definition 5 (Basis-averaged relative resolvability II)**.** *Consider a spectrum $\boldsymbol{\lambda} \in \mathbb{R}^{2^n}_{\geq 0}$ with unit ℓ_1-norm, which corresponds to a noisy reference state. Then define the unitarily-averaged relative resolvability as*

$$\widehat{\widehat{\chi}}_{\boldsymbol{\lambda}} = \frac{1}{\gamma}\frac{\langle(\widehat{C}_m(\rho, U_i, O_{\boldsymbol{\lambda}}) - \mathrm{Tr}[O_{\boldsymbol{\lambda}}]/2^n)^2\rangle_{U_i}}{\langle(\widetilde{C}(\rho, U_i, O_{\boldsymbol{\lambda}}) - \mathrm{Tr}[O_{\boldsymbol{\lambda}}]/2^n)^2\rangle_{U_i}}\,, \tag{118}$$

where $\langle\cdot\rangle_{U_i}$ denotes an average over U_i drawn from a unitary 2-design, and where we denote

$$\widetilde{C}(\rho_{\boldsymbol{\lambda}}, U_i, O_{\boldsymbol{\lambda}}) = \mathrm{Tr}[U_i \rho_{\boldsymbol{\lambda}} U_i^\dagger O_{\boldsymbol{\lambda}}] \tag{119}$$

$$\widehat{C}_m(\rho_{\boldsymbol{\lambda}}, U_i, O_{\boldsymbol{\lambda}}) = \mathrm{Tr}[U_i \mathcal{M}(\rho_{\boldsymbol{\lambda}}) U_i^\dagger O_{\boldsymbol{\lambda}}] \tag{120}$$

where $\mathcal{M} : S(\mathcal{H}) \mapsto B(\mathcal{H})$ is the map that describes the action of the error mitigation protocol.

For this averaged relative resolvability we present a result for Zero-Noise Extrpolation.

Supplemental Proposition 1 (Basis-averaged relative resolvability II with Zero-Noise Extrapolation)**.** *Consider a Richardson extrapolation strategy based on Eq. (48), an exponential extrapolation strategy based on Eq. (52) and a NIBP extrapolation strategy based on Eq. (54). We presume perfect access to an augmented noisy circuit where the noise rate is increased by factor $a > 1$. Denote the output state at the base and augmented noise levels as $\rho(1)$ and $\rho(a)$ respectively. Then we have*

$$\widehat{\widehat{\chi}}_{\boldsymbol{\lambda}} \leq \frac{c^2 + \frac{P(a)-1/2^n}{P(1)-1/2^n}}{c^2+1}\,, \tag{121}$$

where $\widehat{\widehat{\chi}}_{\boldsymbol{\lambda}}$ is the averaged relative resolvability defined in Definition 5, $P(1)$ is the purity of $\rho(1)$, $P(a)$ is the purity of $\rho(a)$, and

$$c = \begin{cases} a & \textit{for Richardson extrapolation} \\ \frac{ar(\varepsilon)^{t(\varepsilon)}}{r(a\varepsilon)^{t(a\varepsilon)}} & \textit{for exponential extrapolation} \\ a^{-(L+1)} & \textit{for NIBP extrapolation}\,. \end{cases} \tag{122}$$

Thus, $\widehat{\widehat{\chi}}_{\boldsymbol{\lambda}} \leq 1$ when $P(a) \leq P(1)$.

Proof. We denote reference states $\widetilde{\rho}(\varepsilon)$ and $\widetilde{\rho}(a\varepsilon)$ as states with purity $P(\varepsilon)$ and $P(a\varepsilon)$ respectively. Moreover, denote the noisy cost function values $\widetilde{C}(U_i, \varepsilon) = \mathrm{Tr}[U_i \widetilde{\rho}(\varepsilon) U_i^\dagger O]$ and $\widetilde{C}(U_i, a\varepsilon) = \mathrm{Tr}[U_i \widetilde{\rho}(a\varepsilon) U_i^\dagger O]$ and further denote $C_m(U_i)$ as the corresponding error mitigated estimator. We start again by noting that in all three Zero-Noise Extrapolation strategies the estimator takes the form

$$C_m(U_i) = \frac{A \cdot \widetilde{C}(U_i, 1) - B \cdot \widetilde{C}(U_i, a)}{D} + E \tag{123}$$

where $A, B \geq 0$ (see Eqs. (49), (53) and (55)) and we have adopted the notation of (99). We first evaluate the relevant expectation values which correspond to integrals over the Haar distribution over the unitary group of degree 2^n. We now proceed to derive the result for Richardson/exponential extrapolation, however, we note that the proof follows in a similar way for NIBP extrapolation with the simple substitution $a\varepsilon \mapsto q/a$. Utilizing Lemma 3, we have

$$\langle \widetilde{C}(U_i,\varepsilon)\rangle_{U_i} = \frac{1}{2^n}\mathrm{Tr}[\widetilde{\rho}(\varepsilon)]\mathrm{Tr}[O] = \frac{1}{2^n}\mathrm{Tr}[O]\,, \tag{124}$$

$$\langle(\Delta\widetilde{C}(U_i,\varepsilon))^2\rangle_{U_i} = \langle\left(\widetilde{C}(U_i,\varepsilon) - \langle\widetilde{C}(U_j,\varepsilon)\rangle_{U_j}\right)^2\rangle_{U_i} = \frac{\left(\mathrm{Tr}[O^2] - \frac{1}{2^n}\mathrm{Tr}[O]^2\right)\left(\mathrm{Tr}[\widetilde{\rho}^2(\varepsilon)] - \frac{1}{2^n}\mathrm{Tr}[\widetilde{\rho}(\varepsilon)]^2\right)}{2^{2n}-1} \tag{125}$$

$$= \frac{\mathrm{Tr}[O^2] - \frac{1}{2^n}\mathrm{Tr}[O]^2}{2^{2n}-1}\left(P(\varepsilon) - \frac{1}{2^n}\right)\,, \tag{126}$$

$$\langle(\Delta\widetilde{C}(U_i,\varepsilon))(\Delta\widetilde{C}(U_i,a\varepsilon))\rangle_{U_i} = \frac{\left(\mathrm{Tr}[O^2] - \frac{1}{2^n}\mathrm{Tr}[O]^2\right)\left(\mathrm{Tr}[\widetilde{\rho}(\varepsilon)\widetilde{\rho}(a\varepsilon)] - \frac{1}{2^n}\mathrm{Tr}[\widetilde{\rho}(\varepsilon)]\mathrm{Tr}[\widetilde{\rho}(a\varepsilon)]\right)}{2^{2n}-1} \tag{127}$$

$$\geq 0\,, \tag{128}$$

where the inequality comes by observing that $\mathrm{Tr}[\widetilde{\rho}(\varepsilon)] = \mathrm{Tr}[\widetilde{\rho}(a\varepsilon)] = 1$ and further applying Cauchy-Schwarz to $\mathrm{Tr}[\widetilde{\rho}(\varepsilon)\widetilde{\rho}(a\varepsilon)]$ and noting that the purity of an n-qubit state is lower bounded by $1/2^n$. Inspecting Eq. (123) we have

$$\langle C_m(U_i)\rangle_{U_i} = \frac{1}{2^n}\frac{A-B}{D}\mathrm{Tr}[O] + E\,, \tag{129}$$

$$\langle\left(C_m(U_i) - \langle C_m(U_j)\rangle_{U_j}\right)^2\rangle_{U_i} = \left\langle\left(\frac{A\cdot\widetilde{C}(U_i,\varepsilon) - B\cdot\widetilde{C}(U_i,a\varepsilon)}{D} + E - \left(\frac{1}{2^n}\frac{A-B}{D}\mathrm{Tr}[O] + E\right)\right)^2\right\rangle_{U_i} \tag{130}$$

$$= \left\langle\left(\frac{A\cdot\Delta\widetilde{C}(U_i,\varepsilon) - B\cdot\Delta\widetilde{C}(U_i,a\varepsilon)}{D}\right)^2\right\rangle_{U_i} \tag{131}$$

$$= \frac{A^2\langle(\Delta\widetilde{C}(U_i,\varepsilon))^2\rangle_{U_i} + B^2\langle(\Delta\widetilde{C}(U_i,a\varepsilon))^2\rangle_{U_i} - 2AB\langle\Delta\widetilde{C}(U_i,\varepsilon)\Delta\widetilde{C}(U_i,a\varepsilon)\rangle_{U_i}}{D^2} \tag{132}$$

$$\leq \frac{\mathrm{Tr}[O^2] - \frac{1}{2^n}\mathrm{Tr}[O]^2}{D^2(2^{2n}-1)}\left(A^2\left(P(\varepsilon) - \frac{1}{2^n}\right) + B^2\left(P(a\varepsilon) - \frac{1}{2^n}\right)\right)\,. \tag{133}$$

The inequality comes by substituting in the expressions for $\langle(\Delta\widetilde{C}(U_i,\varepsilon))^2\rangle_{U_i}$ and $\langle(\Delta\widetilde{C}(U_i,a\varepsilon))^2\rangle_{U_i}$ obtained in (126), and dropping the third term in the numerator, where we have used Eq. (128). Finally, we note that Eq. (103) gives $\gamma^{-1} = \frac{D^2}{A^2+B^2}$. Substituting the obtained expressions for γ^{-1}, (133) and (126) into Definition 5 we obtain

$$\overline{\overline{\chi}}_{\boldsymbol{\lambda}} \leq \frac{A^2 + B^2\frac{P(a\varepsilon)-1/2^n}{P(\varepsilon)-1/2^n}}{A^2+B^2}\,, \tag{134}$$

where we can define $c = A/B$ to obtain the desired result. Further, the explicit form of c for Richardson, exponential and NIBP extrapolation can be respectively found by inspecting Eqs. (49), (53) and (55). □

E.1.3 Richardson extrapolation with 3 noise levels

In this section we focus on Richardson extrapolation (see Appendix B.1.1 for review) and investigate the change in resolvability under an extrapolation strategy that utilizes 3 distinct noise levels.

Supplemental Proposition 2 (Relative resolvability of Richardson extrapolation with global depolarizing noise, 3 noise levels)**.** *Consider L instances of global depolarizing noise of the form (34) acting through a circuit. Consider a Richardson extrapolation strategy based on Eq. (48), an exponential extrapolation strategy based on Eq. (52) and a NIBP extrapolation strategy based on Eq. (54) in the appendix. We presume access to two augmented noisy circuits where the error probability is perfectly increased by factors $a_2 > a_1 > 1$ as $p \to a_1 p$ and $p \to a_2 p$ respectively. Then for all three extrapolation strategies and any such choices of a_2 and a_1, we have*

$$\chi_{depol} \leq 1\,, \tag{135}$$

where χ_{depol} is the relative resolvability (see Definition 2) for global depolarizing noise.

Proof. We start by noting that under L instances of global depolarizing noise with error probability p (of the form (34)), noisy cost differences are given by

$$\Delta\widetilde{C}(a) = (1-ap)^L \Delta C\,, \tag{136}$$

for any noise augmentation factor a and any pair of cost function points, where ΔC is the corresponding noise-free cost difference.

The error-mitigated cost function difference $\Delta C_m(\boldsymbol{\theta}_{1,2}) = C_m(\boldsymbol{\theta}_1) - C_m(\boldsymbol{\theta}_2)$ between two arbitrary points constructed under Richardson extrapolation with 3 noise levels is given by

$$\begin{aligned}\Delta C_m &= \frac{a_1a_2(a_2-a_1)\Delta\widetilde{C}(p) - a_2(a_2-1)\Delta\widetilde{C}(a_1p) + a_1(a_1-1)\Delta\widetilde{C}(a_2p)}{(a_1-1)(a_2-1)(a_2-a_1)} &(137)\\ &= \frac{a_1a_2(a_2-a_1)(1-p)^L\Delta C - a_2(a_2-1)(1-a_1p)^L\Delta C + a_1(a_1-1)(1-a_2p)^L\Delta C}{(a_1-1)(a_2-1)(a_2-a_1)}\,, &(138)\end{aligned}$$

where in order to obtain the first equality we have used (51). The second equality comes by substituting in (136). Inspecting (105), we see that the error mitigation cost can be bounded simply as

$$\begin{aligned}\gamma &= \frac{a_1^2a_2^2(a_2-a_1)^2 + a_2^2(a_2-1)^2\frac{\mathrm{Var}[\widetilde{C}(\boldsymbol{\theta},a_1p)]}{\mathrm{Var}[\widetilde{C}(\boldsymbol{\theta},p)]} + a_1^2(a_1-1)^2\frac{\mathrm{Var}[\widetilde{C}(\boldsymbol{\theta},a_2p)]}{\mathrm{Var}[\widetilde{C}(\boldsymbol{\theta},p)]}}{(a_1-1)^2(a_2-1)^2(a_2-a_1)^2} &(139)\\ &\geq \frac{a_1^2a_2^2(a_2-a_1)^2 + a_2^2(a_2-1)^2 + a_1^2(a_1-1)^2}{(a_1-1)^2(a_2-1)^2(a_2-a_1)^2} &(140)\end{aligned}$$

for any $\boldsymbol{\theta}$, where the inequality comes from our core assumption (100). Inserting our expressions for γ, ΔC_m and $\Delta\widetilde{C}(1)$ into Definition 2, we have

$$\begin{aligned}\chi_{depol} = \frac{1}{\gamma}\left(\frac{\Delta C_m}{\Delta\widetilde{C}(1)}\right)^2 &\leq \frac{\left(a_1a_2(a_2-a_1)(1-p)^L - a_2(a_2-1)(1-a_2p)^L + a_1(a_1-1)(1-a_2p)^L\right)^2}{\left(a_1^2a_2^2(a_2-a_1)^2 + a_2^2(a_2-1)^2 + a_1^2(a_1-1)^2\right)(1-p)^{2L}} &(141)\\ &= \frac{\left(a_1a_2(a_2-a_1) - a_2(a_2-1)\frac{(1-a_1p)^L}{(1-p)^L} + a_1(a_1-1)\frac{(1-a_2p)^L}{(1-p)^L}\right)^2}{a_1^2a_2^2(a_2-a_1)^2 + a_2^2(a_2-1)^2 + a_1^2(a_1-1)^2}\,. &(142)\end{aligned}$$

The desired result can be observed by noting that $a_2(a_2-1) > a_1(a_1-1)$ and that $\frac{(1-a_1p)^L}{(1-p)^L} > \frac{(1-a_2p)^L}{(1-p)^L}$. □

Supplemental Proposition 3 (Average resolvability of Richardson extrapolation, 3 noise levels)**.** *Consider a Richardson extrapolation strategy based on Eq. (48), an exponential extrapolation strategy based on Eq. (52) and a NIBP extrapolation strategy based on Eq. (54) in the appendix. We presume perfect access to two augmented noisy circuits where the noise rate is increased by factors $a_2 > a_1 > 1$.*

We denote $\boldsymbol{\theta}_{\varepsilon *}$ as the parameter corresponding to the global cost minimum at base noise parameter ε. Further denote $\frac{\langle \Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon *}, a_1\varepsilon)\rangle_i}{\langle \Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon *}, \varepsilon)\rangle_i} = z_1$ and $\frac{\langle \Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon *}, a_2\varepsilon)\rangle_i}{\langle \Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon *}, \varepsilon)\rangle_i} = z_2$. Any such noise model has an average relative resolvability

$$\overline{\chi} \leq \frac{(a_1a_2(a_2-a_1) - a_2(a_2-1)z_1 + a_1(a_1-1)z_2)^2}{a_1^2a_2^2(a_2-a_1)^2 + a_2^2(a_2-1)^2 + a_1^2(a_1-1)^2}\,, \tag{143}$$

where $\overline{\chi}$ is the averaged relative resolvability (see Definition 2). Thus, under the assumption that $z_2 \leq z_1 \leq 1$ (on average the cost concentrates with increasing noise level) and $\langle \Delta\widetilde{C}_{i,\varepsilon *}(a_1\varepsilon)\rangle_i, \langle \Delta\widetilde{C}_{i,\varepsilon *}(a_2\varepsilon)\rangle_i \geq 0$ (boosting the noise level does not shift the cost value of the global minimum above the average cost value), then $\overline{\chi} \leq 1$.

Proof. The averaged error-mitigated cost function difference $\langle \Delta C_m(\boldsymbol{\theta}_{i,\varepsilon *})\rangle_i = \langle C_m(\boldsymbol{\theta}_i) - C_m(\boldsymbol{\theta}_{\varepsilon *})\rangle_i$ between two arbitrary points constructed under Richardson extrapolation with 3 noise levels is given by

$$\langle \Delta C_m(\boldsymbol{\theta}_{i,\varepsilon *})\rangle_i = \left\langle \left(\frac{a_1a_2(a_2-a_1)\Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon *}, p) - a_2(a_2-1)\Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon *}, a_1p) + a_1(a_1-1)\Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon *}, a_2p)}{(a_1-1)(a_2-1)(a_2-a_1)} \right)_{i,\varepsilon *} \right\rangle_i \tag{144}$$

$$= \frac{a_1a_2(a_2-a_1)\langle\Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon *}, p)\rangle_i - a_2(a_2-1)\langle\Delta\widetilde{C}((\boldsymbol{\theta}_{i,\varepsilon *}, a_1p)\rangle_i + a_1(a_1-1)\langle\Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon *}, a_2p)\rangle_i}{(a_1-1)(a_2-1)(a_2-a_1)}\,. \tag{145}$$

As in the previous proof, we can inspect (105) and we see that the error mitigation cost can be bounded simply as

$$\gamma = \frac{a_1^2a_2^2(a_2-a_1)^2 + a_2^2(a_2-1)^2\frac{\mathrm{Var}[\widetilde{C}(\boldsymbol{\theta},a_1p)]}{\mathrm{Var}[\widetilde{C}(\boldsymbol{\theta},p)]} + a_1^2(a_1-1)^2\frac{\mathrm{Var}[\widetilde{C}(\boldsymbol{\theta},a_2p)]}{\mathrm{Var}[\widetilde{C}(\boldsymbol{\theta},p)]}}{(a_1-1)^2(a_2-1)^2(a_2-a_1)^2} \tag{146}$$

$$\geq \frac{a_1^2a_2^2(a_2-a_1)^2 + a_2^2(a_2-1)^2 + a_1^2(a_1-1)^2}{(a_1-1)^2(a_2-1)^2(a_2-a_1)^2} \tag{147}$$

for any $\boldsymbol{\theta}$, where the inequality comes from our core assumption (100). Inserting our expressions for γ and ΔC_m into Definition 2, we have

$$\overline{\chi} = \frac{1}{\gamma}\left(\frac{\Delta C_m}{\Delta\widetilde{C}(1)}\right)^2 \leq \frac{\left(a_1a_2(a_2-a_1) - a_2(a_2-1)\frac{\langle \Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon *}, a_1\varepsilon)\rangle_i}{\langle \Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon *}, \varepsilon)\rangle_i} + a_1(a_1-1)\frac{\langle \Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon *}, a_2\varepsilon)\rangle_i}{\langle \Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon *}, \varepsilon)\rangle_i}\right)^2}{a_1^2a_2^2(a_2-a_1)^2 + a_2^2(a_2-1)^2 + a_1^2(a_1-1)^2}\,, \tag{148}$$

and the desired result comes by denoting $\frac{\langle \Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon *}, a_1\varepsilon)\rangle_i}{\langle \Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon *}, \varepsilon)\rangle_i} = z_1$ and $\frac{\langle \Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon *}, a_2\varepsilon)\rangle_i}{\langle \Delta\widetilde{C}(\boldsymbol{\theta}_{i,\varepsilon *}, \varepsilon)\rangle_i} = z_2$. □

As with the results of Proposition 2 we see that $\overline{\chi}$ decreases with increasing cost concentration.

E.2 Virtual Distillation

E.2.1 Bounds on error mitigation cost

We recall the two error mitigation protocols of Ref. [47], denoted "A" and "B" respectively, to prepare

$$C_m^{(A)}(\boldsymbol{\theta}_i) = \mathrm{Tr}[\widetilde{\rho}_i^M O]/\mathrm{Tr}[\widetilde{\rho}_i^M]\,, \tag{149}$$

and

$$C_m^{(B)}(\boldsymbol{\theta}_i) = \mathrm{Tr}[\widetilde{\rho}_i^M O]/\lambda_i^M\,, \tag{150}$$

where λ_i is the dominant eigenvalue of $\widetilde{\rho}_i \equiv \widetilde{\rho}(\boldsymbol{\theta}_i)$. The protocols considered explicitly construct these quantities as

$$\begin{aligned} \mathrm{Tr}[\widetilde{\rho}_i O] &= 2\mathrm{prob}_{1,i} - 1\,, && (151)\\ \mathrm{Tr}[\widetilde{\rho}_i^M O] &= 2\mathrm{prob}_{M,i} - 1\,, && (152)\\ \mathrm{Tr}[\widetilde{\rho}_i^M] &= 2\mathrm{prob}'_{M,i} - 1\,, && (153) \end{aligned}$$

where $\mathrm{prob}_{1,i}$, $\mathrm{prob}_{M,i}$ and $\mathrm{prob}'_{M,i}$ are expectation values of a Pauli-Z measurement on a qubit ancillary subsystem. In order to obtain our results we will hereon make the core assumption

$$\mathrm{Var}[\mathrm{prob}_{M,i}] \geq \mathrm{Var}[\mathrm{prob}_{1,i}] \quad \forall i,\, M \geq 2\,, \tag{154}$$

that is, the statistical uncertainty of the measurement outcomes of the circuit that prepares $\widetilde{\rho}_i^M$ are at best equal to that of $\widetilde{\rho}_i$. In the case of large M we expect $\mathrm{Var}[\mathrm{prob}_{M,i}]$ to be large, as for any (non-pure) $\widetilde{\rho}$, the quantity $\mathrm{Tr}[\widetilde{\rho}^M O]$ is close to zero for large M. This corresponds to $\mathrm{prob}_{M,i} = \frac{1}{2}$, which maximizes the variance for a binomial distribution.

Lemma 4 (Bounds on error mitigation cost of virtual distillation)**.** *Denote the error mitigation cost (see Definition 1) corresponding to (149) and (150) as $\gamma^{(A)}$ and $\gamma^{(B)}$ respectively. We have*

$$\gamma^{(A)} \geq \frac{1}{(\mathrm{Tr}[\widetilde{\rho}^M])^2}\,, \quad \gamma^{(B)} \geq \frac{1}{\lambda^{2M}}\,. \tag{155}$$

Proof. For $\gamma^{(A)}$ and $\gamma^{(B)}$ we need to compute the variances of the estimators of $C_m^{(A)}$, $C_m^{(B)}$ respectively and likewise $\widetilde{C} = \mathrm{Tr}[\widetilde{\rho}O]$. We have

$$\begin{aligned} \mathrm{Var}[\widetilde{C}] = \mathrm{Var}[\mathrm{Tr}[\widetilde{\rho}O]] &= \mathrm{Var}[2\mathrm{prob}_1 - 1]\,, && (156)\\ &= 4\mathrm{Var}[\mathrm{prob}_1]\,, && (157) \end{aligned}$$

$$\begin{aligned} \mathrm{Var}[C_m^{(B)}] = \mathrm{Var}\Big[\frac{\mathrm{Tr}[\widetilde{\rho}^M O]}{\lambda^M}\Big] &= \frac{1}{\lambda^{2M}}\mathrm{Var}[2\mathrm{prob}_M - 1]\,, && (158)\\ &= \frac{4}{\lambda^{2M}}\mathrm{Var}[\mathrm{prob}_M]\,, && (159) \end{aligned}$$

$$\begin{aligned} \mathrm{Var}[C_m^{(A)}] = \mathrm{Var}\Big[\frac{\mathrm{Tr}[\widetilde{\rho}^M O]}{\mathrm{Tr}[\widetilde{\rho}^M]}\Big] &= 4\mathrm{Var}[\mathrm{prob}_M]\left(\mathbb{E}\left[\frac{1}{2\mathrm{prob}'_M - 1}\right]\right)^2 + 4\mathrm{Tr}[\widetilde{\rho}^M O]^2\,\mathrm{Var}\left[\frac{1}{2\mathrm{prob}'_M - 1}\right] && (160)\\ &\quad + 4\mathrm{Var}[\mathrm{prob}_M]\,\mathrm{Var}\left[\frac{1}{2\mathrm{prob}'_M - 1}\right]\\ &\geq 4\mathrm{Var}[\mathrm{prob}_M]\left(\mathbb{E}\left[\frac{1}{2\mathrm{prob}'_M - 1}\right]\right)^2 && (161)\\ &\geq 4\mathrm{Var}[\mathrm{prob}_M]\,\frac{1}{\left(\mathbb{E}\left[2\mathrm{prob}'_M - 1\right]\right)^2} && (162)\\ &= 4\mathrm{Var}[\mathrm{prob}_M]\,\frac{1}{\left(\mathrm{Tr}[\widetilde{\rho}^M]\right)^2}\,. && (163) \end{aligned}$$

Equation (160) comes from the standard formula for the variance of the product of two independent random variables. To obtain the first inequality we simply drop the second and third terms, which are positive. The second inequality is an application of Jensen's inequality. Recalling the definition of error mitigation cost (Definition 1), the above three equations enable us to write

$$\gamma^{(A)} \geq \frac{1}{(\mathrm{Tr}[\widetilde{\rho}^M])^2}\,, \quad \gamma^{(B)} \geq \frac{1}{\lambda^{2M}}\,, \tag{164}$$

where we have used our core assumption that $\mathrm{Var}[\mathrm{prob}_1] \leq \mathrm{Var}[\mathrm{prob}_M]$. □

E.2.2 Relative resolvability for global depolarizing noise

Here we present a proof of Proposition 3, in which we upper bound the relative resolvability for Virtual Distillation, for any two cost function points under global depolarizing noise.

Proposition 3 (Relative resolvability of Virtual Distillation with global depolarizing noise)**.** *Consider l instances of global depolarizing noise $\mathcal{D}$ of the form*

$$\rho \xrightarrow{\mathcal{D}} \widetilde{\rho} = q^l \rho + (1 - q^l)\frac{\mathbb{1}}{2^n} \tag{165}$$

acting on some pure state ρ with some noise parameter $q \in [0, 1)$. We consider the two error mitigation protocols of Ref. [47] (denoted "A" and "B") to respectively prepare (149) and (150). The relative resolvability of any pair of arbitrary cost function points satisfies

$$\chi^{(A)} \leq \chi^{(B)} = \Gamma(n, M, q^l) \tag{166}$$

for all $n \geq 1$, $M \geq 2$, $q^l \in [0, 1]$, and where

$$\Gamma(n, M, q^l) \leq 1\,, \tag{167}$$

is a monotonically decreasing function in M (with asymptotically exponential decay) in the quadrant $n \geq 1$, $M \geq 2$. The bound is saturated as $\Gamma(1, 2, p) = 1$ for all p.

Proof. In this proof we consider arbitrary cost function differences, that is, given two arbitrarily chosen points in parameter space $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, we consider

$$\Delta C = C(\boldsymbol{\theta}_1) - C(\boldsymbol{\theta}_2)\,, \tag{168}$$

and the respective differences for the noisy cost $\widetilde{C}(\boldsymbol{\theta})$ and the mitigated costs $C_m^{(A)}(\boldsymbol{\theta})$, $C_m^{(B)}(\boldsymbol{\theta})$. In order to evaluate χ_A and χ_B we need to first evaluate the following quantities:

$$\Delta\widetilde{C}\,,\ \Delta C_m^{(A)}\,,\ \Delta C_m^{(B)}\,,\ \gamma^{(A),(B)} \tag{169}$$

that is, the noisy cost function difference between two points, the difference between the virtual distillation estimators for the same points for both protocols, and the error mitigation rate for both protocols. The noisy cost function difference under global depolarizing noise is simply related the noiseless difference as

$$\Delta\widetilde{C} = q^l \Delta C\,. \tag{170}$$

To evaluate the other quantities we note that

$$\widetilde{\rho} = \Big[q^l + \frac{1}{2^n}(1 - q^l)\Big]\rho + \Big[\frac{1}{2^n}(1 - q^l)\Big](\mathbb{1} - \rho)\,, \tag{171}$$

$$\widetilde{\rho}^M = \left[\Big[q^l + \frac{1}{2^n}(1 - q^l)\Big]^M - \Big[\frac{1}{2^n}(1 - q^l)\Big]^M\right]\rho + \frac{2^n}{2^{nM}}(1 - q^l)^M\frac{\mathbb{1}}{2^n}\,, \tag{172}$$

$$\mathrm{Tr}[\widetilde{\rho}^M] = \Big[q^l + \frac{1}{2^n}(1 - q^l)\Big]^M + \frac{2^n - 1}{2^{nM}}(1 - q^l)^M\,, \tag{173}$$

$$\mathrm{Tr}[\widetilde{\rho}^M O] = \left[\Big[q^l + \frac{1}{2^n}(1 - q^l)\Big]^M - \Big[\frac{1}{2^n}(1 - q^l)\Big]^M\right]\mathrm{Tr}[\rho O] + \frac{1}{2^{nM}}(1 - q^l)^M\mathrm{Tr}[O]\,. \tag{174}$$

In particular, we highlight that the expression for $\mathrm{Tr}[\widetilde{\rho}^M]$ is independent of the noise-free output state ρ. As the dominant noisy eigenvalue λ and $\mathrm{Tr}[\widetilde{\rho}^M]$ are state independent we have

$$\Delta C_m^{(A)} = \frac{1}{\mathrm{Tr}[\widetilde{\rho}^M]}\left[\Big[q^l + \frac{1}{2^n}(1 - q^l)\Big]^M - \Big[\frac{1}{2^n}(1 - q^l)\Big]^M\right]\Delta C\,, \tag{175}$$

$$\Delta C_m^{(B)} = \frac{1}{\lambda^M}\left[\Big[q^l + \frac{1}{2^n}(1 - q^l)\Big]^M - \Big[\frac{1}{2^n}(1 - q^l)\Big]^M\right]\Delta C\,, \tag{176}$$

where the choice of $\widetilde{\rho}$ is arbitrary.

Now, using Definition 2 and combining (175), (176), (170) along with the result of Lemma 4, we have

$$\chi^{(A)} \leq \chi^{(B)} = \Gamma(n, M, q^l)\,, \tag{177}$$

where we define the function

$$\Gamma(n, M, q^l) = \frac{1}{q^{2l}} \left[\left[q^l + \frac{1}{2^n}(1-q^l) \right]^M - \left[\frac{1}{2^n}(1-q^l) \right]^M \right]^2 . \tag{178}$$

First, note that for $M = 2$

$$\Gamma(q^l, n, 2) = \frac{1}{q^{2l}} \left[\left[q^l + \frac{1}{2^n}(1-q^l) \right]^2 - \left[\frac{1}{2^n}(1-q^l) \right]^2 \right]^2 \tag{179}$$

$$= \left(q^l + \frac{2}{2^n}(1-q^l) \right)^2 . \tag{180}$$

For $n = 1$, $\Gamma(n, 2, q^l) = 1$. For all $n > 1$, $\Gamma(n, 2, q^l) < 1$ as $(1 - q^l) > 0$. Thus,

$$\Gamma(n, 2, q^l) \leq 1 \quad \forall\, n \geq 1\,. \tag{181}$$

We complete the proof by showing that $\Gamma(n, M, q^l)$ monotonically decreases with M for all $n \geq 1$, $M \geq 2$. This can be seen by inspecting the partial derivative (making the decomposition $\Gamma = (\Gamma^{1/2})^2$ due to the square in (178))

$$\frac{\partial \Gamma}{\partial M} = 2\Gamma^{1/2} \frac{\partial \Gamma^{1/2}}{\partial M}\,. \tag{182}$$

We investigate when this quantity is negative. As $\Gamma^{1/2}$ is always positive, negativity is determined by the sign of $\frac{\partial \Gamma^{1/2}}{\partial M}$. Denoting $\delta = \frac{1}{2^n}(1 - q^l)$, we have

$$\frac{\partial \Gamma^{1/2}}{\partial M} = \frac{1}{q^l} \left[(q^l + \delta)^M \ln(q^l + \delta) - \delta^M \ln \delta \right] \tag{183}$$

$$= \frac{1}{q^l} \left[\delta^M \left(\ln(q^l + \delta) - \ln \delta \right) + \ln(q^l + \delta) \left((q^l + \delta)^M - \delta^M \right) \right] \tag{184}$$

$$\leq \frac{1}{q^l} \left[q^l \delta^{M-1} + (q^l + \delta - 1) \left((q^l + \delta)^M - \delta^M \right) \right] \tag{185}$$

$$= \frac{1}{q^l} \left[q^l \delta^{M-1} - (2^n - 1)\delta \left((q^l + \delta)^M - \delta^M \right) \right] \tag{186}$$

$$\leq \frac{1}{q^l} \left[q^l \delta^{M-1} - (2^n - 1)\delta \left(M q^l \delta^{M-1} + \frac{1}{2} M(M-1) q^{2l} \delta^{M-2} \right) \right] \tag{187}$$

$$= \frac{1}{q^l} \left[q^l \delta^{M-1} - (2^n - 1)\delta \left(\frac{1}{2^n} M(1-q^l) q^l \delta^{M-2} + \frac{1}{2} M(M-1) q^{2l} \delta^{M-2} \right) \right] \tag{188}$$

$$= \delta^{M-1} \left[1 - \left(1 - \frac{1}{2^n} \right) M - \frac{1}{2}(2^n - 1) M \left(M - 1 - \frac{2}{2^n} \right) q^l \right] \tag{189}$$

$$\leq \delta^{M-1} \left[1 - \frac{1}{2} M - \frac{1}{2} M\,(M-2)\, q^l \right] \quad \forall\, n \geq 1\,, \tag{190}$$

where in order to obtain the first inequality we use the inequalities $\ln(q^l + \delta) - \ln \delta \leq q^l/\delta$ and $\ln(q^l + \delta) \leq q^l + \delta - 1$. The second inequality comes from observing that the expansion of $\left((q^l + \delta)^M - \delta^M \right)$ is a sum of positive terms, and considering only two such terms. The above implies that

$$\frac{\partial \Gamma}{\partial M} \leq 0 \quad \forall\, n \geq 1, M \geq 2\,, \tag{191}$$

that is, Γ is monotonically decreasing with M in the quadrant $n \geq 1$ $M \geq 2$. Combined with (181), we have the proof as required. □

E.2.3 Average relative resolvability

Here we present a proof of Proposition 4, in which we upper bound the 2-design-averaged relative resolvability for Virtual Distillation.

Proposition 4 (Average relative resolvability of Virtual Distillation)**.** *Consider an error mitigation protocol that prepares estimator $C_m(\boldsymbol{\theta}_i) = \mathrm{Tr}[\widetilde{\rho}_i^M O]/\mathrm{Tr}[\widetilde{\rho}_i^M]$ from some noisy parameterized quantum state $\widetilde{\rho}_i \equiv \widetilde{\rho}(\boldsymbol{\theta}_i)$. Consider the average relative resolvability for noisy states of some spectrum $\boldsymbol{\lambda}$ with purity $P_{\boldsymbol{\lambda}}$ as defined in Definition 3. We have*

$$\overline{\overline{\chi}}_{\boldsymbol{\lambda}} \leq G(n, M, P) \leq 1 \,, \tag{192}$$

where $G(n, M, P)$ is a monotonically decreasing function in M (with asymptotically exponential decay) for all $n \geq 1$, $M \geq 2$. Within this region the bound is saturated as $G(1, 2, P) = 1$ for all P and $G(n, M, 1) = 1$ for all $n \geq 1, M \geq 2$. Explicitly, we have for $n = 1$

$$G(n = 1, M, P) = \frac{1}{2^{2M}} \frac{\left[(1 + \sqrt{2P-1})^M - (1 - \sqrt{2P-1})^M\right]^2}{2P - 1} \,. \tag{193}$$

For $n \geq 2$ and $M = 2$

$$G(n \geq 2, M = 2, P) = \min\left(\frac{4}{2^{2n}} + \frac{4}{2^{n/2}} g_2 \sqrt{P - \frac{1}{2^n}} + 2^n g_2^2 \left(P - \frac{1}{2}\right),\ \frac{P^2}{P - \frac{1}{2^n}}\left(1 - \frac{1}{2^n}\right)\right) \,, \tag{194}$$

where we denote $g_k = \left(\frac{2^n - 1}{2^n}\right)^k + \left(\frac{1}{2^n}\right)^k$. Further, for $n \geq 2$ and $M \geq 3$ we have

$$G(n \geq 2, M \geq 3, P) = \min\left(\frac{2^n}{4} \frac{\left[\left(\sqrt{2\left(P - \frac{1}{2^n}\right)} + \frac{1}{2^n}\right)^M - \left(\frac{1}{2^n}\right)^M\right]^2}{P - \frac{1}{2^n}},\ \frac{P^M}{P - \frac{1}{2^n}}\left(1 - \frac{1}{2^n}\right)\right) \,. \tag{195}$$

Proof. From Definition 4 we have

$$\overline{\overline{\chi}}_{\boldsymbol{\lambda}} = \frac{1}{\gamma(\boldsymbol{\lambda})} \frac{\left\langle \left(C_m(\rho_{\boldsymbol{\lambda}}, U_i) - \mathrm{Tr}[O]/2^n\right)^2 \right\rangle_{U_i}}{\left\langle \left(\widetilde{C}(\rho_{\boldsymbol{\lambda}}, U_i) - \mathrm{Tr}[O]/2^n\right)^2 \right\rangle_{U_i}} \,. \tag{196}$$

Let us first evaluate the required averages over unitary 2-designs. The relevant first moments for virtual distillation are given by

$$\langle \mathrm{Tr}[U \rho_{\boldsymbol{\lambda}} U^\dagger O] \rangle_U = \mathrm{Tr}[O]/2^n \,, \tag{197}$$

$$\langle \mathrm{Tr}[U \rho_{\boldsymbol{\lambda}}^M U^\dagger O] \rangle_U = \mathrm{Tr}[\rho_{\boldsymbol{\lambda}}^M] \mathrm{Tr}[O]/2^n \,, \tag{198}$$

where we have used Lemma 3. Thus we can see that the numerator and denominator of (196) correspond to variances which we now evaluate. Again, utilizing Lemma 3, the second moments are given by

$$\left\langle \left(\widetilde{C}(U_i) - \langle \widetilde{C}(U_j) \rangle_{U_j}\right)^2 \right\rangle_{U_i} = \langle (\mathrm{Tr}[U \rho_{\boldsymbol{\lambda}} U^\dagger O])^2 \rangle_U - (\mathrm{Tr}[O]/2^n)^2 \tag{199}$$

$$= \frac{\left(\mathrm{Tr}[O^2] - \frac{1}{2^n}\mathrm{Tr}[O]^2\right)\left(\mathrm{Tr}[\rho_{\boldsymbol{\lambda}}^2] - \frac{1}{2^n}\mathrm{Tr}[\rho_{\boldsymbol{\lambda}}]^2\right)}{2^{2n} - 1} \tag{200}$$

$$\left\langle \left(C_m(U_i) - \langle C_m(U_j) \rangle_{U_j}\right)^2 \right\rangle_{U_i} = \left\langle \left(\frac{\mathrm{Tr}[U \rho_{\boldsymbol{\lambda}}^M U^\dagger O]}{\mathrm{Tr}[\rho_{\boldsymbol{\lambda}}^M]}\right)^2 \right\rangle_U - (\mathrm{Tr}[O]/2^n)^2 \tag{201}$$

$$= \frac{\left(\mathrm{Tr}[O^2] - \frac{1}{2^n}\mathrm{Tr}[O]^2\right)\left(\frac{\mathrm{Tr}[\rho_{\boldsymbol{\lambda}}^{2M}]}{\mathrm{Tr}[\rho_{\boldsymbol{\lambda}}^M]^2} - \frac{1}{2^n}\right)}{2^{2n} - 1} \,, \tag{202}$$

where in the final equality we have used the fact that $\mathrm{Tr}\left[\frac{\rho_{\boldsymbol{\lambda}}^M}{\mathrm{Tr}[\rho_{\boldsymbol{\lambda}}^M]}\right] = 1$. Using the definition of the basis-averaged relative resolvability (Definition 4), we can arrive at a bound written explicitly in terms of $\rho_{\boldsymbol{\lambda}}$ as

$$\overline{\overline{\chi}}_{\boldsymbol{\lambda}} = \frac{1}{\gamma}\frac{\langle (C_m(U_i) - \langle C_m(U_j)\rangle_{U_j})^2\rangle_{U_i}}{\langle (\widetilde{C}(U_i) - \langle \widetilde{C}(U_j)\rangle_{U_j})^2\rangle_{U_i}} \leq \frac{\mathrm{Tr}[\rho_{\boldsymbol{\lambda}}^{2M}] - \frac{1}{2^n}\mathrm{Tr}[\rho_{\boldsymbol{\lambda}}^M]^2}{\mathrm{Tr}[\rho_{\boldsymbol{\lambda}}^2] - \frac{1}{2^n}\mathrm{Tr}[\rho_{\boldsymbol{\lambda}}]^2}\,, \tag{203}$$

where we have used the fact that the error mitigation cost $\gamma \geq 1/(\mathrm{Tr}[\rho_{\boldsymbol{\lambda}}^M])^2$.

The goal is to now investigate whether or not $f(M) = \mathrm{Tr}[\rho_{\boldsymbol{\lambda}}^{2M}] - \frac{1}{2^n}\mathrm{Tr}[\rho_{\boldsymbol{\lambda}}^M]^2$ is monotonically decreasing for $M \in \mathbb{N}_+$. This quantity has two interpretations. First, it can be seen to be a Hilbert Schmidt distance between $\rho_{\boldsymbol{\lambda}}^M$ and $\mathrm{Tr}[\rho_{\boldsymbol{\lambda}}^M]\frac{\mathbb{1}}{2^n}$. Second, by considering the eigenvalue decomposition of ρ, it can be seen to be proportional to the population variance of the distribution $\{\lambda_i^M\}$, where λ_i are the eigenvalues of $\rho_{\boldsymbol{\lambda}}$, that is,

$$f(M) = 2^n \mathrm{Var}[\boldsymbol{\lambda}^{(M)}] = \sum_i \lambda_i^{2M} - \frac{1}{2^n}\left(\sum_i \lambda_i^M\right)^2\,, \tag{204}$$

where here $\mathrm{Var}[(.)]$ denotes the population variance of the contained vector. Thus, we can rewrite Eq. (203) as

$$\overline{\overline{\chi}}_{\boldsymbol{\lambda}}(M) \leq \frac{f(M)}{f(1)} = \frac{\mathrm{Var}[\boldsymbol{\lambda}^{(M)}]}{\mathrm{Var}[\boldsymbol{\lambda}^{(1)}]}\,. \tag{205}$$

Let us first treat the qubit setting of $n = 1$. Consider eigenvalue decomposition $\rho_{\boldsymbol{\lambda}} = \lambda|\psi\rangle\langle\psi| + (1-\lambda)|\psi_\perp\rangle\langle\psi_\perp|$, where we have defined $\lambda_1 = 1-\lambda$, $\lambda_2 = \lambda$ and without loss of generality we fix $1 - \lambda \geq \lambda$. We define $G(1, M, P) = f(M)/f(1)$ and will determine $f(M)$ exactly for single-qubit states. For generic M we have

$$\begin{aligned} f(M) &= \lambda^{2M} + (1-\lambda)^{2M} - \frac{1}{2}\left(\lambda^M + (1-\lambda)^M\right)^2 && (206)\\ &= \frac{1}{2}((1-\lambda)^M - \lambda^M)^2 && (207)\\ &= \frac{1}{2^{2M+1}}\left[(1+\sqrt{2P-1})^M - (1-\sqrt{2P-1})^M\right]^2\,, && (208) \end{aligned}$$

where in the final equality we have used the fact that for single-qubit states $\lambda = \frac{1}{2}(1 - \sqrt{2P-1})$. Further, using $f(1) = P - \frac{1}{2}$ we have the bound as required.

Now let us consider the setting of $n \geq 2$. We will construct two bounds, for the respective high purity and low purity limits. We start with the bound for high purity states. We can write the right hand side of Eq. (205) explicitly as

$$\begin{aligned} \frac{\mathrm{Var}[\boldsymbol{\lambda}^{(M)}]}{\mathrm{Var}[\boldsymbol{\lambda}^{(1)}]} &= \frac{\frac{1}{2^n}\sum_i \lambda_i^{2M} - (\frac{1}{2^n}\sum_i \lambda_i^M)^2}{\frac{1}{2^n}\sum_i \lambda_i^2 - \frac{1}{2^{2n}}} && (209)\\ &= \frac{\frac{1}{2^n}\sum_i \lambda_i^{2M} - \frac{1}{2^{2n}}\sum_i \lambda_i^{2M} - \frac{1}{2^{2n}}\sum_{i\neq j}\lambda_i^M\lambda_j^M}{\frac{1}{2^n}\sum_i \lambda_i^2 - \frac{1}{2^{2n}}} && (210)\\ &\leq \frac{(2^n-1)(\sum_i \lambda_i^{2M})}{2^n\sum_i \lambda_i^2 - 1} && (211)\\ &\leq \frac{(2^n-1)(\sum_i \lambda_i^2)^M}{2^n\sum_i \lambda_i^2 - 1} && (212)\\ &= \frac{P^M}{P - \frac{1}{2^n}}\left(1 - \frac{1}{2^n}\right)\,, && (213) \end{aligned}$$

where in order to obtain the first inequality we have dropped the cross terms $\frac{1}{2^{2n}}\sum_{i\neq j}\lambda_i^M\lambda_j^M$, and in the second inequality we have introduced new cross terms. The final equality comes by substituting in the definition of the purity P. We note this first bound is upper-bounded by 1 for all $P \geq \frac{1}{2^n-1}$. Thus, we seek a tighter bound for $P \leq \frac{1}{2^n-1}$.

We can now construct our second bound for strongly mixed states (those states with purity close to $1/2^n$). We will consider bounds on $\mathrm{Var}[X^M]$ where a random variable X when it is known that it takes values close to its mean μ. We consider the decomposition

$$\begin{aligned} X^M &= ((X-\mu)-\mu)^M && (214)\\ &= \mu^M + \sum_{k=1}^{M} Y_k && (215) \end{aligned}$$

where we have defined the random variables $Y_k = \binom{M}{k}\mu^{M-k}(X-\mu)^k$. Further, we can write

$$\begin{aligned} \mathrm{Var}[X^M] &= \mathrm{Var}\Big[\sum_{k=1}^{M} Y_k\Big] && (216)\\ &= \mathbb{E}\left[\Big(\sum_k Y_k - \mathbb{E}\Big[\sum_k Y_k\Big]\Big)\Big(\sum_j Y_j - \mathbb{E}\Big[\sum_j Y_j\Big]\Big)\right] && (217)\\ &= \sum_{k,j} \mathbb{E}\left[\Big(Y_k - \mathbb{E}[Y_k]\Big)\Big(Y_j - \mathbb{E}[Y_j]\Big)\right] && (218)\\ &= \sum_{k,j} \mathrm{Cov}\left[Y_k, Y_j\right] && (219)\\ &\leq \sum_{k,j} \sqrt{\mathrm{Var}\left[Y_k\right]\mathrm{Var}\left[Y_j\right]}\,, && (220) \end{aligned}$$

where the inequality is due to Cauchy-Schwarz. We now take X to be the random variable which takes values $\{\lambda_i\}_i$ with uniform probability and mean $\mu = \frac{1}{2^n}$. We will bound $\mathrm{Var}\left[Y_k\right]$ under the assumption that $\{\lambda_i\}_i$ are close in value to the maximally mixed value $\frac{1}{2^n}$.

First, we note that each Y_k is a function of $(X-\mu)^k$, and so we must investigate the shifted spectrum which we denote $\hat{\boldsymbol{\lambda}}$ where $\hat{\lambda}_i = \lambda_i - \frac{1}{2^n}$ for all i. Using Popoviciu's inequality, we have the bound

$$\mathrm{Var}\left[(X-\mu)^k\right] \leq \frac{1}{4}\left(\hat{\lambda}^k_{max} - \hat{\lambda}^k_{min}\right)^2 . \tag{221}$$

Now suppose that we have the constraint

$$\lambda_{max} - \lambda_{min} = 2b \tag{222}$$

for some $b \geq 0$. For any k, we have

$$\hat{\lambda}^k_{max} - \hat{\lambda}^k_{min} \leq \left|\hat{\lambda}_{max}\right|^k + \left|\hat{\lambda}_{min}\right|^k . \tag{223}$$

Let us now bound the quantity on the right by considering its maximum value over all spectra with constraint (222). The quantity on the right is maximized by the choice of vector $(|\hat{\lambda}_{max}|, |\hat{\lambda}_{min}|)$ that majorizes all others, given some fixed value of $|\hat{\lambda}_{max}| + |\hat{\lambda}_{min}|$. Indeed, $|\hat{\lambda}_{max}| + |\hat{\lambda}_{min}| = b$ is fixed by our constraint (222) ($\hat{\lambda}_{min}$ must be negative in order to preserve trace). Thus the quantity on the right hand side of (223) can be bounded by maximizing $\hat{\lambda}_{max}$ and minimizing $|\hat{\lambda}_{min}|$. This is achieved by setting all other $\hat{\lambda}_i$ equal to $\hat{\lambda}_{min}$. We then have pair of constraints

$$\begin{aligned} \hat{\lambda}_{max} + (2^n - 1)\hat{\lambda}_{min} &= 0\,, && (224)\\ \hat{\lambda}_{max} - \hat{\lambda}_{min} &= 2b\,, && (225) \end{aligned}$$

where the first constraint comes from preservation of trace, and the second is our original constraint. This is a linear system of equations with solution

$$\hat{\lambda}^*_{max} = 2b\frac{2^n-1}{2^n}\,, \quad \hat{\lambda}^*_{min} = -2b\frac{1}{2^n} \tag{226}$$

substituting these values into (223) we have the bound

$$\begin{aligned}\hat{\lambda}^k_{max} - \hat{\lambda}^k_{min} &\leq (2b)^k\left(\left(\frac{2^n-1}{2^n}\right)^k + \left(\frac{1}{2^n}\right)^k\right) &(227)\\ &\leq (2b)^k\,. &(228)\end{aligned}$$

We will find it necessary to use the tighter bound (227) in the case of $M = 2$, but the looser bound (228) will enable us to write a bound with a more compact form for $M \geq 3$.

We now relate b to the purity. We can write a general spectrum that satisfies the constraint in (222) as $\boldsymbol{\lambda}_{b,c,\boldsymbol{a}} = (\frac{1}{2^n} + b + c,\, \frac{1}{2^n} - b + c,\, \frac{1}{2^n} - a_1,\, ...,\, \frac{1}{2^n} - a_{d-2})$, for some c and set $\{a_i\}_i$ that satisfy $\sum_i^{d-2} a_i = 2c$ (in order to preserve trace). The purity that corresponds to this spectrum is given by

$$\begin{aligned}P(\boldsymbol{\lambda}_{b,c,\boldsymbol{a}}) &= \left(\frac{1}{2^n} + b + c\right)^2 + \left(\frac{1}{2^n} - b + c\right)^2 + \sum_{i=1}^{d-2}\left(\frac{1}{2^n} - a_i\right)^2 &(229)\\ &= \frac{1}{2^n} + 2b^2 + c^2 + \sum_i a_i^2 + \frac{2}{2^n}\left[2c - \sum_{i=1}^{d-2} a_i\right] &(230)\\ &\geq \frac{1}{2^n} + 2b^2\,. &(231)\end{aligned}$$

Moreover, this purity bound is achievable by the spectrum $\boldsymbol{\lambda}_{b,0,\boldsymbol{0}} = (\frac{1}{2^n} + b, \frac{1}{2^n} - b, \frac{1}{2^n}, ..., \frac{1}{2^n})$ if we have $b \leq \frac{1}{2^n}$. We conclude that for any spectrum $\boldsymbol{\lambda}_b$ that satisfies the constraint (222), we have

$$b \leq \sqrt{\frac{1}{2}\left(P(\boldsymbol{\lambda}_b) - \frac{1}{2^n}\right)}\,. \tag{232}$$

And we now have all the tools to bound $\mathrm{Var}[Y_k]$ for all k and subsequently $\mathrm{Var}[X^M]$

By combining the bounds (221) and (227) we have

$$\mathrm{Var}\left[(X-\mu)^k\right] \leq \frac{1}{4}(2b)^{2k} g_k^2 \tag{233}$$

where we have denoted $g_k = \left(\frac{2^n-1}{2^n}\right)^k + \left(\frac{1}{2^n}\right)^k \leq 1$. This allows us to bound $\mathrm{Var}[Y_k]$ by writing

$$\begin{aligned}\sqrt{\mathrm{Var}[Y_k]} &= \binom{M}{k}\mu^{M-k}\sqrt{\mathrm{Var}\left[(X-\mu)^k\right]} &(234)\\ &\leq \frac{1}{2}\binom{M}{k}\mu^{M-k}(2b)^k g_k\,. &(235)\end{aligned}$$

We first pursue a bound for general $M \in \mathbb{N}$ and replace each g_k with 1. We observe that the quantities $\{\binom{M}{k}\mu^{M-k}(2b)^k\}_{k=1}^M$ are simply the terms in the expansion of $(2b-\mu)^M - \mu^M$, that is,

$$\sum_k \sqrt{\mathrm{Var}[Y_k]} \leq \frac{1}{2}\left((2b-\mu)^M - \mu^M\right)\,. \tag{236}$$

Returning to (220), we have

$$\begin{aligned}\mathrm{Var}[X^M] &\leq \frac{1}{4}\left((2b-\mu)^M - \mu^M\right)^2 &(237)\\ &\leq \frac{1}{4}\left[\left(2\sqrt{\frac{1}{2}\left(P - \frac{1}{2^n}\right)} - \mu\right)^M - \mu^M\right]^2 &(238)\end{aligned}$$

where in order to obtain the second inequality we have used (232) to substitute b with its bound in terms of the purity. We further note that $\mathrm{Var}[X]=\frac{1}{2^n}(P-\frac{1}{2^n})$, and dividing the two quantities we obtain

$$\overline{\overline{\chi}} \leq \frac{2^n}{4}\frac{\left[\left(\sqrt{2\left(P-\frac{1}{2^n}\right)}+\frac{1}{2^n}\right)^M-\left(\frac{1}{2^n}\right)^M\right]^2}{P-\frac{1}{2^n}} \tag{239}$$

as required. To summarize, combining the two bounds for high purity and low purity, so far we have

$$G'(n\geq 2, M\geq 2, P)=\min\left(\frac{2^n}{4}\frac{\left[\left(\sqrt{2\left(P-\frac{1}{2^n}\right)}+\frac{1}{2^n}\right)^M-\left(\frac{1}{2^n}\right)^M\right]^2}{P-\frac{1}{2^n}},\ \frac{P^M}{P-\frac{1}{2^n}}\left(1-\frac{1}{2^n}\right)\right). \tag{240}$$

Now we discuss the magnitude of our bound obtained thus far, as well as its monotonicity with respect to M. In particular, we will show that its value can exceed 1 for $M=2$, and so we will pursue a tighter bound for $M=2$. We can evaluate $G'(n\geq 2, M\geq 2, P)$ explicitly at $P=\frac{1}{2^n-1}$ as

$$G'\Big(n\geq 2, M\geq 2, P=\frac{1}{2^n-1}\Big)=\min\left(\frac{2^{2n}(2^n-1)}{4}\left[\left(\sqrt{\frac{2}{2^n(2^n-1)}}+\frac{1}{2^n}\right)^M-\left(\frac{1}{2^n}\right)^M\right]^2,\ \frac{(2^n-1)^2}{(2^n-1)^M}\right). \tag{241}$$

Firstly, by inspection this is a decreasing function in n for all $M\geq 2$, so in order to bound its magnitude we can consider $n=2$. At $M=2$ we have

$$G'\Big(2,2,\frac{1}{2^2-1}\Big)=\min\left(1,\ \frac{5+2\sqrt{6}}{6}\right)=1\,, \tag{242}$$

where we note $\frac{5+2\sqrt{6}}{6}\geq 1$. As the first function in the minimization of (240) has negative gradient for $P<\frac{1}{2^n-1}$ for $n\geq 2, M=2$, this implies that there exists a set of values $P=\frac{1}{2^n-1}-\delta$, where $\delta>0$ is small, such that the first function has value exceeding 1. The second function also has value exceeding 1 in such a region as it is continuous. Thus, there exist values of P for which the bound $G'>1$ at $M=2$. Moving on to $M=3$, we can numerically verify that $G'\Big(2,3,\frac{1}{2^n-1}\Big)\leq 1$ with both functions in the minimization having value below 1. As functions of the form $f(x)=a^x-b^x$ where $b\leq a\leq 1$ only have one stationary point which is a maximum, this implies that $G'\Big(2,M,\frac{1}{2^n-1}\Big)$ is decreasing for all $M\geq 3$ and thus $G'\Big(2,M\geq 3,\frac{1}{2^n-1}\Big)\leq 1$.

We will replace $G'\Big(n\geq 2,2,P\Big)$ with a tighter bound that is less than 1 for all $n\geq 2$. We return to (234) and now explicitly consider the g_k terms. Substituting this into (220) for $M=2$ we have

$$\begin{aligned}
\mathrm{Var}[X^2] &\leq \mathrm{Var}[Y_1]+\mathrm{Var}[Y_2]+2\sqrt{\mathrm{Var}[Y_1]\mathrm{Var}[Y_2]} && (243)\\
&= (2\mu)^2\mathrm{Var}[X-\mu]+\mathrm{Var}\left[(X-\mu)^2\right]+4\mu\sqrt{\mathrm{Var}\left[(X-\mu)^2\right]\mathrm{Var}[X-\mu]} && (244)\\
&\leq (2\mu)^2\mathrm{Var}[X]+\frac{1}{4}(2b)^4g_2^2+4\mu\sqrt{\frac{1}{4}(2b)^4g_2^2\mathrm{Var}[X]} && (245)\\
&\leq \frac{4}{2^{2n}}\mathrm{Var}[X]+\left(P-\frac{1}{2^n}\right)^2g_2^2+\frac{4}{2^n}g_2\left(P-\frac{1}{2^n}\right)\sqrt{\mathrm{Var}[X]} && (246)
\end{aligned}$$

where in the first equality we use the definition of Y_k for $M=2$, in the first inequality we use (233) along with the fact that $g_1=1$, and in the final inequality we use (232). Finally, dividing by $\mathrm{Var}[X]$

we have

$$\overline{\overline{\chi}}(M=2) \leq \frac{4}{2^{2n}} + \frac{1}{\mathrm{Var}[X]}\left(P - \frac{1}{2^n}\right)^2 g_2^2 + \frac{4}{2^n} g_2 \sqrt{P - \frac{1}{2^n}}\sqrt{\frac{1}{\mathrm{Var}[X]}} \tag{247}$$

$$= \frac{4}{2^{2n}} + 2^n g_2^2 \left(P - \frac{1}{2^n}\right) + \frac{4}{2^{n/2}} g_2 \sqrt{P - \frac{1}{2^n}}\,, \tag{248}$$

where we have used $\mathrm{Var}[X] = \frac{1}{2^n}(P - \frac{1}{2^n})$. □

We note in the following remark that outside of the low purity regime, as purity decreases, our bounds monotonically decrease.

Remark 1. *The bounds in Proposition 4 are monotonically increasing with purity P for all $P \geq \frac{1}{2^n}\max(\frac{2^n}{2^n-1}, \frac{M}{M-1})$.*

Proof. For $n \geq 2$, this can be seen by inspecting the partial derivative of the high purity bound, which is given by

$$\frac{\partial G(n \geq 2, M, P \geq \frac{1}{2^n-1})}{\partial P} = \frac{(M+1)P^M - \frac{1}{2^n} M P^{M-1}}{(P - \frac{1}{2^n})^2}\left(1 - \frac{1}{2^n}\right) \tag{249}$$

which is positive for all $P > \frac{M}{M-1}\frac{1}{2^n}$. Similarly, for $n = 1$ the bound can be shown to be monotonically increasing in P for all physically allowable values of P. We note that the bound for $n = 1$ satisfies

$$\sqrt{2^{2M} G(P, n=1, M)} = \frac{(1+x)^M - (1-x)^M}{x} \tag{250}$$

where we have denoted $x = \sqrt{2P - 1}$. The derivative of the numerator with respect to x takes the value

$$\frac{d\left((1+x)^M - (1-x)^M\right)}{dx} = M\left((1+x)^{M-1} + (1-x)^{M-1}\right) \tag{251}$$

$$\geq M(1 + (M-1)x + 1 - (M-1)x) \tag{252}$$

$$= 2M \tag{253}$$

where in order to obtain the inequality we have used the standard inequality $(1+x)^n \geq 1 + nx$, $\forall x \geq -1, n > 1$. Thus, as we only consider $M \geq 2$, the numerator of (250) increases at a faster rate than the denominator. Moreover, the second derivative of the numerator is positive, and both the numerator and denominator of (250) take value 0 at $x = 0$ ($P = 1/2^n$). Thus, (250) is an increasing function in the purity P. □

We plot the bounds obtained in Proposition 4 on the 2-design-averaged resolvability in Fig. 7. First, in the left figure we plot the intermediate bound $\frac{\mathrm{Var}[\boldsymbol{\lambda}^{(M)}]}{\mathrm{Var}[\boldsymbol{\lambda}^{(1)}]}$ in (205) for states with 100 randomly generated spectra for increasing number of qubits n and number of state copies M. We see that all values lie below 1. Moreover, this plot visualizes the exponential scaling with M for fixed n and we observe that broadly, the bound is decreasing with increasing number of qubits n for fixed M. Further, as expected, the bound is always less than or equal to 1. Second, in order to demonstrate the behaviour of our final upper bound (192) we plot increasing number of state copies M ranging from 2 to 4 for $n = 2$. For each M, we randomly generate 10000 states and plot $\frac{\mathrm{Var}[\boldsymbol{\lambda}^{(M)}]}{\mathrm{Var}[\boldsymbol{\lambda}^{(1)}]}$ against the purity of the state as separate points. The final upper bound is then plotted as a line for each value of M.

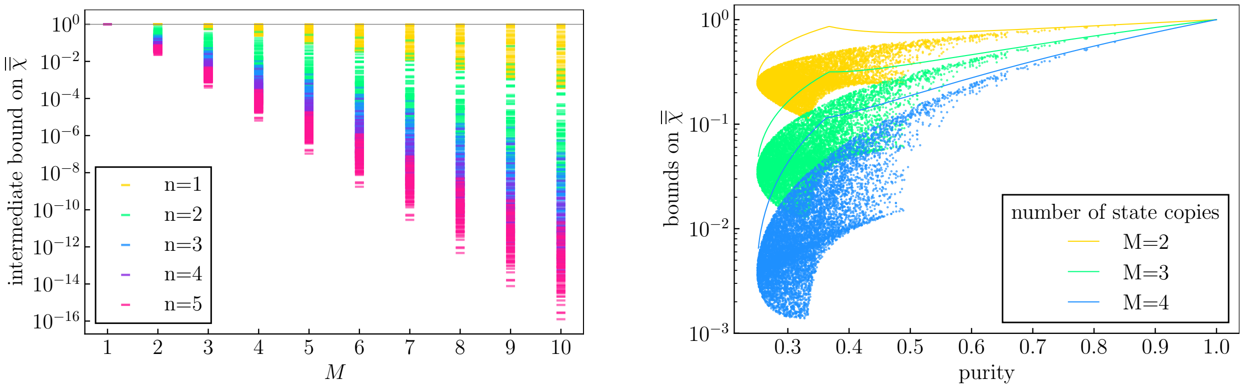

Figure 7: **Bounds on $\overline{\overline{\chi}}$ for VD.** (Left): We plot the intermediate upper bound $\frac{\mathrm{Var}[\boldsymbol{\lambda}^{(M)}]}{\mathrm{Var}[\boldsymbol{\lambda}^{(1)}]}$ for randomly generated states with increasing number of qubits n and number of state copies M. (Right): We plot as points the intermediate upper bound $\frac{\mathrm{Var}[\boldsymbol{\lambda}^{(M)}]}{\mathrm{Var}[\boldsymbol{\lambda}^{(1)}]}$ against purity for randomly generated states at different values of M at $n = 2$. We also plot the final bound (192), which is a function of purity, as a line.

E.3 Probabilistic Error Cancellation

E.3.1 Error mitigation of multiple gates

We first consider the error mitigation cost of mitigating multiple noise channels. Suppose we have two noisy gates which we represent as the channel

$$\mathcal{N}' \circ \mathcal{U}' \circ \mathcal{N} \circ \mathcal{U} \tag{254}$$

where $\{\mathcal{U}',\mathcal{U}\}$ are channels that represent the ideal gates and $\{\mathcal{N}',\mathcal{N}\}$ are noise channels. Note that this framework also includes as a special case the scenario where two gates act in parallel on different subsystems. Given a set of basis gates $\{\mathcal{B}_\alpha\}_\alpha$, we can construct a quasiprobability distribution for the ideal channel as

$$\mathcal{U}' \circ \mathcal{U} = \sum_{\alpha,\beta} k_\alpha k_\beta \, \mathcal{B}_\alpha \circ \mathcal{N}' \circ \mathcal{U}' \circ \mathcal{B}_\beta \circ \mathcal{N} \circ \mathcal{U} \,. \tag{255}$$

where we have used (58). From (255) we see that the error mitigation cost is

$$\gamma_{tot} = \sum_{\alpha,\beta} k_\alpha^2 k_\beta^2 = \gamma\gamma' \tag{256}$$

where γ, γ' are the individual error mitigation costs for each gate. We can see the above reasoning can be extended inductively to show that the error mitgation cost of a collection of gates with the probabilistic error cancellation is equal to the product of the individual error mitigation costs.

E.3.2 Global depolarizing noise

Proposition 5 (Relative resolvability of Probabilistic Error Cancellation for global depolarizing noise)**.** *Consider a quasi-probability method that corrects global depolarizing noise of the form* (34)*. For any pair of states corresponding to points on the cost function landscape, the optimal quasiprobability scheme gives*

$$\chi_{depol} = \frac{2^{2n}}{2^{2n} - p(2-p)} \geq 1 \,, \tag{257}$$

for all $n \geq 1$, $p \in [0,1]$, which is achieved with access to noisy Pauli gates.

Proof. Ref. [94] gives the optimal quasi-probability decomposition for the inverse noise channel as

$$\mathcal{D}^{-1} = \left(1 + \frac{(2^{2n}-1)p}{2^{2n}(1-p)}\right)\mathcal{I} - \sum_{i=1}^{2^{2n}-1} \frac{p}{2^{2n}(1-p)}\mathcal{P}_i \,, \tag{258}$$

where $\mathcal{I}$ is the identity channel and $\mathcal{P}_i$ is the Pauli channel corresponding to the ith Pauli tensor product. This has corresponding error mitigation cost

$$\gamma = \frac{2^{2n} - 2p + p^2}{2^{2n}(1-p)^2}. \tag{259}$$

Assuming perfect correction we have $\Delta\widetilde{C} = (1-p)\Delta C$ which implies

$$\chi_{depol} = \frac{2^{2n}}{2^{2n} - 2p + p^2}\,, \tag{260}$$

which is greater than or equal to 1 as $-2p + p^2 \leq 0$ for all $0 \leq p \leq 1$. □

E.3.3 Local depolarizing noise

Here we consider a model of cost concentration due to a single instance of local depolarizing noise in a circuit. We presume that the concentration follows a similar form of scaling to global depolarizing noise and a tensor product of local depolarizing noise (see Eq. (6)). We show that, under this assumption, the relative resolvability has regimes of being greater than 1 or less than 1, depending on the strength of the cost concentration.

Supplemental Proposition 4 (Relative resolvability of Probabilistic Error Cancellation with one instance of local depolarizing noise)**.** *Consider a single instance of local depolarizing noise occurring with error probability p acting at an arbitrary point in the parameterized circuit. Suppose that due to this noise channel we have*

$$\langle \Delta\widetilde{C}(\boldsymbol{\theta}_{i,*})\rangle_i \geq (1 - b_\alpha p)\langle \Delta C(\boldsymbol{\theta}_{i,*})\rangle_i \tag{261}$$

for all $p \in [0,1]$ where $\langle\cdot\rangle_i$ denotes an average over all avaliable parameters and b_α where is some positive constant. Then the optimal quasiprobability scheme gives:

- *for $b_\alpha \leq \frac{3}{4}$,*

$$\overline{\chi} \leq 1\,, \quad \forall p \in [0,1], \tag{262}$$

- *for $\frac{3}{4} < b_\alpha \leq 1$,*

$$\overline{\chi} \leq 1 + \frac{1}{4}p(2-p) + \mathcal{O}(p^2)\,, \quad \forall p \in [0,1]\,, \tag{263}$$

$$\overline{\chi} > 1\,, \quad \forall p \in \left(0,\, 1 - \sqrt[3]{3(b^{-1} - 1)}\right]\,, \tag{264}$$

- *for $b_\alpha > 1$,*

$$\overline{\chi} > 1 + \frac{p(2-p)}{4 - p(2-p)}\,, \quad \forall p \in (0,\, 1/b_\alpha]\,. \tag{265}$$

Proof. From Eq. (259), we can write the optimal error mitigation cost for one instance of local depolarizing noise acting on one qubit as

$$\gamma = \frac{4 - 2p + p^2}{4(1-p)^2}\,. \tag{266}$$

Now, due to our assumption (261) and assuming perfect implementation of the basis of noisy gates (leading to perfect correction of the noise) we have

$$\overline{\chi} \leq \frac{4(1-p)^2}{(4 - 2p + p^2)(1 - bp)^2}\,, \tag{267}$$

and we denote the quantity on the right hand side as $h(p)$. Note that for any value of b, $h(p=0)=1$ and $h(p=1)=0$. The partial derivative can be found to satisfy

$$\frac{\partial h}{\partial p} \propto (1-p)\,(1-bp)\left(-p^3+3p^2-3p+\frac{1}{b}-4\left(\frac{1}{b}-1\right)\right)\,, \tag{268}$$

where the proportionality factor we omit is positive for all $b \geq 0$ and $p \in [0,1]$. The third bracket is a cubic form with discriminant

$$\Delta = \frac{108}{b^2}\left(-8b^2+11b-4\right)\,, \tag{269}$$

which is strictly negative for all b. Thus, the cubic form only has one real root and, inspecting its behaviour for large p, we can conclude it has negative gradient for all p. The cubic form has root at $p=0$ when $b=3/4$. More generally, the root can be found to take the form

$$p' = 1+\sqrt[3]{3(1-b^{-1})}\,. \tag{270}$$

By evaluating the second derivative of $h(p)$, this root can be seen to correspond to a local maximum of $h(p)$. We now find the maximum value of $h(p)$ over the interval $p \in [0,1]$ for different regimes of cost concentration strength b.

First, we inspect the regime where $b \leq 3/4$. In this case $p' \leq 0$ and thus $\frac{\partial h}{\partial p} \leq 0$ for $p \in [0,1]$. Thus the maximum value of $h(p)$ on the interval $p \in [0,1]$ is $h(0)=1$. We can then conclude that $\overline{\chi} \leq 1$ with bound saturated in the limit of zero error probability.

Now consider the regime $3/4 < b \leq 1$. In this case $0 < p' \leq 1$ and $\frac{\partial h}{\partial p} > 0$ for small values of p. Specifically, it is clear that $\overline{\chi} > 1$ for $0 < p \leq 1+\sqrt[3]{3(1-b^{-1})}$. The upper limit on p can be raised, however, the exact interval is obtained by solving a quartic equation which we omit here as it is not very insightful. Moreover, the upper limit is tight in the limit $b \to 1$ and we obtain the result that when $b=1$, $\overline{\chi} > 0$ for all $p \in (0,1)$.

Finally, consider the regime $b \geq 1$. Now $\frac{\partial h}{\partial p}$ has a different root $p'' = 1/b$ due to the second bracket in (268). Again, this can be shown to correspond to a maximum of $\overline{\chi}$ and we can write $\overline{\chi} > 1$ for $0 < p \leq 1/b$. □

Proposition 6 (Scaling of Probabilistic Error Cancellation with local depolarizing noise)**.** *Consider tensor-product local depolarizing noise with local depolarizing probability p acting in L instances through a depth L circuit as in Eq. (3). Suppose that the effect of this noise is to cause cost concentration*

$$\langle \Delta\widetilde{C}(\boldsymbol{\theta}_{i,*})\rangle_i = Aq^L \langle \Delta C(\boldsymbol{\theta}_{i,*})\rangle_i\,, \tag{271}$$

for some constant A and noise parameter $q \in [0,1)$. The optimal quasiprobability method to mitigate the depolarizing noise in the circuit yields

$$\overline{\chi} = \frac{1}{A^2 q^{2L}}\left(Q(p)\right)^{nL}\,, \tag{272}$$

where $0 \leq Q(p) \leq 1$ for all p. Thus, the average relative resolvability has unfavourable scaling with system size.

Proof. As shown in Section E.3.1, error mitigation cost of multiple gates with probabilistic error cancellation is the product of the individual error mitigation costs. Thus, for the collection of gates considered, we have total error mitigation cost

$$\gamma_{tot} = \left(\frac{4(1-p)^2}{4-2p+p^2}\right)^{nL}\,, \tag{273}$$

where we have used Eq. (266). We suppose that mitigation perfectly corrects the error, such that $\Delta C_m(\boldsymbol{\theta}_{i,*}) = \Delta C(\boldsymbol{\theta}_{i,*})$. Combining this with our assumption (271) we obtain the desired result, where we denote

$$Q(p) = \frac{4-2p+p^2}{4(1-p)^2} = 1-\frac{3p(2-p)}{4-p(2-p)}\,, \tag{274}$$

which clearly satisfies $0 \leq Q(p) \leq 1$. □

E.4 Linear Ansatz Methods

E.4.1 Global depolarizing noise is exactly correctable

Consider the linear ansatz

$$C_m(\boldsymbol{a}) = a_1\widetilde{C} + a_2\,, \tag{275}$$

where we denote $\boldsymbol{a} = (a_1, a_2)$. As shown in [45], this ansatz is particularly suited to global depolarizing noise and the ansatz can correct the noise exactly. Namely, the n-qubit noise channel

$$\rho \xrightarrow{\mathcal{D}} (1-p)^L\rho + (1-(1-p)^L)\frac{\mathbb{1}}{2^n} \tag{276}$$

can be exactly corrected by using

$$a_1 = \frac{1}{(1-p)^L}\,, \quad a_2 = -\frac{(1-(1-p)^L)}{(1-p)^L}\mathrm{Tr}[O]/2^n\,. \tag{277}$$

As correction is exact, $\Delta C_m = \Delta C$. It can also be seen that $\Delta\widetilde{C} = (1-p)^L\Delta C$ and the error mitigation cost is $\gamma = 1/(1-p)^{2L}$. This gives $\chi = 1$ for any pair of cost function points.

Note in this discussion we have neglected the shot burden of training. Whilst this may be significant and difficult to quantify for other noise channels, in the case of global depolarizing noise this is minimal as only two training data points are required and the ansatz is universal for any state.

E.4.2 Relative resolvability between two points with same ansatz applied

Proposition 7 (Linear ansatz methods)**.** *Consider any error mitigation strategy that mitigates noisy cost function value $\widetilde{C}(\boldsymbol{\theta})$ by constructing an estimator $C_m(\boldsymbol{\theta})$ of the form (16). For any two noisy cost function points to which the same ansatz is applied, we have*

$$\chi = 1\,, \tag{278}$$

for any noise process.

Proof. By applying the same ansatz of the form (16) to two noisy cost function points corresponding to parameter sets $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, one can write

$$C_m(\boldsymbol{\theta}_1, \boldsymbol{a}) = a_1\widetilde{C}(\boldsymbol{\theta}_1) + a_2\,, \tag{279}$$

$$C_m(\boldsymbol{\theta}_2, \boldsymbol{a}) = a_1\widetilde{C}(\boldsymbol{\theta}_2) + a_2\,. \tag{280}$$

This gives $\gamma = a_1$ and $\Delta C_m = a_1\Delta\widetilde{C}$. Thus, substituting these quantities into Definition 2 one obtains $\chi = 1$ as required. □

F Numerical simulations - implementation details

We perform our optimizations using the MATLAB implementation of the Nelder-Mead algorithm [105]. For each MaxCut graph, we perform optimization independently for N_i random choices of an initial simplex. We evaluate the cost function by performing perfect sampling of the simulated state with N_s shots. After each iteration of the Nelder-Mead algorithm, we compute the total cost of optimization per graph N_{tot} by summing the shot budget spent for all N_i instances of the optimization. To analyze the convergence of results with N_{tot}, as shown in Figs. 5, 8, we take the optimization results after N_{tot} shots to be the best of N_i instances according to the optimized cost function. The optimization is terminated for $n = 5$ ($n = 8$) when N_{tot} exceeds 1.5×10^8 (7×10^7).

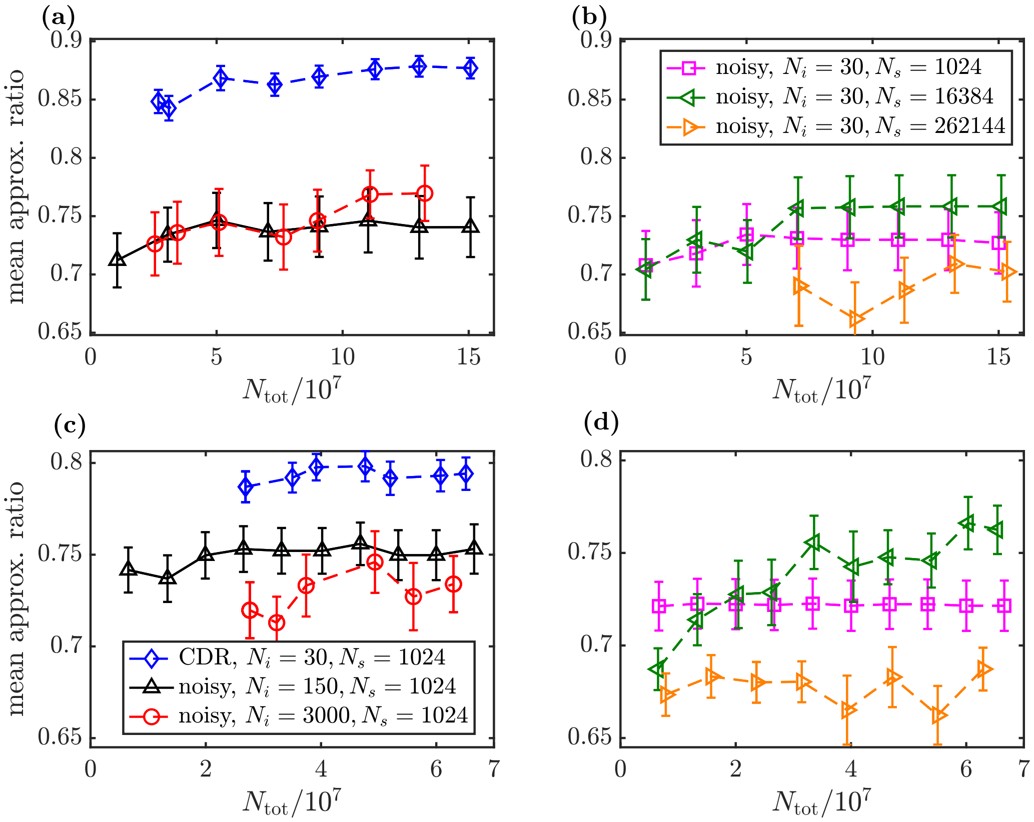

Figure 8: **Benchmarking various implementations of the noisy optimization for 5-qubit and 8-qubit MaxCut QAOA with $p = 4$.** In **(a,b)** we plot the approximation ratio averaged over 36 $n = 5$ Max-Cut graphs chosen randomly from the Erdös-Rényi ensemble as a function of N_{tot}. The error bars are computed as in Fig. 5. In **(c,d)** we show in a similar manner the results for 30 $n = 8$ random Max-Cut Erdös-Rényi graphs. We compare the results for various numbers N_i of optimization instances initialized randomly and various numbers of shots N_s per cost function evaluation. As a reference, we show the results of CDR optimization for $p = 4$. For the 5-qubit (8-qubit) case we have $N_{\text{tot}} = 10^7$ to 1.5×10^8 ($N_{\text{tot}} = 1 \times 10^7$ to 7×10^7) as in Fig. 5 (Fig. 6). Additionally, as in Fig. 5 we use the approximation ratio computed with the exact energy to benchmark the optimization, and in the case of $N_i > 1$ we choose as the result of optimization the best instance determined according to the optimized cost function. The error bars are computed as in Fig. 5. We consider various values of $N_s = 1024,\ 16384,\ 262144$ for $N_i = 30$ and various values of $N_i = 30,\ 150,\ 3000$ for $N_s = 1024$. For the 5-qubit case, we find that $N_i = 3000, N_s = 1024$ yields the best results although differences in quality between most of the noisy optimization implementations are relatively small in comparison to the CDR mitigated optimization. In the case of $n = 8$ the best noisy results are obtained for $N_i = 30, N_s = 16384$, but again different choices of N_i and N_s lead to similar quality of the solutions.

F.1 CDR-mitigated optimization

We perform CDR-mitigated optimization with $N_i = 30$ and $N_s = 1024$. We use training circuits constructed with a non-Clifford gates projection algorithm of [82]. To construct the training circuits we decompose $e^{i\gamma_j H_{\text{MaxCut}}}$ to native gates of an IBM quantum computer using a decomposition from [106]. In order to account for linear connectivity of the simulated device we use SWAP gates to implement $e^{-i\gamma_j Z_k Z_l}$ for non nearest-neighbors terms. The training circuits contain 100 near-Clifford circuits with at most 30 non-Clifford gates. In the case of circuits with fewer than 60 non-Clifford gates, we construct training circuits with half of the non-Clifford gates replaced by Clifford gates. We evaluate the cost function for the training circuits using perfect sampling and $N_s = 1024$ shots. We perform CDR mitigation for each 2-body term of H_{MaxCut} independently. In general, in order to maximize the quality of the mitigation one should construct the training circuits independently for each new set of QAOA angles. Here, for the sake of shot efficiency, for each new set of parameters we compute the training set from scratch only if the 1-norm distance of its QAOA angles $(\gamma_1, \beta_1, \gamma_2, \beta_2, \dots \gamma_p, \beta_p)$ from the closest point of a simplex is larger than 0.01. Otherwise, we use the CDR linear ansatz for the

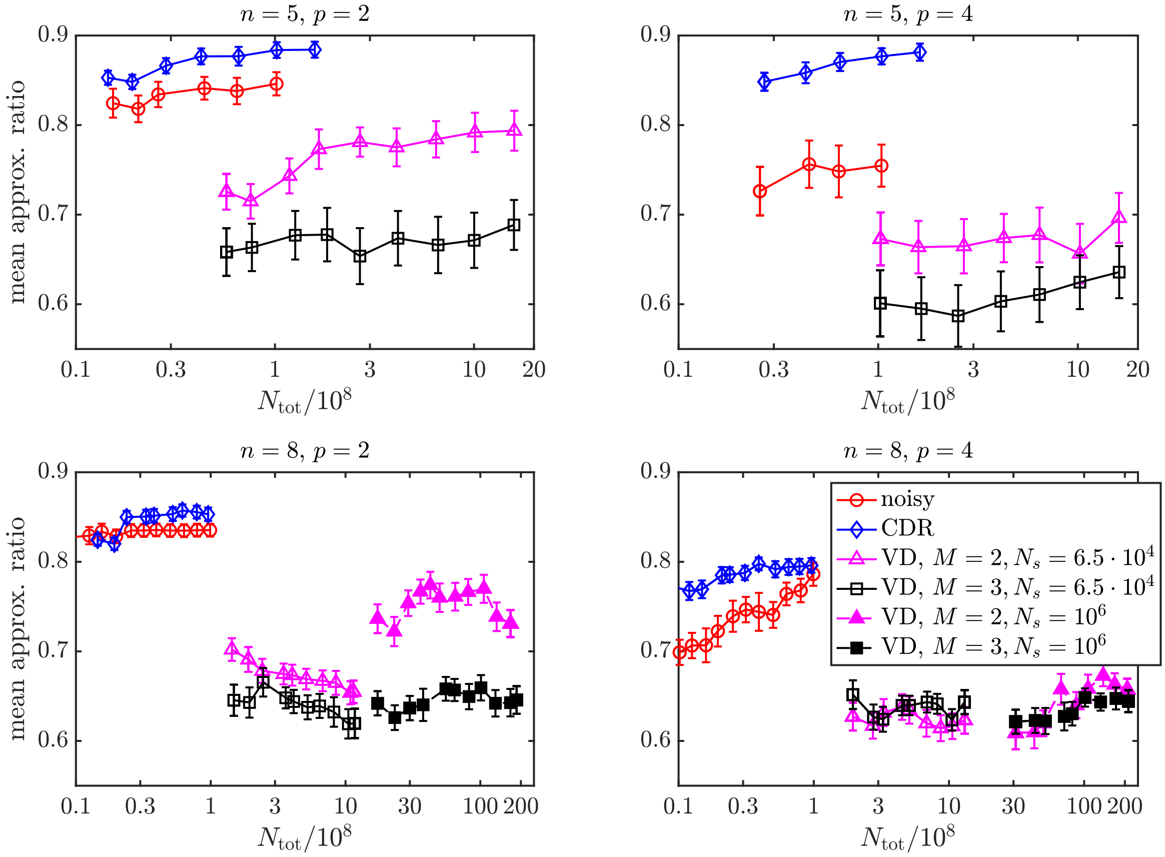

Figure 9: **Virtual Distillation mitigated optimization for $p = 2, 4$, 5-qubit and 8-qubit MaxCut QAOA.** We plot the approximation ratio averaged over instances of Max-Cut graphs randomly chosen from the Erdös-Rényi ensemble as a function of total shot number N_{tot}. For $n = 5$ ($n = 8$) we choose 36 (30) graphs. The results were obtained with $N_i = 30$ initializations and $\widetilde{N}_s = 65536$ shots per $\mathrm{Tr}[\widetilde{\rho}^M Z_i Z_j]$ and $\mathrm{Tr}[\widetilde{\rho}^M]$ estimation. For larger $n = 8$ graphs we also show results obtained with $\widetilde{N}_s = 10^6$, $N_i = 30$. For reference we also present our results of CDR-mitigated and noisy optimization from Figs. 5 and 6. The error bars are computed as described in the caption of Fig. 5. We observe that for this setting the optimization with Virtual Distillation does not outperform the noisy or CDR-mitigated optimization.

closest point of the simplex.

For the noisy (unmitigated) optimization we benchmark various combinations of N_i and N_s values for $n = 5$, $p = 4$. In particular we consider increasing N_s for $N_i = 30$ and increasing N_i for $N_s = 1024$. We gather the results in Fig. 8. We find that for $n = 5$ ($n = 8$) while using N_{tot} ranging from 10^7 to 1.5×10^8 (from 10^7 to 7×10^7), as considered in Fig. 5 (6), the best results are obtained for $N_i = 3000$, $N_s = 1024$ ($N_i = 30$, $N_s = 16384$). We use these values for the noisy optimization presented in Figs. 5 and 6, respectively.

F.2 Optimization with Virtual Distillation

In this Appendix, we compare 5-qubit and 8-qubit MaxCut QAOA optimization of the VD-mitigated cost function with optimization of the noisy and CDR-mitigated cost function for $p = 2, 4$. We perform the comparison using the same randomly chosen graphs from the Erdös-Rényi ensemble as in Figs. 5, 6. We perform VD mitigation for each expectation value of a 2-site term of H_{MaxCut} according to (13). Therefore, a key assumption is that we neglect derangement noise, which would affect realistic VD implementation on hardware [47]. We use the Nelder-Mead algorithm as described above. We have $N_i = 30$ as for CDR simulations from Figs. 5, 6 and assign $\widetilde{N}_s = 65536, 10^6$ shots in order to estimate $\mathrm{Tr}[\widetilde{\rho}^M Z_i Z_j]$ for each 2-site term of H_{MaxCut} and $\mathrm{Tr}[\widetilde{\rho}^M]$. Consequently, the total shot cost of the mitigated cost function estimation is $(n_e + 1) \times \widetilde{N}_s$, where n_e is the number of Max-Cut graph edges.

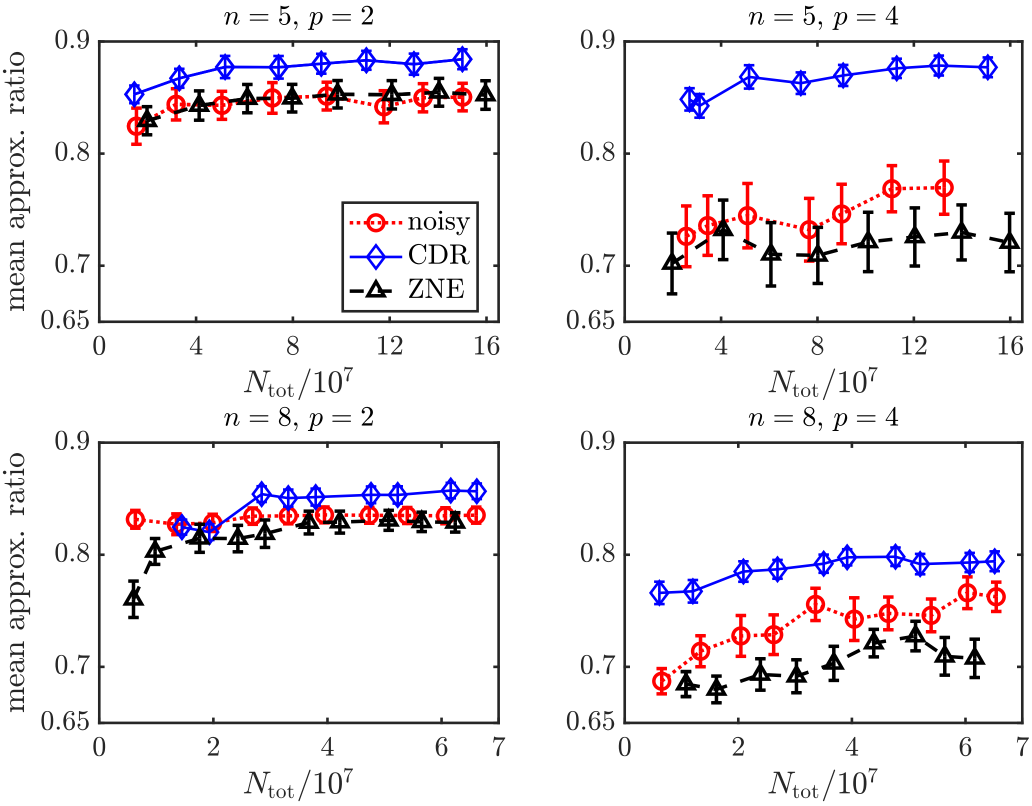

Figure 10: **Zero noise extrapolation mitigated optimization for $p = 2, 4$, 5-qubit and 8-qubit MaxCut QAOA.** We plot the approximation ratio averaged over instances of Max-Cut graphs randomly chosen from the Erdös-Rényi ensemble as a function of total shot number N_{tot}. For $n = 5$ ($n = 8$), we choose 36 (30) graphs. The results were obtained with $N_i = 30$ initializations and 30000 shots per ZNE-mitigated cost function evaluation. ZNE is performed by CNOT identity insertions with noise levels amplified by factors $a_0 = 1, a_1 = 3, a_2 = 5$, and with a linear extrapolation. The error bars are computed as in Fig. 5. For reference, we also present the results of CDR-mitigated and noisy optimization from Figs. 5, 6.

We consider $M = 2, 3$ state copies as the shot cost of VD mitigation for given precision grows with increasing M [92] and $M = 2, 3$ was shown to be sufficient for typical applications [107]. We find that for this setup $M = 2$ gives better results than $M = 3$ similar to our analytical results. Here we allow for N_{tot} up to 2×10^9 for $n = 5$ and up to 2×10^{10} for $n = 8$, i.e. up to 1 and 2 order of magnitudes more shots than considered for CDR-mitigated and noisy optimization in Figs. 5, 6, respectively.

We gather the results in Figs. 9 comparing them with noisy and CDR mitigated optimization from Figs. 5, 6. We find that even with smaller N_{tot} the noisy and CDR-mitigated optimization outperforms the VD-mitigated optimization. This example shows that even for circuits outside of the NIBP regime there is no guarantee that using an error-mitigated cost function leads to better performance than noisy cost function optimization. We note that this result does not prohibit VD-mitigated optimization advantage for different choices of N_{tot}, M or the shot number per cost function evaluation.

F.3 Optimization with Zero-Noise Extrapolation (ZNE)

Here, we analyze the optimization of ZNE-mitigated cost function for 5-qubit and 8-qubit MaxCut QAOA optimization. As in Figs. 5, 6, 9 for CDR and VD, we investigate number of rounds satisfying $p = 2, 4$. More precisely, we have used the same Erdös-Rényi graphs as the ones in the benchmark simulations described above. We use the Nelder-Mead algorithm with $N_i = 30$ initializations, the same as for the CDR- and VD-mitigated optimization. For each noise level used to perform an extrapolation to the zero-noise limit, we evaluate the cost function with $N_s = 10000$ shots. Consequently, the shot cost of ZNE-mitigated cost function evaluation is $n_l N_s$ where n_l is the number of noise levels. Here we investigate N_{tot} similar to N_{tot} for the noisy and CDR-mitigated optimization, i.e., $N_{\text{tot}} = (2-16)\times 10^7$ for $n = 5$, and $N_{\text{tot}} = (1 - 7) \times 10^7$ for $n = 8$. We have found that for considered here N_{tot} and $n = 5$ values, our choice of N_s leads to a better quality of the ZNE-mitigated optimization than $N_s = 1000$ and $N_s = 100000$.

ZNE is performed by CNOT identity insertions, and we consider noise levels amplified by factors $a_0 = 1, a_1 = 3, a_2 = 5$, i.e., a CNOT gate in the original circuit is replaced by a_i CNOTs [98]. We use linear extrapolation to obtain the ZNE-mitigated expectation values for each term in the Hamiltonian. Such an extrapolation in the presence of more than two noise levels has been proposed to improve the robustness of ZNE results [98] for realistic noise whose strength is challenging to scale accurately and has been applied in real-hardware ZNE implementations [99]. Furthermore, we find that for $n = 5$ and $N_{\text{tot}} = (1 - 10) \times 10^7$, using all three values of a_l for linear extrpolation outperforms ZNE-mitigated optimization with $a_0 = 1$, $a_1 = 3$. We also find that this choice outperforms an approach with $a_0 = 1$, $a_1 = 3$, $a_2 = 5$ using quadratic extrapolation for the considered problem parameters. We note that a detailed characterization of the effects of choice of the noise levels on performance of ZNE-mitigated optimization is beyond the scope of this work. However, we explore a range a hyperparameters in order to quickly gauge the power of a relatively simple ZNE approach in comparison to CDR and VD approaches analyzed above.

We gather the results in Fig. 10 plotting the approximation ratio averaged over graph instances versus N_{tot}, the same as in Figs. 5, 6, 9, and comparing it to the noisy and CDR-mitigated results. For $p = 2$, the ZNE-mitigated optimization gives results similar to the noisy one. For $p = 4$, the ZNE-mitigated approximation ratios are slightly worse than the ones obtained by the noisy optimization.