Beyond unital noise in variational quantum algorithms: noise-induced barren plateaus and limit sets

Phattharaporn Singkanipa[1] and Daniel A. Lidar[2]

[1]Department of Physics and Center for Quantum Information Science & Technology, University of Southern California, Los Angeles, CA 90089, USA

[2]Departments of Electrical Engineering, Chemistry, Physics & Astronomy, and Center for Quantum Information Science & Technology, University of Southern California, Los Angeles, CA 90089, USA

Variational quantum algorithms (VQAs) hold much promise but face the challenge of exponentially small gradients. Unmitigated, this barren plateau (BP) phenomenon leads to an exponential training overhead for VQAs. Perhaps the most pernicious are noise-induced barren plateaus (NIBPs), a type of unavoidable BP arising from open system effects, which have so far been shown to exist for unital noise maps. Here, we generalize the study of NIBPs to more general completely positive, trace-preserving maps, investigating the existence of NIBPs in the unital case and a class of non-unital maps we call Hilbert-Schmidt (HS)-contractive. The latter includes amplitude damping. We identify the associated phenomenon of noise-induced limit sets (NILS) of the VQA cost function and prove its existence for both unital and HS-contractive non-unital noise maps. Along the way, we extend the parameter shift rule of VQAs to the noisy setting. We provide rigorous bounds in terms of the relevant variables that give rise to NIBPs and NILSs, along with numerical simulations of the depolarizing and amplitude-damping maps that illustrate our analytical results.

1 Introduction

Variational quantum algorithms (VQAs) are promising applications of quantum computing in the NISQ era [1, 2, 3, 4]. These algorithms leverage a customizable quantum circuit design, integrating both quantum and classical computation capabilities. Using parameterized quantum circuits, they compute problem-specific cost functions, followed by classical optimization to iteratively update the parameters. This hybrid quantum-classical optimization process continues until predefined termination criteria are met.

Previous studies have demonstrated that VQA circuits operable within the existing noise levels and hardware connectivity limitations of the NISQ era, already find applications across diverse domains, such as quantum optimization [5, 6, 7], quantum optimal control [8], linear systems [9, 10, 11], quantum metrology [12, 13], quantum compiling [14, 15], quantum error correction [16, 17], quantum machine learning [18, 19] and quantum simulation [20, 21, 22]. Moreover, VQA has been established as a universal model of quantum computation [23].

Despite their comparable computational power to other quantum models and demonstrated advantages, VQAs exhibit inherent constraints that present scalability challenges for problems of arbitrary scale. Specifically, VQAs for random circuits suffer from exponentially vanishing gradients, commonly referred to as the Barren Plateau (BP) phenomenon [24, 25, 26, 27, 28, 29, 30, 31, 32, 33]. This phenomenon renders the parameter training step asymptotically impossible for circuits with a sufficiently large number of qubits n, even at shallow circuit depth.

Here, we study noise-induced barren plateaus (NIBPs), which emerge under decoherence-induced noise [34]. NIBPs were previously shown to be present in sufficiently deep circuits subjected to unital maps [31]. This holds true even in constant-width or non-random circuits. Alternatively, NIBPs exist under strictly contracting noise maps when the parameter shift rule (PSR) [18, 35] is applicable [36].

Unlike other BP types, for which mitigation strategies have been proposed [37, 25, 38, 39, 40, 41, 42, 43, 44, 45, 46], it remains unclear whether NIBPs can be similarly mitigated. Experimental investigations on small systems have suggested that error mitigation (EM) techniques enable VQAs to more closely approach the true ground-state energy [47]. Clifford Data Regression has proven effective in mitigating errors and reversing the concentration of cost function values [48]. Nevertheless, it is noteworthy that the majority of EM protocols do not enhance trainability or even exacerbate the lack of trainability. Additionally, post-processing expectation values of noisy circuits is not advantageous in the context of NIBP [49]. Previous work suggests that stochastic noise could be helpful for training VQAs [50]; an interesting open question that remains is whether there is an intermediate noise regime where we can train VQAs by exploiting noise.

In this work, we extend the study of NIBPs to completely positive trace-preserving (CPTP) maps, including both unital maps and a class of non-unital maps we call HS-contractive. A rigorous definition of this class is given in Definition 1 below, but intuitively, this is the class of maps under which the Bloch vector (or more generally, the coherence vector) is shrunk before it is shifted, just as in the case of the amplitude damping map. We analytically derive the scaling of the cost function gradient as a function of circuit width n, circuit depth L, and noise strength. We find that HS-contractive non-unital noise need not necessarily give rise to NIBPs, but instead exhibits a different phenomenon, which we refer to as a noise-induced limit set (NILS). Moreover, we simplify the NIBP derivation compared to Ref. [31], guided by the intuition gained by considering the effect of noise on the single-qubit Bloch sphere. We generalize this to n-qubit systems via the coherence vector and compute derivatives of the cost function via the PSR. In addition, we investigate the applicability of the PSR under control noise and random unitary noise and assess the impact of these noise types on the bounds we derive. We find analytical expressions for the dependence of the circuit depth L on relevant noise and circuit parameters that give rise to NIBP and NILS. Our analytical results are supported by numerical simulations.

This paper is organized as follows. Background results regarding VQA, PSR, the coherence vector, and characteristics of CPTP noise maps are presented in Section 2. Section 3 extends the PSR analysis to scenarios involving noise. We study the effects of HS-contractive non-unital and general unital noise in Section 4 and Section 5, respectively. In the unital case, we reprove that NIBP is always present. In the HS-contractive non-unital case, we find new results, in particular the phenomenon of NILS. Our theoretical findings are supported by numerical simulations in Section 6. We summarize our findings in Section 7.

2 Preliminaries

In this section, we review pertinent technical details and establish the notation we use to derive our results.

2.1 VQA and PSR

We adopt the variational quantum algorithm (VQA) framework of Ref. [31] and consider a general class of parameterized unitary ansatzes:

$$U(\boldsymbol{\theta}) = \Pi_{l=L}^{1} U_l(\boldsymbol{\theta}_l), \tag{1a}$$

$$U_l(\boldsymbol{\theta}_l) = \Pi_{m=G_l}^{1} e^{-\frac{i}{2}\theta_{lm} H_{lm}} W_{lm}, \tag{1b}$$

where L is the circuit depth and the $U_l(\boldsymbol{\theta}_l)$ are unitaries sequentially applied by layers. The l'th layer consists of G_l gates: the unparametrized gates denoted by W_{lm} (such as CNOT) and the gates generated by dimensionless Hamiltonians denoted by H_{lm} (the m'th gate in the l'th layer). In writing the various gates in Eq. (1) we implicitly assume that they are in a tensor product with the identity operator acting on all the qubits that are not explicitly labeled. The set $\boldsymbol{\theta} = \{\boldsymbol{\theta}_l\}_{l=1}^{L}$ consists of vectors of dimensionless continuous parameters $\boldsymbol{\theta}_l = \{\theta_{lm}\}_{m=1}^{G_l}$ that are optimized to minimize a cost function C_Ω expressed as the expectation value of an operator Ω:

$$C_\Omega(\boldsymbol{\theta}) = \mathrm{Tr}[\Omega\, \mathcal{U}(\boldsymbol{\theta})(\rho_0)]. \tag{2}$$

Here,

$$\mathcal{U}(\boldsymbol{\theta})(\rho_0) \equiv U(\boldsymbol{\theta})\rho_0 U^\dagger(\boldsymbol{\theta}) = \rho(\boldsymbol{\theta}) \tag{3}$$

is the unitary superoperator acting on the initial state ρ_0. An important special case, which we

focus on, is when $\Omega = H$, the "problem Hamiltonian" whose energy one is trying to minimize, and in this case we simply write C for the cost function.

For n qubits, we can always parametrize the traceless gate-generating Hamiltonians as

$$H_{lm} = \sum_{j=1}^{d^2-1} h_{lmj} P_j = \boldsymbol{h}_{lm} \cdot \boldsymbol{P}, \tag{4}$$

where $P_j \in \{I, \sigma^x, \sigma^y, \sigma^z\}^{\otimes n}$ is a Pauli string, i.e., a tensor product of up to n Pauli matrices, $P_0 = I^{\otimes n}$, I is the identity operator, and $d = 2^n$. We assume that the P_j's are ordered such that j increases with the Hamming weight of the Pauli string, i.e., the number of non-identity terms in P_j (the manner in which j increases at fixed Hamming weight does not matter for our purposes), and $\boldsymbol{P} = (P_1, \ldots, P_{d^2-1})$.

In most cases of interest, the $h_{lmj} \in \mathbb{R}$ vanish for strings involving more than two Pauli matrices, i.e., the Hamiltonians are two-local. This framework includes the Quantum Approximate Optimization Algorithm or Quantum Alternating Operator Ansatz (QAOA) [5, 7], where $H_l = H_1 \delta_{l,\text{odd}} + H_2 \delta_{l,\text{even}}$ $\forall l, m$ with $[H_1, H_2] \neq 0$, the Unitary Coupled Cluster (UCC) ansatz [51], where the h_{lmj} are coefficients derived from one- and two-electron integrals, which is used in the Variational Quantum Eigensolver (VQE) algorithm [20] with applications in quantum chemistry [52], and the Hardware Efficient VQE Ansatz, which tries to minimize the circuit depth (i.e., the set of non-zero θ_{lm}) given a predefined gate-set tailored to particular hardware [21].

The parameter shift rule (PSR) is frequently used in evaluating derivatives of cost functions in VQAs [8, 18, 35]. For a cost function $C(\boldsymbol{\theta})$ as in Eq. (2), the PSR states that (see, e.g., [35, Table 2]):

$$\frac{\partial C(\boldsymbol{\theta})}{\partial \theta_{lm}} = \frac{1}{2}[C(\boldsymbol{\theta} + \boldsymbol{\theta}^{\pi/2}_{lm}) - C(\boldsymbol{\theta} - \boldsymbol{\theta}^{\pi/2}_{lm})], \tag{5}$$

where

$$\boldsymbol{\theta}^{\pi/2}_{lm} = \frac{\pi}{2} \hat{e}_{lm} \tag{6}$$

and $\{\hat{e}_{lm}\}$ are standard unit vectors [i.e., the (l,m)th component of $\boldsymbol{\theta}^{\pi/2}_{lm}$ is $\pi/2$ and the rest are zero]. We reprove this result in Section 3.1. The essential point is that we can compute the derivative by means of a finite difference.

2.2 Nice operator basis and Schatten p-norms

Consider a Hilbert space $\mathcal{H}$ of dimension $d < \infty$. The space of bounded linear operators acting on $\mathcal{H}$ is denoted $\mathcal{B}(\mathcal{H})$. Let $\mathcal{M}(d, \mathbb{F})$ denote the vector space of $d \times d$ matrices with coefficients in $\mathbb{F}$, where $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. For our purposes it suffices to identify $\mathcal{B}(\mathcal{H})$ with $\mathcal{M}(d, \mathbb{C})$. The Hilbert-Schmidt inner product is $\langle A, B\rangle \equiv \text{Tr}(A^\dagger B)$ for any two operators $A, B \in \mathcal{B}(\mathcal{H})$.

We define a "nice operator basis" as a set $\{F_j\}_{j=0}^{d^2-1} \in \mathcal{B}(\mathcal{H})$, where $F_0 = \frac{1}{\sqrt{d}} I$, $\text{Tr}(F_j) = 0$ $\forall j \geq 1$, that in addition satisfies the following properties:

$$F_j = F_j^\dagger, \quad \langle F_j, F_k\rangle = \text{Tr}(F_j F_k) = \delta_{jk} \quad \forall j, k. \tag{7}$$

The normalized Pauli strings $\{\frac{1}{\sqrt{d}} P_j\}_{j=0}^{d^2-1}$ (where $d = 2^n$) are a convenient explicit choice for the nice operator basis. Another convenient choice is the set of generalized $d \times d$ Gell-Mann matrices [53, 54], normalized such that $\text{Tr}(F_j F_k) = \delta_{jk}$ is satisfied.

Let $|A| \equiv \sqrt{A^\dagger A}$. Recall that the Schatten p-norm $\|A\|_p$ is given, for $1 \leq p < \infty$, by the p-norm of the singular values σ_i of the operator $A \in \mathcal{B}(\mathcal{H})$:

$$\|A\|_p = \text{Tr}(|A|^p)^{1/p} = (\sum_i \sigma_i^p)^{1/p}. \tag{8}$$

$\|A\|_1 = \text{Tr}(|A|)$ is the trace norm (sum of the singular values), $\|A\|_2 = \sqrt{\langle A, A\rangle} = \sqrt{\text{Tr}(|A|^2)}$ is the Hilbert-Schmidt or Frobenius norm, and $\|A\|_\infty$ is the operator norm (largest singular value). Without risk of confusion, we use $\|A\|$ to denote $\|A\|_\infty$ and also use $\|\boldsymbol{v}\|$ to denote the Euclidian norm (i.e., 2-norm) of any *vector* $\boldsymbol{v} \in \mathcal{H}$ from hereon.

2.3 The coherence vector

Quantum states are represented by density operators $\rho \in \mathcal{B}_+(\mathcal{H})$ (positive trace-class operators acting on $\mathcal{H}$) with unit trace: $\text{Tr}\rho = 1$. Elements of $\mathcal{B}[\mathcal{B}(\mathcal{H})]$, i.e., linear transformations $\mathcal{N} : \mathcal{B}(\mathcal{H}) \mapsto \mathcal{B}(\mathcal{H})$, are called superoperators, or maps.

Complete positivity of a superoperator $\mathcal{N}$ is equivalent to the statement that $\mathcal{N}$ has a Kraus representation [55]: $\forall X \in \mathcal{B}(\mathcal{H})$,

$$\mathcal{N}(X) = \sum_\alpha K_\alpha X K_\alpha^\dagger, \tag{9}$$

where the $\{K_\alpha\}$ are called Kraus operators. When they satisfy $\sum_\alpha K_\alpha^\dagger K_\alpha = I$, the map $\mathcal{N}$ is trace-preserving.

The density operator can be expanded in an arbitrary nice operator basis as

$$\rho = \frac{1}{d} I + \sum_{j=1}^{d^2-1} v_j F_j = \frac{1}{\sqrt{d}} F_0 + \boldsymbol{v} \cdot \boldsymbol{F}, \tag{10}$$

where $\boldsymbol{F} = \{F_1, \ldots, F_{d^2-1}\}^T$, and $\boldsymbol{v} = \{v_1, \ldots, v_{d^2-1}\}$ is called the *coherence vector*.

We summarize two well-known facts about the coherence vector. First, the purity

$$P \equiv \mathrm{Tr}\rho^2 = \langle \rho, \rho \rangle = \|\rho\|_2^2 \tag{11}$$

is bounded by

$$0 \leq \|\boldsymbol{v}\| = \left(P - \frac{1}{d}\right)^{1/2} \leq \left(1 - \frac{1}{d}\right)^{1/2} < 1 \ . \tag{12}$$

See Appendix A for a proof.

Second, let $\mathcal{M}(d, F)$ denote the vector space of $d \times d$ matrices with coefficients in the field F.

Proposition 1. *The CPTP map $\rho' = \mathcal{N}(\rho)$ is equivalent to the affine coherence vector transformation*

$$\boldsymbol{v}' = M\boldsymbol{v} + \boldsymbol{c}, \tag{13}$$

where $M \in \mathcal{M}(d^2 - 1, \mathbb{R})$ and $\boldsymbol{c} \in \mathbb{R}^{d^2-1}$ have elements given by

$$M_{ij} = \langle F_i, \mathcal{N}(F_j) \rangle = \sum_\alpha \mathrm{Tr}(F_i K_\alpha F_j K_\alpha^\dagger) \tag{14a}$$

$$c_i = \frac{1}{d} \langle F_i, \mathcal{N}(I) \rangle = \frac{1}{d} \sum_\alpha \mathrm{Tr}(F_i K_\alpha K_\alpha^\dagger) \ . \tag{14b}$$

See Appendix B for a proof.

The Gell-Mann matrices reduce to the standard Pauli matrices for $d = 2$, normalized such that $\boldsymbol{F} = \boldsymbol{\sigma}/\sqrt{2} = (\sigma^x, \sigma^y, \sigma^z)/\sqrt{2}$. Therefore, in the case of single qubit, we can write ρ in the well-known form

$$\rho = \frac{1}{2}(I + \bar{\boldsymbol{v}} \cdot \boldsymbol{\sigma}) = \frac{1}{\sqrt{2}} F_0 + \boldsymbol{v} \cdot \boldsymbol{F}, \tag{15}$$

where $\boldsymbol{v} = \bar{\boldsymbol{v}}/\sqrt{2}$. Note that $\|\bar{\boldsymbol{v}}\| \leq 1$, which is the convention for the Bloch sphere representation. We avoid this normalization and instead use the nice operator basis convention from here on, even for $d = 2$, so that $\|\boldsymbol{v}\| \leq 1/\sqrt{2}$ [Eq. (12)].

2.4 Unital maps

A unital map $\mathcal{N}$ is defined as satisfying $\mathcal{N}(I) = I$, hence $\sum_\alpha K_\alpha K_\alpha^\dagger = I$.

Lemma 1. *Unital CPTP maps are purity non-increasing: $P' \leq P$, where P and P' are, respectively, the purity of ρ and $\rho' = \mathcal{N}(\rho)$. Equality holds for all ρ iff the map is unitary.*

See Appendix C for a proof.

Lemma 2. *For unital CPTP maps $\mathcal{N}$ we have:*

$$\boldsymbol{c} = \boldsymbol{0}, \tag{16a}$$

$$\|\boldsymbol{v}'\| = \|M\boldsymbol{v}\| \leq \|\boldsymbol{v}\|. \tag{16b}$$

Equality in Eq. (16b) holds for all $\boldsymbol{v}$ iff $\mathcal{N}$ is unitary, in which case M is norm-preserving (hence orthogonal).

Proof. To prove Eq. (16a), note that it follows from Eq. (14) that $c_i = \frac{1}{d}\mathrm{Tr}[F_i \sum_\alpha K_\alpha K_\alpha^\dagger] = \frac{1}{d}\mathrm{Tr}[F_i] = 0$.

To prove Eq. (16b), note that from Eqs. (12) and (13) we have $\|M\boldsymbol{v}\| = \|\boldsymbol{v}'\| = \sqrt{P' - 1/d}$, where $P' = \mathrm{Tr}[(\rho')^2]$, $\rho' = \mathcal{N}(\rho)$. If $\|M\boldsymbol{v}\| > \|\boldsymbol{v}\| = \sqrt{P - 1/d}$ then $P' > P$, which contradicts Lemma 1. Moreover, from Lemma 1, $P' = P$ iff $\mathcal{N}$ is unitary. Since $P' = P$ is equivalent to $\|M\boldsymbol{v}\| = \|\boldsymbol{v}'\| = \|\boldsymbol{v}\|$, we have equality for all $\boldsymbol{v}$ iff $\mathcal{N}$ is unitary. □

2.5 Non-unital maps

From here on, when we consider non-unital noise maps we restrict our analysis to the following class, which are contractive under the Hilbert-Schmidt (HS) norm:

Definition 1. *A (finite-dimensional) map $\mathcal{N}$ is called* HS-contractive *if $\exists r < 1$ s.t. for all states $\rho_1 \neq \rho_2$ we have $\|\mathcal{N}(\rho_1) - \mathcal{N}(\rho_2)\|_2 \leq r\|\rho_1 - \rho_2\|_2$.*

This definition of contractivity is different from the standard one for CPTP maps, which are well known to be contractive under the trace norm, i.e., $\|\mathcal{N}(\rho_1) - \mathcal{N}(\rho_2)\|_1 \leq \|\rho_1 - \rho_2\|_1$ for any pair of states ρ_1, ρ_2 and any CPTP map $\mathcal{N}$, including all non-unital maps [56]. Other notions of contractivity also exist, such as for general positive maps between matrix spaces that include the zero element, in which case a non-unital map is always non-contractive [57]; see Appendix D for details.

Lemma 3. *A map is HS-contractive if and only if its matrix M satisfies $\|M\| < 1$.*

Proof. In terms of a nice operator basis and the corresponding coherence vector, we can write $\|\rho_1 - \rho_2\|_2 = \|(\boldsymbol{v}_1 - \boldsymbol{v}_2)\cdot\boldsymbol{F}\|_2$. Let $\boldsymbol{v} = \boldsymbol{v}_1 - \boldsymbol{v}_2$ and using the properties of $\{F_i\}$ in Eq. (7), we have $\|\boldsymbol{v}\cdot\boldsymbol{F}\|_2 = \|\boldsymbol{v}\|$. Let $\rho' = \mathcal{N}(\rho)$ and $\boldsymbol{v}' = M\boldsymbol{v} + \boldsymbol{c}$. We can write $\|\rho_1' - \rho_2'\|_2 = \|(\boldsymbol{v}_1' - \boldsymbol{v}_2')\cdot\boldsymbol{F}\|_2 = \|(M\boldsymbol{v}_1 - M\boldsymbol{v}_2)\cdot\boldsymbol{F}\|_2 = \|M\boldsymbol{v}\|$.

($\Rightarrow$) By definition of $\mathcal{N}$ being HS-contractive we have $\|M\boldsymbol{v}\| = \|\mathcal{N}(\rho_1) - \mathcal{N}(\rho_2)\|_2 \leq r\|\rho_1 - \rho_2\|_2 = r\|\boldsymbol{v}\| < \|\boldsymbol{v}\|$, $\forall \boldsymbol{v} \neq \boldsymbol{0}$. This is true in particular for the vector $\boldsymbol{v}$ that achieves the supremum in $\sup_{\boldsymbol{v}\neq\boldsymbol{0}} \|M\boldsymbol{v}\|/\|\boldsymbol{v}\| = \|M\|$, which implies that $\|M\| < 1$.

($\Leftarrow$) Let $r = \|M\|$. By assumption, $r < 1$. We have $\|M\boldsymbol{v}\| \leq \|M\|\|\boldsymbol{v}\| = r\|\boldsymbol{v}\|$ for any $\boldsymbol{v}$. Taking two arbitrary density matrices ρ_1, ρ_2 with respective coherence vectors $\boldsymbol{v}_1, \boldsymbol{v}_2$ and applying this to $\boldsymbol{v} = \boldsymbol{v}_1 - \boldsymbol{v}_2$, we obtain $\|\mathcal{N}(\rho_1) - \mathcal{N}(\rho_2)\|_2 \leq r\|\rho_1 - \rho_2\|_2$. □

Using the polar decomposition, let us decompose M in Eq. (14) as $M = OS$, where O is orthogonal and $S = |M|$ is positive semidefinite. Let $\sigma_{\max\,/\,\min}(M)$ denote the largest/smallest singular value of M (the largest/smallest eigenvalue of S). We interpret O as a rotation, S as a dilation, and $\boldsymbol{c}$ [Eq. (13)] as an affine shift.

Lemma 4. *For HS-contractive non-unital CPTP maps $\mathcal{N}$, we have:*

$$\boldsymbol{c} \neq \boldsymbol{0} \tag{17a}$$

$$\|\boldsymbol{c}\| \leq 1/\sqrt{1 - 1/d} \tag{17b}$$

$$\|M\| < 1 \tag{17c}$$

$$\|M\boldsymbol{v}\| < \|\boldsymbol{v}\|, \forall \boldsymbol{v} \neq \boldsymbol{0}. \tag{17d}$$

Lemma 5. *Any single-qubit ($d = 2$) non-unital map is always HS-contractive.*

See Appendix E for proofs of these two Lemmas.

Corollary 1. *If $\mathcal{N}$ is an HS-contractive non-unitary CPTP map and $\mathcal{U}$ is unitary, with corresponding coherence vector transformations $\boldsymbol{v}' = M\boldsymbol{v} + \boldsymbol{c}$ and $\boldsymbol{v}' = O\boldsymbol{v}$, where O is orthogonal, then for the maps $\mathcal{N}\circ\mathcal{U}$ and $\mathcal{U}\circ\mathcal{N}$ we have:*

$$\|MO\boldsymbol{v}\|, \|OM\boldsymbol{v}\| < \|\boldsymbol{v}\|. \tag{18}$$

Proof. This is an immediate consequence of Lemma 2 and Lemma 4, since whether $\mathcal{N}$ is unital or HS-contractive non-unital we have $\|M\boldsymbol{v}\| < \|\boldsymbol{v}\|$, so that: $\|M(O\boldsymbol{v})\| < \|O\boldsymbol{v}\| = \|\boldsymbol{v}\|$, and $\|O(M\boldsymbol{v})\| = \|M\boldsymbol{v}\| < \|\boldsymbol{v}\|$. □

3 Parameter shift rule in the presence of noise

The PSR given in Eq. (5) is valid for closed systems undergoing unitary evolution without noise. This section presents a brief rederivation for the noiseless setting, which we then adapt to accommodate scenarios involving control noise and random unitary noise. We also bound the gradient of the cost function in both of these cases.

3.1 The noiseless PSR case

For simplicity (and w.l.o.g., but at the expense of increasing the circuit depth), let us assume that each of the gate Hamiltonians H_{lm} [Eq. (4)] is a single Pauli string, i.e., we can write the terms in Eq. (1b) as

$$\exp(-\frac{i}{2}\theta_{lm}H_{lm}) = \exp(-\frac{i}{2}\theta_\mu P_{j(\mu)}) \equiv U(\theta_\mu), \tag{19}$$

where $\mu = (l, m)$ is the location of the gate in the circuit (the m'th gate in the l'th layer), we wrote $j(\mu)$ since the Pauli string type depends on the location μ, and we dropped the subscript j on U since, in the calculation below, the type of Pauli string will not matter. From now on we sometimes also write j instead of $j(\mu)$ for notational simplicity.

Recall Eq. (1) and let us write $\boldsymbol{\theta} = \{\boldsymbol{\theta}_a, \theta_\mu, \boldsymbol{\theta}_b\}$, where $\boldsymbol{\theta}_b$ and $\boldsymbol{\theta}_a$ collect the rotation angles before and after θ_μ, respectively. Thus, $\mathcal{U}(\boldsymbol{\theta}) = \mathcal{U}(\boldsymbol{\theta}_a)\circ\mathcal{U}(\theta_\mu)\circ\mathcal{U}(\boldsymbol{\theta}_b)$, where we used the unitary superoperator notation of Eq. (3).

Anticipating the derivative with respect to θ_μ, we rewrite the cost function as follows:

$$C(\boldsymbol{\theta}) = \mathrm{Tr}[H\rho(\boldsymbol{\theta})] = \mathrm{Tr}[H\mathcal{U}(\boldsymbol{\theta})(\rho_0)] \tag{20a}$$

$$= \mathrm{Tr}[H\mathcal{U}(\boldsymbol{\theta}_a)\circ\mathcal{U}(\theta_\mu)\circ\mathcal{U}(\boldsymbol{\theta}_b)(\rho_0)] \tag{20b}$$

$$= \mathrm{Tr}[\tilde{H}\mathcal{U}(\theta_\mu)(\tilde{\rho})], \tag{20c}$$

where

$$\tilde{H} = \mathcal{U}^\dagger(\boldsymbol{\theta}_a)(H) = U^\dagger(\boldsymbol{\theta}_a)HU(\boldsymbol{\theta}_a) \tag{21a}$$

$$\tilde{\rho} = \mathcal{U}(\boldsymbol{\theta}_b)(\rho_0). \tag{21b}$$

Thus, since $\partial_\theta \mathcal{U}(\theta)(\cdot) = -\frac{i}{2}\mathcal{U}(\theta)([P_j, \cdot])$ for $\mathcal{U}(\theta)(\cdot) = e^{-i\theta P_j/2} \cdot e^{i\theta P_j/2}$:

$$\frac{\partial C(\boldsymbol{\theta})}{\partial \theta_\mu} = -\frac{i}{2}\mathrm{Tr}[\tilde{H}\mathcal{U}(\theta_\mu)([P_{j(\mu)}, \tilde{\rho}])]. \tag{22}$$

The following identity holds for $U(\theta) = \exp(-i\theta P_j/2)$, where P_j is an arbitrary Pauli string [18]:

$$[P_j, \rho] = i\mathcal{U}\left(\frac{\pi}{2}\right)(\rho) - i\mathcal{U}\left(-\frac{\pi}{2}\right)(\rho). \tag{23}$$

Using this identity in Eq. (22) we have:

$$\begin{aligned}
&\frac{\partial C(\boldsymbol{\theta})}{\partial \theta_\mu} \\
&= \frac{1}{2}\mathrm{Tr}\left[\tilde{H}\mathcal{U}(\theta_\mu)\left(\mathcal{U}\left(\frac{\pi}{2}\right)(\tilde{\rho}) - \mathcal{U}\left(-\frac{\pi}{2}\right)(\tilde{\rho})\right)\right] && \text{(24a)}\\
&= \frac{1}{2}\mathrm{Tr}[\tilde{H}\mathcal{U}\left(\theta_\mu + \frac{\pi}{2}\right)(\tilde{\rho})] \\
&\qquad - \frac{1}{2}\mathrm{Tr}[\tilde{H}\mathcal{U}\left(\theta_\mu - \frac{\pi}{2}\right)(\tilde{\rho})] && \text{(24b)}\\
&= \frac{1}{2}[C(\{\boldsymbol{\theta}_a, \theta_\mu + \frac{\pi}{2}, \boldsymbol{\theta}_b\}) - C(\{\boldsymbol{\theta}_a, \theta_\mu - \frac{\pi}{2}, \boldsymbol{\theta}_b\})], && \text{(24c)}
\end{aligned}$$

where in the last line we used Eq. (20c). This is the closed system PSR, Eq. (5).

3.2 Control noise

Now consider adding a small perturbation to the ideal Pauli generator of a gate, i.e., $P_j \mapsto P_j + A_j$, such that $\|A_j\| \ll 1$. In analogy to Eq. (4), we can decompose A_j in the Pauli basis such that

$$A_j = \sum_{k=0}^{d^2-1} a_{jk} P_k, \tag{25}$$

where $a_{jk} \in \mathbb{R}$. This amounts to control noise that perturbs the intended gate Hamiltonian P_j by a bounded (but not necessarily local) operator. We may now write the noisy version of the gate as:

$$U'(\theta_\mu) = \exp(-i\theta_\mu(P_{j(\mu)} + A_{j(\mu)})/2), \tag{26}$$

where the prime indicates the presence of noise. Note that this noise model includes both under/over-rotation and axis-angle errors. In the former case $a_{jk} = a\delta_{jk}$, so that $\theta_\mu \mapsto \theta_\mu + a$, while in the latter case $a_{jk} \neq \delta_{jk}$, so that the rotation axis is no longer P_j. Since $j = j(\mu)$, this noise model also accounts for the location of the errors in the circuit. The errors can be deterministic or stochastic, but our model assumes that they are constant throughout the duration of each gate.

Using Eq. (20c) and Eq. (22), the noisy version of the cost function and its gradient with respect to θ_μ can be written as:

$$\begin{aligned}
C'(\boldsymbol{\theta}) &= \mathrm{Tr}[\tilde{H}\mathcal{U}'(\theta_\mu)(\tilde{\rho})] && \text{(27a)}\\
\frac{\partial C'(\boldsymbol{\theta})}{\partial \theta_\mu} &= -\frac{i}{2}\mathrm{Tr}[\tilde{H}\mathcal{U}'(\theta_\mu)([P_{j(\mu)} + A_{j(\mu)}, \tilde{\rho}])]. && \text{(27b)}
\end{aligned}$$

Expanding Eq. (27b) using Eq. (23):

$$\begin{aligned}
\frac{\partial C'(\boldsymbol{\theta})}{\partial \theta_\mu} &= -\frac{i}{2}\mathrm{Tr}[\tilde{H}\mathcal{U}'(\theta_\mu)([P_j + A_j, \tilde{\rho}])] && \text{(28a)}\\
&= -\frac{i}{2}\mathrm{Tr}[\tilde{H}\mathcal{U}'(\theta_\mu)([P_j, \tilde{\rho}])] \\
&\quad - \frac{i}{2}\sum_k a_{jk}\mathrm{Tr}[\tilde{H}\mathcal{U}'(\theta_\mu)([P_k, \tilde{\rho}])] && \text{(28b)}\\
&= \frac{1}{2}\mathrm{Tr}[\tilde{H}\mathcal{U}'(\theta_\mu)\left(\mathcal{U}_j\left(\frac{\pi}{2}\right)(\tilde{\rho}) - \mathcal{U}_j\left(-\frac{\pi}{2}\right)(\tilde{\rho})\right)] \\
&\quad + \frac{1}{2}\sum_k a_{jk}\mathrm{Tr}[\tilde{H}\mathcal{U}'(\theta_\mu)\times && \text{(28c)}\\
&\qquad \left(\mathcal{U}_k\left(\frac{\pi}{2}\right)(\tilde{\rho}) - \mathcal{U}_k\left(-\frac{\pi}{2}\right)(\tilde{\rho})\right)],
\end{aligned}$$

where, in the last two lines, we have used $\mathcal{U}_j$ to denote a map generated by P_j, so that it is distinguished from $\mathcal{U}'$, which is generated by $P_j + A_j$ as given in Eq. (26). We also denote

$$\tilde{\rho}^{\pm}_{j(\mu)} \equiv \mathcal{U}'(\theta_\mu)\mathcal{U}_{j(\mu)}\left(\pm\frac{\pi}{2}\right)(\tilde{\rho}) = \frac{1}{\sqrt{d}}F_0 + \tilde{\boldsymbol{v}}^{\pm}_{j(\mu)} \cdot \boldsymbol{F}, \tag{29}$$

where $\tilde{\boldsymbol{v}}^{\pm}_{j(\mu)}$ is the corresponding coherence vector after an expansion in a nice operator basis. This bifurcation into the two paths labeled $\pm$ will turn out to be key to the NIBP phenomenon.

Let

$$\begin{aligned}
\tilde{\boldsymbol{w}}_{j(\mu)} &= \tilde{\boldsymbol{v}}^{+}_{j(\mu)} - \tilde{\boldsymbol{v}}^{-}_{j(\mu)}, && \text{(30a)}\\
\tilde{\xi}_{j(\mu)} &= \tilde{\rho}^{+}_{j(\mu)} - \tilde{\rho}^{-}_{j(\mu)} = \tilde{\boldsymbol{w}}_{j(\mu)} \cdot \boldsymbol{F}. && \text{(30b)}
\end{aligned}$$

Substituting Eq. (29) into Eq. (28), we have:

$$\frac{\partial C'(\boldsymbol{\theta})}{\partial \theta_\mu} \tag{31a}$$

$$= \frac{1}{2}\mathrm{Tr}[\tilde{H}(\tilde{\rho}_j^+ - \tilde{\rho}_j^-)] + \frac{1}{2}\sum_k a_{jk}\mathrm{Tr}[\tilde{H}(\tilde{\rho}_k^+ - \tilde{\rho}_k^-)]$$

$$= \frac{1}{2}\mathrm{Tr}(\tilde{H}\tilde{\xi}_j) + \frac{1}{2}\sum_k a_{jk}\mathrm{Tr}(\tilde{H}\tilde{\xi}_k) \tag{31b}$$

$$= \frac{1}{2}\mathrm{Tr}(\tilde{H}\tilde{\boldsymbol{w}}_j \cdot \boldsymbol{F}) + \frac{1}{2}\sum_k a_{jk}\mathrm{Tr}(\tilde{H}\tilde{\boldsymbol{w}}_k \cdot \boldsymbol{F}) \tag{31c}$$

$$= \frac{1}{2}\sum_{l=1}^{d^2-1}(\tilde{\boldsymbol{w}}_j)_l\tilde{h}_l + \frac{1}{2}\sum_k a_{jk}\sum_{l=1}^{d^2-1}(\tilde{\boldsymbol{w}}_k)_l\tilde{h}_l \tag{31d}$$

$$= \frac{1}{2}\tilde{\boldsymbol{w}}_{j(\mu)} \cdot \tilde{\boldsymbol{h}} + \frac{1}{2}\sum_k a_{jk}\tilde{\boldsymbol{w}}_{k(\mu)} \cdot \tilde{\boldsymbol{h}}, \tag{31e}$$

where we denoted $\tilde{h}_l \equiv \mathrm{Tr}(\tilde{H}F_l)$ and selectively added the μ-dependence for emphasis (though both j and k depend on μ). Since $\tilde{H}$ is related to H via a unitary transformation [Eq. (21a)] they have the same norm, and likewise $\|\tilde{\boldsymbol{h}}\| = \|\boldsymbol{h}\|$. We can now bound the derivative as follows:

$$\left|\frac{\partial C'(\boldsymbol{\theta})}{\partial \theta_\mu}\right| \leq \frac{1}{2}\|\tilde{\boldsymbol{w}}_{j(\mu)}\|\|\boldsymbol{h}\| + \frac{1}{2}\sum_k |a_{jk}|\|\tilde{\boldsymbol{w}}_{k(\mu)}\|\|\boldsymbol{h}\|. \tag{32}$$

The noise term in Eq. (32) (the second term) makes this bound looser than the noiseless case (just the first term), and it might seem that this could prevent the gradient from becoming vanishingly small. However, as long as $\|\tilde{\boldsymbol{w}}_{k(\mu)}\|$ is exponentially suppressed, it is not loose enough to escape the NIBP. We discuss this in more detail in Section 5.

3.3 Random unitary noise

The noise model in Section 3.2 is semiclassical, in that the bath is treated classically; a fully quantum noise model would replace Eq. (25) by $\sum_{k=0}^{d^2-1} a_{jk}P_k \otimes B_k + H_B$, where $\{B_k\}$ and H_B are, respectively, bath operators and the bath Hamiltonian. This would change the system-only unitary $U'(\theta_\mu)$ into a system-bath unitary, but we do not consider this case here. Instead, to account for a quantum noise model of faulty gates, we consider a (unital) noise map defined by a set of unitary operators and their corresponding probabilities $\{p_k, U_k\}$. I.e., with probability p_k the unitary $U_k = \exp(-i\theta_{\mu,k}P_k/2)$ is applied. Only one of these unitaries is the intended one; w.l.o.g., we call this index $j = j(\mu)$. We assume that $\sum_{k\neq j} p_k \ll p_j$ (i.e., $p_j \lesssim 1$).

For simplicity, we may assume that $\theta_{\mu,k} = \theta_\mu$, since the case $\theta_{\mu,k} = \theta_\mu + (\Delta\theta)_k$ was already accounted for in Section 3.2. Hence, each U_k is implicitly $U_k(\theta_\mu)$, which has the same angle dependence as $U_l(\theta_\mu)$ for $k \neq l$.

This noise model can be written in the Kraus representation as

$$\mathcal{V}(X) = \sum_k p_k U_k X U_k^\dagger \tag{33a}$$

$$= p_j\mathcal{U}(X) + \sum_{k\neq j} p_k \mathcal{V}_k'(X), \tag{33b}$$

where $\mathcal{V}_k'(X) = U_k X U_k^\dagger$ for $k \neq j$ and $\{\sqrt{p_k}U_k\}$ are the Kraus operators.

Using a similar approach as in the derivation of Eq. (32), we find:

$$\left|\frac{\partial C'(\boldsymbol{\theta})}{\partial \theta_\mu}\right| \leq p_{j(\mu)}\left|\frac{\partial C(\boldsymbol{\theta})}{\partial \theta_\mu}\right| + \frac{1}{2}\sum_{k\neq j} p_{k(\mu)}\|\tilde{\boldsymbol{w}}_{k(\mu)}\|\|\boldsymbol{h}\|. \tag{34}$$

Details are given in Appendix F. The conclusion regarding this bound is similar to the case discussed in Section 3.2.

3.4 Bounds on $\|\boldsymbol{h}\|$

Both Eqs. (32) and (34) involve the Hamiltonian norm $\|\boldsymbol{h}\|$, so we next bound this quantity.

Consider the scaling of the HS norm of the problem Hamiltonian. Writing such Hamiltonians as $H = \sum_{j=0}^{d^2-1} h_j F_j$, we have $\|H\|_2^2 = \sum_{j=0}^{d^2-1} h_j^2 = h_0^2 + \|\boldsymbol{h}\|^2$. In practice, we choose the nice operator basis $\{F_j\}$ as the normalized Pauli basis, $\{P_j\}/\sqrt{d}$.

Assume that the Pauli strings are ordered by Hamming weight k. We say that H is K-local if $h_j = 0$ for $j > K$ with K a constant independent of n. The number of non-zero h_j terms in $\boldsymbol{h}$ is at most $\sum_{k=1}^K \binom{n}{k}$, and a crude upper bound for $K \leq n/2$ (which holds in the K-local case) is

$$\sum_{k=1}^K \binom{n}{k} \leq K \max_{k\in[0,K]}\binom{n}{k} = K\binom{n}{K} \tag{35a}$$

$$= K\frac{n(n-1)\cdots(n-K+1)}{K!} \tag{35b}$$

$$\leq \frac{n^K}{(K-1)!}. \tag{35c}$$

Thus,

$$\|\boldsymbol{h}\| \leq \frac{h}{\sqrt{(K-1)!}} n^{K/2}, \quad h \equiv \max_{j>0} h_j. \tag{36}$$

4 Cost Function Concentration and Noise-Induced Limit Sets

Under the unital map setting of generalized Pauli noise, Ref. [31] has shown that the cost function concentrates. We now extend this to the action of HS-contractive non-unital noise maps and show furthermore that a new phenomenon of noise-induced limit sets arises in this context.

4.1 Cost function concentration under non-unital noise

Expanding H in a nice operator basis, we have:

$$H = \sum_{j=0}^{d^2-1} h_j F_j = h_0 F_0 + \boldsymbol{h} \cdot \boldsymbol{F}, \tag{37}$$

so that $\boldsymbol{h}$ collects the coordinates of the traceless component of H. The cost function can be written as:

$$C(\boldsymbol{\theta}) = \mathrm{Tr}[H\rho(\boldsymbol{\theta})] = \frac{1}{d}\mathrm{Tr}(H) + \boldsymbol{v} \cdot \boldsymbol{h}. \tag{38}$$

We already showed that when the noise is part of the gate, the gradient of the noisy cost function is bounded as in Eqs. (32) and (34). We now consider noise between gate applications, which we model as a concatenation of non-unitary CPTP maps $\mathcal{N}_l$ after applying all the noisy gates in the l'th layer:

$$\rho_{l+1} = [\mathcal{N}_{l+1} \circ \mathcal{U}'(\boldsymbol{\theta}_{l+1})](\rho_l) \quad \forall l \geq 0. \tag{39}$$

I.e., for the evolution of a circuit of depth L, with the initial state ρ_0, we have

$$\rho(\boldsymbol{\theta}) \equiv \rho_L(\boldsymbol{\theta}) = [\mathcal{N}_L \circ \mathcal{U}'(\boldsymbol{\theta}_L) \circ \cdots \circ \mathcal{N}_1 \circ \mathcal{U}'(\boldsymbol{\theta}_1)](\rho_0), \tag{40}$$

where, as before $\boldsymbol{\theta} = \{\boldsymbol{\theta}_l\}_{l=1}^L$, and $\mathcal{U}'(\boldsymbol{\theta}_l)$ denotes the noisy unitary superoperator in the l'th layer, formed from gates of the form given in Eq. (26).

Let $\boldsymbol{v}_l$ be the coherence vector corresponding to ρ_l in Eq. (39). Let us denote the transformed coherence vector after $\mathcal{N}_l \circ \mathcal{U}'_l(\boldsymbol{\theta}_l)$ by

$$\boldsymbol{v}_l = \Omega_l \boldsymbol{v}_{l-1} + \boldsymbol{c}_l \ , \quad \Omega_l \equiv M_l O_l, \tag{41}$$

with O_l the orthogonal rotation corresponding to $\mathcal{U}'_l(\boldsymbol{\theta}_l)$, and M_l the rotation+dilation and $\boldsymbol{c}_l$ the affine shift corresponding to $\mathcal{N}_l$ (see Section 2.5). Thus, O_l is norm-preserving and $\boldsymbol{c}_l$ is either zero or non-zero, depending on whether $\mathcal{N}_l$ is unital or non-unital, respectively (Lemma 2 and Lemma 4). The dilation part of M_l contracts the vector it acts on. Let q_l be the contractivity factor associated with Ω_l, i.e., for any vector $\boldsymbol{v}$,

$$\|\Omega_l \boldsymbol{v}\| = q_l \|\boldsymbol{v}\|, \quad 0 \leq q_l < 1. \tag{42}$$

Expanding the recursion given by Eq. (41), with the initial condition $\boldsymbol{v}_0$ corresponding to ρ_0, we obtain, with $\boldsymbol{d}_1 \equiv \boldsymbol{c}_1$ and $l \leq L$:

$$\boldsymbol{v}_j = \Omega_j \cdots \Omega_1 \boldsymbol{v}_0 + \boldsymbol{d}_j \ , \qquad 1 \leq j \leq L \tag{43a}$$

$$\boldsymbol{d}_j = \Omega_j \cdots \Omega_2 \boldsymbol{c}_1 + \Omega_j \cdots \Omega_3 \boldsymbol{c}_2 + \cdots + \Omega_j \boldsymbol{c}_{j-1} + \boldsymbol{c}_j \tag{43b}$$

$$= \sum_{r=1}^{j-1} \left(\prod_{s=j}^{r+1} \Omega_s \right) \boldsymbol{c}_r + \boldsymbol{c}_j, \quad 1 \leq j \leq L. \tag{43c}$$

Note that $\boldsymbol{d}_j$ is entirely a property of the maps and does not depend on the system state $\boldsymbol{v}_j$. In other words, $\boldsymbol{d}_j$ contains no useful information about the state of the computation carried out by the circuit. Using Eq. (43), we can write a transformed $\boldsymbol{v}$ at the end of the circuit as

$$\boldsymbol{v}_L = \Omega_L \cdots \Omega_1 \boldsymbol{v}_0 + \boldsymbol{d}_L, \tag{44}$$

where $\boldsymbol{d}_L$ arises from the noise map combined with the unitary VQA map. Substituting this into Eq. (38) with $\boldsymbol{v} \equiv \boldsymbol{v}_L$, we have:

$$C(\boldsymbol{\theta}) = \frac{1}{d}\mathrm{Tr}(H) + \Omega_L \cdots \Omega_1 \boldsymbol{v}_0 \cdot \boldsymbol{h} + \boldsymbol{d}_L \cdot \boldsymbol{h}. \tag{45}$$

Grouping together the terms that do not get contracted over the layers of the VQA circuit, we find:

$$|C(\boldsymbol{\theta}) - \frac{1}{d}\mathrm{Tr}(H) - \boldsymbol{d}_L \cdot \boldsymbol{h}| = q^L |\boldsymbol{v}_0 \cdot \boldsymbol{h}| \tag{46a}$$

$$\leq q^L \|\boldsymbol{v}_0\| \|\boldsymbol{h}\| \tag{46b}$$

$$\leq q^L \|\boldsymbol{h}\|, \tag{46c}$$

where

$$q^L \equiv q_1 \cdots q_L \ , \quad 0 \leq q_l < 1 \ \forall l. \tag{47}$$

Here q_l is the contractivity factor associated with Ω_l, so that the effective contractivity factor $q <$

1. Moreover, recall from Eq. (36) that $\|\boldsymbol{h}\| \leq \frac{h}{\sqrt{(K-1)!}} n^{K/2}$, where K (a constant) is the locality of H. We have thus proven that the cost function landscape concentrates:

Theorem 1. *The cost function of HS-contractive non-unital maps concentrates for any VQA circuit with greater than logarithmic depth, i.e., if* $L \in \omega[\log(n)]$ *then*[1]

$$|C(\boldsymbol{\theta}) - C_L| \leq \frac{h}{\sqrt{(K-1)!}} n^{K/2} q^L \tag{48a}$$

$$\to 0 \quad as \quad L \to \infty \tag{48b}$$

$$C_L \equiv \frac{1}{d}\mathrm{Tr}(H) + \boldsymbol{d}_L \cdot \boldsymbol{h} \tag{48c}$$

The meaning of the $L \in \omega[\log(n)]$ condition is that this result holds as long as L is large enough, i.e., except for circuits that are shallower than logarithmic depth, since then the $\|\boldsymbol{h}\|$ factor can counteract the q^L factor in Eq. (46c).

4.2 Noise-induced limit set

While Theorem 1 shows that the cost function concentrates on $C_\infty = \frac{1}{d}\mathrm{Tr}(H) + \boldsymbol{d}_\infty \cdot \boldsymbol{h}$, this does not mean that it tends to a fixed point. The reason is that the vector $\boldsymbol{d}_L$ is affected by the unitary transformations in the VQA circuit, which rotate it, and by noise in the circuit, which rotates and contracts it. The rotations prevent convergence on a fixed point, and instead C_∞ lies in an interval. We call this interval the *noise-induced limit set* (NILS). Our next result characterizes the NILS.

We can rewrite Eq. (43c) as:

$$\boldsymbol{d}_L = \Omega_L \cdots \Omega_2 \boldsymbol{c}_1 + \Omega_L \cdots \Omega_3 \boldsymbol{c}_2 + \cdots + \Omega_L \boldsymbol{c}_{L-1} + \boldsymbol{c}_L \tag{49a}$$

$$= p_1^{L-1}\Theta_1 \boldsymbol{c}_1 + p_2^{L-2}\Theta_2 \boldsymbol{c}_2 + \cdots + p_{L-1}\Theta_{L-1}\boldsymbol{c}_{L-1} + \boldsymbol{c}_L, \tag{49b}$$

where $\Pi_{l=i+1}^{L}\Omega_l \equiv p_i^{L-i}\Theta_i$, i.e., combining the transformations of $\boldsymbol{c}_i$ into an overall rotation Θ_i and an overall contraction p_i^{L-i}, where $0 \leq p_i < 1, \forall i$. Let $p \equiv \max_i p_i$, a noise contraction factor. Then:

[1]Recall the little omega notation; $f(x) \in \omega(g(x))$ means that for any positive constant c, there exist a real constant x_0 such that $f(x) > cg(x)\ \forall x \geq x_0$.

Proposition 2.

$$|\boldsymbol{d}_L \cdot \boldsymbol{h}| \leq \frac{1-p^L}{1-p}\frac{\|\boldsymbol{h}\|}{\sqrt{1-1/d}} \equiv \Lambda_L \leq \Lambda_\infty. \tag{50}$$

The proof is given in Appendix G.

Taking the limit $L \to \infty$, we can identify the *noise-induced limit set* (NILS) that the cost function concentrates in:

Corollary 2.

$$C_\infty \in [\frac{1}{d}\mathrm{Tr}(H) - \Lambda_\infty, \frac{1}{d}\mathrm{Tr}(H) + \Lambda_\infty] \equiv \mathrm{NILS}. \tag{51}$$

where

$$\Lambda_\infty = \frac{1}{1-p}\frac{\|\boldsymbol{h}\|}{\sqrt{1-1/d}}. \tag{52}$$

Proof. It follows from Eq. (50) that $-\Lambda_\infty \leq \boldsymbol{d}_L \cdot \boldsymbol{h} \leq \Lambda_\infty$; adding $\frac{1}{d}\mathrm{Tr}(H)$ to all sides gives $-\Lambda_\infty + \frac{1}{d}\mathrm{Tr}(H) \leq C_L \leq \Lambda_\infty + \frac{1}{d}\mathrm{Tr}(H)\ \forall L$. □

Thus, the NILS is an interval dictated by the dimension of the system, the problem Hamiltonian, and the noise through the contraction factor p. Note again, that C_∞ and $\boldsymbol{d}_\infty$ are notations we have chosen and are not meant to suggest that a limit should exist as a point. As we have shown, the limit is an *interval* that we call a noise-induced limit set.

4.3 The unital case

Since the unital case is the case for which $\boldsymbol{d}_L = \boldsymbol{0}$ [all the shift vectors $\boldsymbol{c}_j$ vanish in Eq. (43c)], we have:

Corollary 3. *The cost function of unital maps concentrates super-polynomially for any VQA circuit with logarithmic depth, i.e., if* $L = \omega[\log(n)]$*, and exponentially for any VQA circuit with at least linear depth, i.e., if* $L = \Omega(n)$*.*

The latter statement recovers the concentration result for unital maps of Ref. [31, Lemma 1], which holds "whenever the number of layers L scales linearly with the number of qubits", i.e., if $L = \omega(n)$ in our notation.

Note that the NILS in the case of HS-contractive non-unital noise is worse for VQA circuit performance than unital noise since knowledge of C_L requires a precise characterization of each of the intermediate non-unital maps (to determine $\boldsymbol{d}_L$), whereas in the unital case, the NILS

is determined purely by the target Hamiltonian H, and becomes a fixed point:

$$C_{\infty}^{\text{unital}} = \frac{1}{d}\text{Tr}(H) = \langle H \rangle_{I/d}, \tag{53}$$

i.e., the expectation value of the Hamiltonian with respect to the fully mixed state.

Next, we discuss the NIBP phenomenon for unital and HS-contractive non-unital maps and show that it appears for the former (in agreement with Ref. [31]) but not for the latter.

5 NIBP via parameter shift rules

From now on, when we use $C(\boldsymbol{\theta})$, we refer to the cost function in the presence of a noise map. In this section, we bound $|\partial C(\boldsymbol{\theta})/\partial\theta_\mu|$, the magnitude of the cost function gradient, and show that it is exponentially small in the circuit depth L for both unital and non-unital noise, for $n \geq 1$ qubits. We interchangeably use both lm and μ to denote the gate location.

Our proof in the unital case is simpler and more general than that of Ref. [31], but the HS-contractive non-unital case is our main new result. The bound we find has implications for the many applications of VQA where the goal is to learn the optimal parameters, i.e., when one is primarily concerned with trainability, and hence the gradient is a key quantity of interest.

At this point, it is useful to provide a formal definition of noise-induced barren plateaus. The following definition is inspired by Ref. [31, Theorem 1]:

Definition 2. *A cost function $C(\boldsymbol{\theta})$ exhibits a noise-induced barren plateau (NIBP) if the magnitude of its gradient, $\left|\frac{\partial C(\boldsymbol{\theta})}{\partial\theta_\mu}\right|$, decays exponentially as a function of the circuit depth L for all L larger than some constant $L_0 \geq 1$, independently of l and m, even for constant-width circuits.*

Thus, NIBPs flatten the entire control landscape independently of the location $\mu = (l, m)$ of the gate in the circuit at which the derivative is taken. Moreover, we impose the condition that the result holds even for constant-width circuits in order to preclude a measure concentration-type argument, which is typical of noise-free barren plateaus [24]. In this sense, NIBPs are distinct from the latter, for which the global minimum can be embedded inside a deep, narrow valley in the control landscape [26].

Using the PSR [Eq. (5)], and choosing the target operator to be a Hamiltonian H, we have:

$$\frac{\partial C(\boldsymbol{\theta})}{\partial\theta_\mu} = \frac{1}{2}\text{Tr}[H(\rho(\boldsymbol{\theta}_\mu^+) - \rho(\boldsymbol{\theta}_\mu^-))], \tag{54}$$

where $\boldsymbol{\theta}_\mu^\pm \equiv \boldsymbol{\theta} \pm \boldsymbol{\theta}_\mu^{\pi/2}$. We use Eq. (10) to write

$$\rho(\boldsymbol{\theta}_\mu^\pm) = \frac{1}{d}I + \sum_{i=1}^{d^2-1}(v_\mu^\pm)_i F_i = \frac{1}{\sqrt{d}}F_0 + \boldsymbol{v}_\mu^\pm \cdot \boldsymbol{F}\ , \tag{55}$$

so that

$$\rho(\boldsymbol{\theta}_\mu^\pm) = \frac{1}{\sqrt{d}}F_0 + \boldsymbol{v}_\mu^\pm \cdot \boldsymbol{F}. \tag{56}$$

Let

$$\tilde{\boldsymbol{v}}_\mu^L \equiv \boldsymbol{v}_\mu^+ - \boldsymbol{v}_\mu^-, \tag{57}$$

where we added the L subscript as a reminder that $\tilde{\boldsymbol{v}}_\mu^L$ corresponds to the difference between two states obtained at the end of the circuit. Using Eq. (37):

$$\frac{\partial C(\boldsymbol{\theta})}{\partial\theta_\mu} = \frac{1}{2}\text{Tr}(H\tilde{\boldsymbol{v}}_\mu^L \cdot \boldsymbol{F}) \tag{58a}$$

$$= \frac{1}{2}\sum_{i=1}^{d^2-1}(\tilde{\boldsymbol{v}}_\mu^L)_i \text{Tr}(HF_i) \tag{58b}$$

$$= \frac{1}{2}\sum_{i=1}^{d^2-1}(\tilde{\boldsymbol{v}}_\mu^L)_i h_i. \tag{58c}$$

We thus arrive at the key result that the magnitude of the cost function gradient can be written simply as:

$$\left|\frac{\partial C(\boldsymbol{\theta})}{\partial\theta_\mu}\right| = \frac{1}{2}\left|\tilde{\boldsymbol{v}}_\mu^L \cdot \boldsymbol{h}\right|, \tag{59}$$

which expresses the cost function gradient in terms of the overlap of the difference between two coherence vectors with the coordinates of the traceless component of the target Hamiltonian.

5.1 NIBP in the unital case

We now prove the existence of an NIBP for unital maps, in agreement with Ref. [31].

Theorem 2. *Assume that the maps $\{\mathcal{N}_l\}_{l=1}^L$ in the VQA circuits described by Eq. (40) are all unital but non-unitary and that the Hamiltonians generating the gates in the circuit and the control*

noise are K-local. Let n denote the circuit width. Assume, moreover, that either $L = c[\ln(n)]^Q$ with $Q > 1$, or $Q = 1$ and $K < 2c\ln(1/r)$, where $c > 0$ is a constant and $0 < r < 1$ is a contractivity factor associated with the maps maps $\{\mathcal{N}_l\}_{l=1}^L$ [defined in Eq. (63b)]. The cost function of such circuits exhibits an NIBP.

Proof. Using the Cauchy-Schwarz inequality, Eq. (59) yields:

$$\left|\frac{\partial C(\boldsymbol{\theta})}{\partial \theta_\mu}\right| \leq \frac{1}{2}\|\tilde{\boldsymbol{v}}_\mu^L\|\|\boldsymbol{h}\|, \tag{60}$$

Since the maps $\mathcal{N}_l$ in Eq. (40) are all unital, $\mathcal{N}_l \circ \mathcal{U}_l'$ is also unital since $\mathcal{U}_l'$ is unitary and hence unital. Let $\boldsymbol{v}_l$ be the coherence vector corresponding to ρ_l in Eq. (39). It follows from Lemma 2 that $\|\boldsymbol{v}_{l+1}\| = r_l\|\boldsymbol{v}_l\|$, where $0 \leq r_l < 1$ $\forall l$ is the contractivity factor associated with $\mathcal{N}_l$. Since $\boldsymbol{v}_\mu^\pm$ are the coherence vectors corresponding to $\rho(\boldsymbol{\theta}_\mu^\pm)$ – the states obtained at the end of the circuit but with $\boldsymbol{\theta}$ shifted by $\boldsymbol{\theta}_\mu^\pm$ – the same reasoning applies, and we can write:

$$\|\boldsymbol{v}_\mu^+\| = p^L\|\boldsymbol{v}_0\| \ , \quad \|\boldsymbol{v}_\mu^-\| = q^L\|\boldsymbol{v}_0\|, \tag{61}$$

where $\boldsymbol{v}_0$ is the coherence vector of ρ_0 (the initial density matrix), and

$$p^L \equiv p_1 \cdots p_L \ , q^L \equiv q_1 \cdots q_L \ , 0 \leq p_l, q_l < 1 \ \forall l. \tag{62}$$

Here p and q are the effective contractivity factors associated with the two paths involving Eq. (40) and $\boldsymbol{\theta}_\mu^+$ or $\boldsymbol{\theta}_\mu^-$, respectively. Using the elementary inequality $\|\boldsymbol{a} - \boldsymbol{b}\| \leq \|\boldsymbol{a}\| + \|\boldsymbol{b}\|$ we now have

$$\|\tilde{\boldsymbol{v}}_\mu^L\| = \|\boldsymbol{v}_\mu^+ - \boldsymbol{v}_\mu^-\| \leq \|\boldsymbol{v}_0\|(p^L + q^L) \leq 2r^L \tag{63a}$$

$$r \equiv \max(q, p), \tag{63b}$$

where we used $\|\boldsymbol{v}_0\| < 1$ [Eq. (12)]. Thus, using Eq. (60):

$$\left|\frac{\partial C(\boldsymbol{\theta})}{\partial \theta_\mu}\right| \leq \|\boldsymbol{h}\|r^L \ , \quad 0 \leq r < 1. \tag{64}$$

The noise channel in Eq. (40) already includes the situations with ideal gates [$\mathcal{U}$, Eq. (19)] and control noise in the gates [$\mathcal{U}'$, Eq. (26)]. To make this explicit, we can write:

$$\left|\frac{\partial C'(\boldsymbol{\theta})}{\partial \theta_\mu}\right|_{\text{ctrl}} \leq \|\boldsymbol{h}\|r_{\text{c}}^L \ , \quad 0 \leq r_{\text{c}} < 1. \tag{65}$$

We also model the random unitary noise which is a linear combination of unitary noise channels [Eq. (33b)]. Each summand can be modeled by Eq. (40), hence, is bounded as in Eq. (64):

$$\left|\frac{\partial C'(\boldsymbol{\theta})}{\partial \theta_\mu}\right|_{\text{rand}} \leq \sum_{k'} p'_{k'}\|\boldsymbol{h}\|r_{\text{r}}^L, \quad 0 \leq r_{\text{r}} < 1, \tag{66a}$$

which is still bounded since k does not scale exponentially with L.

Recalling Eq. (36), the magnitude of the gradient of the cost function [whether Eq. (64), Eq. (65), or Eq. (66)] satisfies the bound

$$\left|\frac{\partial C(\boldsymbol{\theta})}{\partial \theta_\mu}\right|, \left|\frac{\partial C'(\boldsymbol{\theta})}{\partial \theta_\mu}\right|_{\text{ctrl}}, \left|\frac{\partial C'(\boldsymbol{\theta})}{\partial \theta_\mu}\right|_{\text{rand}} \leq gn^{K/2}r^L, \tag{67}$$

where $g = g'g''$, and $g' = \frac{h}{\sqrt{(K-1)!}}$ and g'' are positive constants independent of n.

If $L = c[\ln(n)]^Q$ ($c > 0$), i.e., the circuit has sublogarithmic ($0 < Q < 1$), logarithmic ($Q = 1$), or superlogarithmic ($Q > 1$) depth, then $n = \exp[(L/c)^{1/Q}]$, so that

$$gn^{K/2}r^L = O\left(\exp[\frac{1}{2}K(L/c)^{1/Q} - \ln(1/r)L]\right). \tag{68}$$

This quantity decays exponentially provided $\frac{1}{2}K(L/c)^{1/Q} < \ln(1/r)L$. Solving this inequality for L, we see that for any $Q > 1$, this is again exponentially suppressed in the circuit depth L for all $L > L_0$, where

$$L_0 = c^{1-Q}\left(\frac{K/2}{\ln(1/r)}\right)^{\frac{Q}{Q-1}}. \tag{69}$$

In both cases the cost function gradient decays exponentially with the circuit depth L, i.e., we have an NIBP as per Definition 2. The circuit under extra control noise and random unitary noise could not escape NIBP if the original noisy circuit exhibits NIBP.

When $Q = 1$ (logarithmic circuit depth), the r.h.s. of Eq. (68) decays exponentially in L if $K < 2c\ln(1/r)$. This can be interpreted as an upper bound on the locality of the Hamiltonian and the control noise in terms of the largest contractivity factor (r) of the unital noise maps in the circuit. If this condition is satisfied then we again find an NIBP. □

Note that if $K \geq 2c\ln(1/r)$, or if $Q < 1$ (sublogarithmic circuit depth), then we cannot

conclude from our bounds that the circuit exhibits an NIBP. In other words, an NIBP may still occur but this cannot be inferred from our analysis.

5.2 The non-unital case

Now assume that at least one of the maps $\mathcal{N}_l$ in Eq. (40) is non-unital. We will show that in contrast to the unital case, in the non-unital case $\left|\frac{\partial C(\boldsymbol{\theta})}{\partial \theta_{lm}}\right|$ can be lower-bounded by a quantity that is non-vanishing even for arbitrarily deep circuits. This means that there is no guarantee of an NIBP in the non-unital case.

Recall Eq. (41). The effect of shifting $\boldsymbol{\theta}$ by $\boldsymbol{\theta}^{\pm}_{lm}$ at a single location $\mu = (l, m)$ in the circuit is that in layer l, and only in this layer, we have two different unitaries $\mathcal{U}'_l(\boldsymbol{\theta}^{\pm}_{lm})$, and correspondingly two different orthogonal rotations $O^{\pm}_{lm}$:

$$\boldsymbol{v}^{\pm,l}_{lm} = \Omega^{\pm}_{lm}\boldsymbol{v}_{l-1} + \boldsymbol{c}_l \ , \quad \Omega^{\pm}_{lm} \equiv M_l O^{\pm}_{lm}. \tag{70}$$

Note that prior to this location, i.e., for all $l' < l$, the bifurcation into the two paths labeled $\pm$ has not yet happened, which is why $\boldsymbol{v}_{l-1}$ does not carry a $\pm$ label.

5.2.1 No guarantee of an NIBP: an example

As a simple example that demonstrates why there is no guarantee of an NIBP in the non-unital case, assume that all $\mathcal{N}_l$ are unital except for the last two, i.e., $\mathcal{N}_l$ is non-unital only for $l = L-1$ and $l = L$. Writing the last two coherence vectors explicitly then gives:

$$\boldsymbol{v}^{\pm,L}_{Lm} = \Omega^{\pm}_{Lm}\boldsymbol{v}_{L-1} + \boldsymbol{c}_L \tag{71a}$$
$$\boldsymbol{v}_{L-1} = \Omega_{L-1}\boldsymbol{v}_{L-2} + \boldsymbol{c}_{L-1}, \tag{71b}$$

and substituting Eq. (71b) into Eq. (71a) yields:

$$\boldsymbol{v}^{\pm,L}_{Lm} = \Omega^{\pm}_{Lm}\Omega_{L-1}\boldsymbol{v}_{L-2} + \Omega^{\pm}_{Lm}\boldsymbol{c}_{L-1} + \boldsymbol{c}_L. \tag{72}$$

The term $\Omega^{\pm}_{Lm}\Omega_{L-1}\boldsymbol{v}_{L-2}$ is identical to the terms that appear in the unital case, so we know from the proof of Theorem 2 that its norm is $O\left(e^{-L}\right)$; therefore, for simplicity, let us neglect it entirely. Subtracting then yields:

$$\tilde{\boldsymbol{v}}^{L}_{Lm} = \boldsymbol{v}^{+,L}_{Lm} - \boldsymbol{v}^{-,L}_{Lm} = M_L(O^{+}_{Lm} - O^{-}_{Lm})\boldsymbol{c}_{L-1}. \tag{73a}$$

We know from Lemma 4 that $0 < \|\boldsymbol{c}_{L-1}\| < 1$. Thus, the vector $(O^{+}_{Lm} - O^{-}_{Lm})\boldsymbol{c}_{L-1}$ has a non-zero L-independent norm determined by the two different rotations $O^{\pm}_{Lm}$. Applying M_L to it can rescale its norm by, at most, a constant (L-independent) factor. Thus, the argument leading to the exponentially small upper bound on $\|\tilde{\boldsymbol{v}}^{L}_{Lm}\|$ in Eq. (63a) does not hold in this case. Instead, we now have, from Eq. (59):

$$\left|\frac{\partial C(\boldsymbol{\theta})}{\partial \theta_{Lm}}\right| = \frac{1}{2}\left|\tilde{\boldsymbol{v}}^{L}_{Lm} \cdot \boldsymbol{h}\right|. \tag{74}$$

It follows from standard Levy's lemma-type arguments that two randomly chosen unit vectors in $\mathbb{R}^D$ have overlap $\sim 1/\sqrt{D}$ (see, e.g., Ref. [58] for a variety of intuitive arguments). In our case, there is no *a priori* relation between $\tilde{\boldsymbol{v}}^{L}_{Lm}$ and $\boldsymbol{h}$ that would compel them to lie in some joint smaller-dimensional subspace, and since the Hamiltonian is K-local, the effective dimension D of $\boldsymbol{h}$ is $\sum_{k=1}^{K}\binom{n}{k} = O(n^K)$ as discussed in Section 3.4. This polynomially small overlap is, however, determined by the circuit width n rather than its depth L, so it will not be an NIBP as per Definition 2. See Appendix H for a proof sketch and simulation results. Of course, any circuit for which width and depth are related will impose a barren plateau-type result via Eq. (74) that does depend on L; this type of noise-free barren plateau is the one first demonstrated in Ref. [24].

5.2.2 No guarantee of an NIBP: general argument

We now show that the example above can be generalized, in particular without assuming that $l = L$. Specifically:

Theorem 3. *Assume that a circuit described by Eq. (40) contains any sequence of HS-contractive non-unital maps $\{\mathcal{N}_i\}_{i=1}^{l-1}$ for which $\max_{1\le i\le l-1}\sigma_{\max}(M_i) \in [0, \mu)$, where $\mu \in (0, 1/2]$, $l \ge 3$, and $\sigma_{\max}(M_i)$ is the largest singular value of M_i, the real rotation+dilation matrix associated with the map $\mathcal{N}_i$. Assume in addition that the last maps $\{\mathcal{N}_i\}_{i=l}^{L}$ in the circuit are all HS-contractive non-unital, $L - l = O(1)$ for large L, and $\{\sigma_{\min}(M_i) > c\}_{i=l}^{L}$, where $c > 0$ and $\sigma_{\min}(M_i)$ denotes the smallest singular value of M_i. The cost function of such a circuit does not exhibit an NIBP.*

Before we present the proof, we remark that Theorem 1 concerns the behavior of the *cost function* in circuits deeper than logarithmic depth ($L \in \omega[\log(n)]$), while Theorem 3 concerns the

behavior of *derivative of the cost function* in the last few layers of deep circuits (large L). The class of deep circuits in Theorem 3 is included in the class considered in Theorem 1 since deep circuits satisfy $L \in \omega[\log(n)]$. The concentration of the cost function in Theorem 1 does not prevent the non-concentration of the derivative in Theorem 3. This is because the cost function in Theorem 1 concentrates within a *finite interval* (the NILS; see Section 4.2), which means that a pair of cost functions $C(\boldsymbol{\theta}_1)$ and $C(\boldsymbol{\theta}_2)$ for circuits of the same depth L can have a finite difference. Using PSRs, this allows for scenarios where the derivative of the cost function in Theorem 3 remains finite, i.e., non-concentrated.

Proof. We start from Eq. (43). There is no additional bifurcation after the l'th layer, so that using Eq. (70), and the recursion given by Eq. (41) again, we obtain for $1 \leq j \leq L-l$:

$$\boldsymbol{v}_{lm}^{\pm,l+j} = \Omega_{l+j}\cdots\Omega_{l+1}\boldsymbol{v}_{lm}^{\pm,l} + \boldsymbol{d}_{l+j} \tag{75a}$$

$$\boldsymbol{d}_{l+j} = \Omega_{l+j}\cdots\Omega_{l+2}\boldsymbol{c}_{l+1} + \cdots + \Omega_{l+j}\boldsymbol{c}_{l+j-1} + \boldsymbol{c}_{l+j}. \tag{75b}$$

Note that it follows from Eq. (17b) applied to Eqs. (43a) and (75a) that $\|\boldsymbol{d}_j\| < 1$ for all $1 \leq j \leq L$.

Combining Eqs. (43) and (75) we obtain:

$$\boldsymbol{v}_{lm}^{\pm,L}$$

$$= \Omega_L\cdots\Omega_{l+1}\boldsymbol{v}_{lm}^{\pm,l} + \boldsymbol{d}_L \tag{76a}$$

$$= \Omega_L\cdots\Omega_{l+1}(\Omega_{lm}^{\pm}\boldsymbol{v}_{l-1} + \boldsymbol{c}_l) + \boldsymbol{d}_L \tag{76b}$$

$$= \Omega_L\cdots\Omega_{l+1}[\Omega_{lm}^{\pm}(\Omega_{l-1}\cdots\Omega_1\boldsymbol{v}_0 + \boldsymbol{d}_{l-1}) + \boldsymbol{c}_l] + \boldsymbol{d}_L \tag{76c}$$

$$= \underbrace{\Omega_L\cdots\Omega_{l+1}\Omega_{lm}^{\pm}\Omega_{l-1}\cdots\Omega_1\boldsymbol{v}_0}_{\boldsymbol{w}_{lm}^{\pm,L}} + \underbrace{\Omega_L\cdots\Omega_{l+1}\Omega_{lm}^{\pm}\boldsymbol{d}_{l-1}}_{\boldsymbol{e}_{lm}^{\pm,L}} + \underbrace{\Omega_L\cdots\Omega_{l+1}\boldsymbol{c}_l + \boldsymbol{d}_L}_{\boldsymbol{f}_L}. \tag{76d}$$

Let $\tilde{\boldsymbol{w}}_{lm}^L \equiv \boldsymbol{w}_{lm}^{+,L} - \boldsymbol{w}_{lm}^{-,L}$ and $\tilde{\boldsymbol{e}}_{lm}^L \equiv \boldsymbol{e}_{lm}^{+,L} - \boldsymbol{e}_{lm}^{-,L}$. Then, using Eq. (76):

$$\tilde{\boldsymbol{v}}_{lm}^L \equiv \boldsymbol{v}_{lm}^{+,L} - \boldsymbol{v}_{lm}^{-,L} = \tilde{\boldsymbol{w}}_{lm}^L + \tilde{\boldsymbol{e}}_{lm}^L. \tag{77}$$

We will show that $\tilde{\boldsymbol{w}}_{lm}^L$ can be neglected but $\tilde{\boldsymbol{e}}_{lm}^L$ cannot. First:

$$\|\tilde{\boldsymbol{w}}_{lm}^L\|$$

$$= \|\Omega_L\cdots\Omega_{l+1}(\Omega_{lm}^{+} - \Omega_{lm}^{-})\Omega_{l-1}\cdots\Omega_1\boldsymbol{v}_0\| \tag{78a}$$

$$\leq \|\Omega_L\cdots\Omega_{l+1}(\Omega_{lm}^{+} - \Omega_{lm}^{-})\Omega_{l-1}\cdots\Omega_1\|\|\boldsymbol{v}_0\| \tag{78b}$$

$$< \|\Omega_L\|\cdots\|\Omega_{l+1}\|\|\Omega_{lm}^{+} - \Omega_{lm}^{-}\|\|\Omega_{l-1}\|\cdots\|\Omega_1\| \tag{78c}$$

$$< 2p_L\cdots p_1 \tag{78d}$$

where the second line follows by definition of the operator norm, in the third line we used submultiplicativity and $\|\boldsymbol{v}_0\| < 1$, and in the last line we defined $p_l \equiv \|M_l\|$ and used $\|\Omega_{lm}^{+} - \Omega_{lm}^{-}\| = \|M_l(O_{lm}^{+} - O_{lm}^{-})\|$ and $\|(O_{lm}^{+} - O_{lm}^{-})\| \leq 2$ since $O_{lm}^{\pm}$ are orthogonal and the eigenvalues of any orthogonal matrix are all ± 1. We have $p_l < 1$ by Lemma 2 and Lemma 4. We may thus write

$$\|\tilde{\boldsymbol{w}}_{lm}^L\| < 2p^L\ , \quad p < 1, \tag{79}$$

where $p \equiv (\prod_{l=1}^{L} p_l)^{1/L}$. Thus, $\|\tilde{\boldsymbol{w}}_{lm}^L\|$ vanishes exponentially in the circuit depth.

Next, let us consider $\tilde{\boldsymbol{e}}_{lm}^L$. We can rewrite Eq. (43c) for $j = l-1$ as:

$$\boldsymbol{d}_{l-1} = \boldsymbol{c}_{l-1} + \sum_{l'=1}^{l-2}\prod_{i=l-1}^{l'+1}\Omega_i\boldsymbol{c}_{l'}\ , \quad l \geq 3\ , \tag{80}$$

where the order in the product reflects operator ordering.

Next, we lower-bound $\|\boldsymbol{d}_{l-1}\|$. As a simple example for which $\|\boldsymbol{d}_{l-1}\|$ is lower bounded by a positive constant, consider the case where only the last map is non-unital, and the rest are unital, i.e., $\boldsymbol{c}_{l-1} \neq \boldsymbol{0}$ but $\{\boldsymbol{c}_j = \boldsymbol{0}\}_{j=1}^{l-2}$. Then $\|\boldsymbol{d}_{l-1}\| = \|\boldsymbol{c}_{l-1}\| > 0$. To make the argument more general, we note that it follows from the triangle inequality $\|A+B\| \geq |\|A\| - \|B\||$ that:

$$\|\sum_{k=1} A_k\| \geq |\|A_1\| - \|\sum_{k=2} A_k\|| \tag{81a}$$

$$\geq \|A_1\| - \|\sum_{k=2} A_k\| \geq \|A_1\| - b, \tag{81b}$$

where $b \geq \|\sum_{k=2} A_k\|$. In the context of bounding $\|\boldsymbol{d}_{l-1}\|$, we identify

$$A_1 \equiv \boldsymbol{c}_{l-1}\ , \quad \{A_k\}_{k=2} \equiv \left\{\prod_{i=l-1}^{l'+1}\Omega_i\boldsymbol{c}_{l'}\right\}_{l'=1}^{l-2}. \tag{82}$$

Using the polar decomposition $M_i = V_i S_i$ with V_i orthogonal and $S_i = \sqrt{M_i^\dagger M_i}$ positive semidefinite and symmetric, we have

$$\|\Omega_i \boldsymbol{c}\| = \|V_i S_i O_i \boldsymbol{c}\| = \|S_i O_i \boldsymbol{c}\| \leq \lambda_{i,1}\|O_i \boldsymbol{c}\| = \lambda_{i,1}\|\boldsymbol{c}\|, \tag{83}$$

where, using Lemma 4, $\lambda_{i,1} = \sigma_{\max}(M_i) < 1$ is the largest singular value of M_i, and $\|\boldsymbol{c}\| < 1$. The same calculation, pulling out one factor from the left at a time, yields:

$$\|\prod_i \Omega_i \boldsymbol{c}\| \leq \prod_i \lambda_{i,1}\|\boldsymbol{c}\|. \tag{84}$$

Thus, letting

$$\tilde{\lambda}_l \equiv \max_{1\leq i\leq l-1} \lambda_{i,1} \ , \quad \tilde{c}_l \equiv \max_{1\leq l'\leq l-1} \|\boldsymbol{c}_{l'}\| \ , \tag{85}$$

we have:

$$\|\sum_{l'=1}^{l-2}\prod_{i=l-1}^{l'+1} \Omega_i \boldsymbol{c}_{l'}\| \leq \sum_{l'=1}^{l-2}\|\prod_{i=l-1}^{l'+1} \Omega_i \boldsymbol{c}_{l'}\| \tag{86a}$$

$$\leq \sum_{l'=1}^{l-2}\prod_{i=l'+1}^{l-1} \lambda_{i,1}\|\boldsymbol{c}_{l'}\| \leq \sum_{l'=1}^{l-2}\|\boldsymbol{c}_{l'}\|\tilde{\lambda}_l^{l-l'-1} \tag{86b}$$

$$\leq \tilde{c}_l r_l \ , \qquad r_l \equiv \frac{\tilde{\lambda}_l - \tilde{\lambda}_l^{l-1}}{1-\tilde{\lambda}_l} \ , \quad l \geq 3, \tag{86c}$$

where in the first inequality in Eq. (86b) we inverted the order in the product since $l'+1 \leq l-1$. Using Eqs. (80) to (82) we thus have:

$$\|\boldsymbol{d}_{l-1}\| \geq |\|\boldsymbol{c}_{l-1}\| - \|\sum_{l'=1}^{l-2}\prod_{i=l-1}^{l'+1} \Omega_i \boldsymbol{c}_{l'}\|| \tag{87a}$$

$$\geq \|\boldsymbol{c}_{l-1}\| - \tilde{c}_l r_l. \tag{87b}$$

We defined $\tilde{c}_l$ so that $\|\boldsymbol{c}_{l-1}\|$ is included in the maximum in Eq. (85). Therefore, if $\tilde{c}_l = \|\boldsymbol{c}_{l-1}\|$ and $r_l < 1$, then the r.h.s. of Eq. (87b) is positive. The condition $r_l < 1$ holds for all $\tilde{\lambda} \in [0, 1/2)$. If, instead, $\tilde{c}_l$ corresponds to $\max_{1\leq l'\leq l-1}\|\boldsymbol{c}_{l'}\|$ with $l' < l-1$, then we still have $\|\boldsymbol{d}_{l-1}\| > 0$, provided $r_l < \|\boldsymbol{c}_{l-1}\|/\tilde{c}_l$; this condition is satisfied for all $\tilde{\lambda}_l \in [0, \mu)$, where $\mu \in (0, 1/2)$ is found by solving the transcendental equation $r_l = \|\boldsymbol{c}_{l-1}\|/\tilde{c}_l$ for $\tilde{\lambda}_l$. Thus, we may conclude that a sufficient condition for $\|\boldsymbol{d}_{l-1}\| > C$, where $C > 0$ is a constant, is that the circuit contains any sequence of HS-contractive non-unital maps $\{\mathcal{N}_i\}_{i=1}^{l-1}$ for which $\max_{1\leq i\leq l-1}\sigma_{\max}(M_i) \in [0, \mu)$ and $l \geq 3$, where $\sigma_{\max}(M_i)$ is the largest singular value of M_i, i.e., the largest eigenvalue of the dilation $S_i = \sqrt{M_i^\dagger M_i}$ corresponding to $\mathcal{N}_i$.

Next, from Eq. (76d) we need to consider $\Omega_{lm}^{\pm}\boldsymbol{d}_{l-1} = M_l O_{lm}^{\pm}\boldsymbol{d}_{l-1}$. The two vectors $O_{lm}^{\pm}\boldsymbol{d}_{l-1}$ are separated by a distance $d_l = \|(O_{lm}^{+} - O_{lm}^{-})\boldsymbol{d}_{l-1}\|$ equal to the sum of the norms of the orthogonal vectors to their projections onto $\boldsymbol{d}_{l-1}$, i.e.,

$$d_l = 2\|\boldsymbol{d}_{l-1}\||\sin(\theta/2)| \tag{88a}$$

$$\cos\theta = \frac{(O_{lm}^{+}\boldsymbol{d}_{l-1})\cdot(O_{lm}^{-}\boldsymbol{d}_{l-1})}{\|\boldsymbol{d}_{l-1}\|^2} \ . \tag{88b}$$

The remaining transformations prescribed by Eq. (76d) are $\Omega_L \cdots \Omega_{l+1} M_l$, where $\Omega_i = M_i O_i$ and, using the polar decomposition again, $M_i = V_i S_i$ with V_i orthogonal and $S_i = \sqrt{M_i^\dagger M_i}$. The orthogonal rotations preserve the norms of $O_{lm}^{\pm}\boldsymbol{d}_{l-1}$; the dilations $\{S_i\}_{i=l}^{L}$ shrink these norms by at most the products of their smallest singular values, $\sigma_{\min}(M_i)$. Thus $d_l \mapsto \sigma_{\min}(M_l)\cdots\sigma_{\min}(M_L)d_l$, and as a result:

$$\|\tilde{\boldsymbol{e}}_{lm}^{L}\| \geq \sigma_{\min}(M_l)\cdots\sigma_{\min}(M_L)d_l \ . \tag{89}$$

The final step is to use Eqs. (77) and (79) and the triangle inequality to write

$$\|\tilde{\boldsymbol{v}}_{lm}^{L}\| = \|\tilde{\boldsymbol{w}}_{lm}^{L} + \tilde{\boldsymbol{e}}_{lm}^{L}\| \geq \|\tilde{\boldsymbol{e}}_{lm}^{L}\| - \|\tilde{\boldsymbol{w}}_{lm}^{L}\| \tag{90a}$$

$$\geq \sigma_{\min}(M_l)\cdots\sigma_{\min}(M_L)d_l - 2p^L. \tag{90b}$$

Therefore, as long as

$$\{\sigma_{\min}(M_i) > c\}_{i=l}^{L} \text{ and } L - l \in O(1), \tag{91}$$

then $\|\tilde{\boldsymbol{v}}_{lm}^{L}\| > C$ where $c, C > 0$ are both constants. This is because $2p^L$ is exponentially small in L, and $\sigma_{\min}(M_l)\cdots\sigma_{\min}(M_L)d_l$ is a product of $O(1)$ positive constants, i.e., itself a positive constant. Thus, $\|\tilde{\boldsymbol{v}}_{lm}^{L}\|$ is lower bounded by a positive constant for L sufficiently large: $L > \log(a^\kappa d_l/2)/\log(p)$, where $\kappa = L - l > 0$ is an $O(1)$ constant and $a = (\Pi_{i=l}^{L}\sigma_{\min}(M_i))^{\frac{1}{L-l}} > 0$ is another constant.

Reverting to Eq. (59), the fact that the lower bound on $\|\tilde{\boldsymbol{v}}_{lm}^{L}\|$ is now a positive constant for any l such that $L - l = O(1)$ then means that there is no NIBP in the HS-contractive non-unital case. As in the case of Eq. (74), the overlap in Eq. (59) could still be exponentially small (Levy's lemma), but as argued above, this is not noise-induced. □

Theorem 3 assumes that $\sigma_{\min}(M_i) > 0$. However, there exist non-unital noise channels for which $\sigma_{\min}(M_i) = 0$, e.g., a composite of a bit-flip or phase-flip channel with Kraus operators $\{\sqrt{p}I, \sqrt{1-p}\sigma\}$ where $\sigma \in \{X, Y, Z\}$, at the special symmetry point $p = 1/2$, followed by an amplitude damping channel in Eq. (92); see Appendix J. We next discuss the possibility of NIBPs in such cases.

When $\sigma_{\min}(M_i) = 0$, there is an in-principle possibility of encountering NIBPs in circuits subject to non-unital noise. To see why the proof of Theorem 3 does not hold in this case, note that $\|\tilde{\boldsymbol{v}}_{lm}^L\|$ in Eq. (90b) is not lower bounded by $C > 0$ if at least one of $\{\sigma_{\min}(M_i) > c\}_{i=l}^L$ is zero. Hence, there is no guarantee that the circuit will escape an NIBP since $\|\tilde{\boldsymbol{v}}_{lm}^L\|$ is not lower bounded above zero.

However, the contraction of $\boldsymbol{d}_l$ by $\sigma_{\min}(M_i)$ only happens when $\boldsymbol{d}_l$ is aligned in specific directions such that the noise channel maps $\boldsymbol{d}_l$ to the zero vector. As the number of qubits in the circuit grows, the dimension of $\boldsymbol{d}_l$ grows exponentially, making this scenario highly unlikely by a similar argument to the one below Eq. (74), using the standard Levy's lemma. Hence, we expect that NIBPs are avoided under non-unital noise even for channels for which one or more $\sigma_{\min}(M_i) = 0$.

5.2.3 Example: amplitude-damping

As a physical example of a HS-contractive non-unital map that prevents an NIBP, consider an amplitude-damping map. It suffices to consider the simple case of the amplitude-damping map for a qubit coupled to a zero-temperature bath. The Kraus operators are

$$K_0 = |0\rangle\langle 0| + \sqrt{1-p}|1\rangle\langle 1|, \quad K_1 = \sqrt{p}|0\rangle\langle 1|, \tag{92}$$

where p is the probability of relaxation from the excited state $|1\rangle$ to the ground state $|0\rangle$ [59]. We find, using Eq. (14), that $\boldsymbol{c} = (0, 0, p)$ and $M = \text{diag}(\sqrt{1-p}, \sqrt{1-p}, 1-p)$, so that $\sigma_{\max}(M) = \sqrt{1-p}$ and $\sigma_{\min}(M) = 1-p$. Thus, according to Theorem 3, for any $p \in [3/4+\varepsilon, 1-\varepsilon]$ and $\varepsilon > 0$, a sequence of L such amplitude-damping maps acting independently on each qubit of a noisy VQA circuit will prevent the cost function of such a circuit from exhibiting an NIBP.

6 Simulations

We performed numerical simulations employing the Qiskit framework [60] to ascertain the ground state energy of specific Hamiltonian under the influence of depolarizing (unital) and amplitude-damping (HS-contractive non-unital) noise. The single-qubit depolarizing map is defined as:

$$\mathcal{N}(\rho) = (1-p)\rho + \frac{p}{3}\sum_{\alpha\in\{x,y,z\}} \sigma^\alpha \rho \sigma^\alpha, \tag{93}$$

where p is the probability of the error. The single-qubit Kraus operators of the amplitude-damping map are given in Eq. (92).

The simulation was conducted utilizing three-qubit VQAs incorporating the `TwoLocal` ansatz, composed of $RY(\theta) = \exp(-i\theta\sigma^y/2)$ rotation gates and CNOTs in a linear entanglement configuration, where qubit i is entangled with qubit $i+1$ $\forall i$ along a chain, with open boundary conditions. The optimization procedure was performed using the Stochastic Perturbation Simulated Annealing (SPSA) algorithm [61] as the classical optimizer, constrained by a maximum iteration limit of 200 (`maxiter=200`). The multi-qubit noise map employed in the simulation was constructed from a composition of n one-qubit noise maps, which are HS-contractive for both of the examples we used to represent unital and non-unital maps.

We employ a stochastic procedure to generate a set of 50 sparse three-qubit Hamiltonians. These Hamiltonians are structured as $H = \sum_{i_1,i_2,i_3} h_{i_1 i_2 i_3}\sigma_{i_1}\otimes\sigma_{i_2}\otimes\sigma_{i_3}$, with the constraint that each element σ_{i_j} is drawn from the set $\sigma^0, \sigma^x, \sigma^z$ and interactions are restricted to be two-local, i.e., there is at least one σ_{i_j} that is σ^0. The magnitude of $h_{i_1 i_2 i_3}$ is uniformly sampled from $[0, 1)$ and later normalized such that $\|H\|_2 = 1$. The Hamiltonian for the main simulation was restricted to a maximum of three qubits to reduce the effect of non-noise-induced forms of BP.

In the simulation to demonstrate other types of BP that are not NIBP, we performed similar simulations using 50 randomly generated n-body Hamiltonians H_n, with $2 \le n \le 9$. We constructed these Hamiltonians so that each has a zero ground state energy and is at most two-local. These Hamiltonians are represented in the format $H_n = \sum_{\boldsymbol{i}} h_{\boldsymbol{i}}\sigma_{i_1}\otimes\cdots\otimes\sigma_{i_n}$, where σ_{i_j} belong to the set $\sigma_0, \sigma_x, \sigma_z$ and $\boldsymbol{i} = (i_1, \ldots, i_n)$.

The pseudo-code to generate random n-qubit Hamiltonians in this simulation is given in Algorithm 1.

procedure GENERATING A 2-LOCAL HAMILTONIAN
    Requirement: $H_0 = 0$
    Randomly generate a 2-local Hamiltonian H of n qubits such that $\|H\|_2 = 1$.
    Find its ground state H_0.
    Rescale the energy of H to have $H_0 = 0$ by subtracting the ground state energy.
    Renormalize H such that $\|H\|_2 = 1$.

Algorithm 1: Generating Hamiltonians for simulations

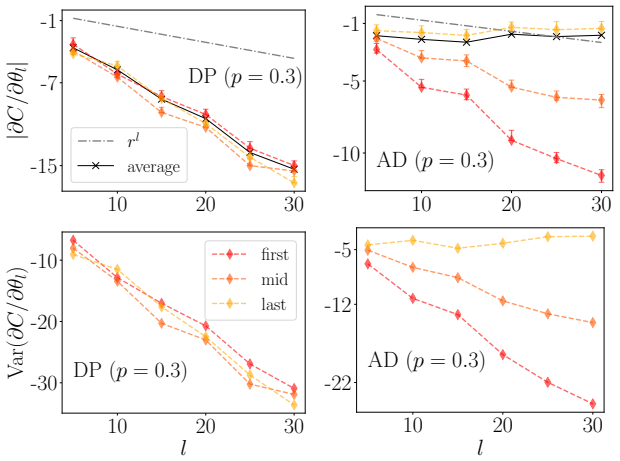

Figure 1: Mean and variance ($\log_{10}$ scale) of the magnitude of the cost function gradient for depolarizing (left) and amplitude-damping (right) maps with noise probability $p = 0.3$ as a function of layers. Error bars in the upper plots represent the range between the maximum and minimum values.

6.1 Demonstration of Theorem 2 and 3

NIBP refers to the phenomenon where the magnitude of the cost function's gradient with respect to control angles diminishes exponentially in the number of layers [Eq. (64)]. Our analysis demonstrates this characteristic to be consistently applicable solely within the unital noise scenario (Theorem 2). Conversely, in the case of HS-contractive non-unital noise, the magnitude of the gradient need not necessarily experience exponential suppression as a function of the number of layers (Theorem 3). We now present simulation results that support these results.

Fig. 1 illustrates the magnitude of the cost function gradient (on a logarithmic scale) as a function of layers of the VQAs. Mean values (variances) are plotted in the top (bottom) two plots. Plots on the left (right) correspond to VQAs under depolarizing noise (amplitude-damping noise), both with a noise probability of $p = 0.3$. We specifically present three angles corresponding to the initial, middle, and final layers of the VQA to emphasize their distinct behaviors. Additionally, the dot-dashed black line, denoted as r^l, is in proportion to the bound derived in Eq. (64) for the unital case. The solid black line is the numerical average over the three angles shown.

Under unital, depolarizing noise (upper and lower left), both the mean and variance of the gradient magnitude exhibit an exponential decay with an increasing number of layers. The error bars displayed in the upper plots represent the range between the maximum and minimum values. This suggests that all simulated mean values remain within the predicted bound. The observed behavior remains consistent irrespective of the specific layer within the VQA from which these angles were selected. The theoretical upper bound (dot-dashed line) is rather loose but consistent with the numerical results.

Under amplitude-damping noise (upper and lower right), distinct behaviors are observed among the three angles. The mean value of the magnitude of the gradient increases as the angle approaches the end of the VQA. The angle selected from the final layer consistently demonstrates a nearly constant gradient magnitude and violates the bound in Eq. (64), consistent with Theorem 3 and as anticipated in our theoretical analysis that angles within the terminal layers of the VQA circuit under HS-contractive non-unital noise may evade the NIBP.

Fig. 2 depicts an alternative view of Eq. (64), where the magnitude of the gradient is plotted as a function of the noise probability. The total number of layers in the VQA is fixed at $L = 20$. The noise probability is varied from $p = 0$ to $p = 0.5$, corresponding to decreasing the parameter r in Eq. (64). The behavior seen in Fig. 2 is similar to Fig. 1, where the magnitude of the gradient fully respects the derived bound only under depolarizing noise.

By fixing the noise probability and varying

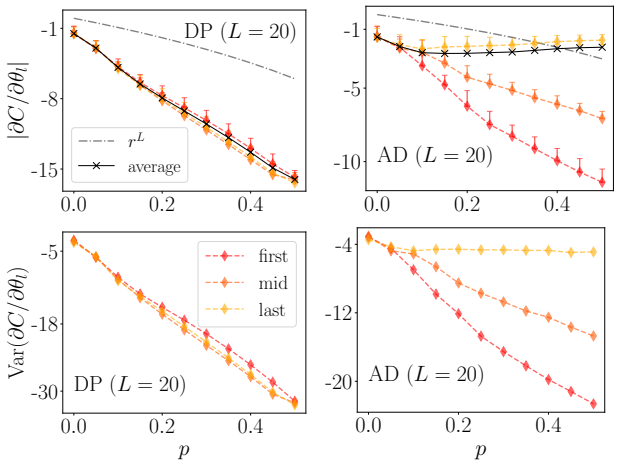

Figure 2: Mean and variance ($\log_{10}$ scale) of the magnitude of the cost function gradient for depolarizing (left) and amplitude-damping (right) maps as a function of noise probability p in a VQA with 20 layers. Error bars in the upper plots represent the range between the maximum and minimum values.

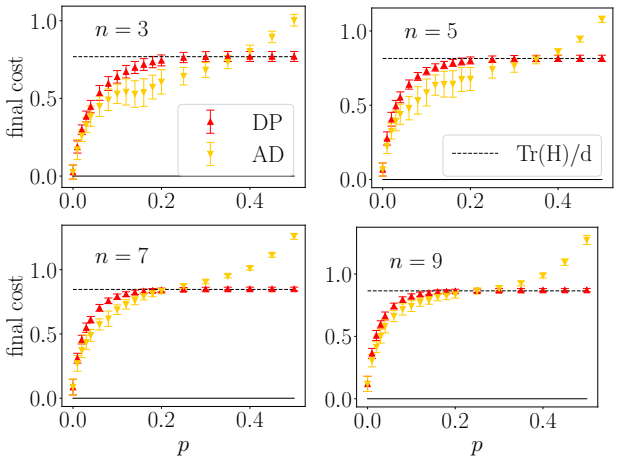

Figure 3: Final cost function averaged over 50 random n-qubit Hamiltonians with zero ground state energy under depolarizing (red up-triangles) and amplitude-damping (yellow down-triangles) maps as a function of noise probability p using VQA with $L = 5$ layers. The solid black line at zero denotes the true minimum of the Hamiltonians used in this simulation. The error bars are the standard deviation of the final cost. The dashed black line is the predicted NILS value in the large circuit depth limit in the unital case, Eq. (53).

the number of layers (Fig. 1) or fixing the number of layers and varying the noise probability (Fig. 2), our simulations consistently demonstrate that Eq. (64) is well-respected in VQAs under unital noise and violated in VQAs under HS-contractive non-unital noise. This observation aligns with our theoretical predictions in Theorem 2 and Theorem 3. For a numerical examination of the standard (noise-free) BP, see Appendix H.

6.2 Cost function behavior

Our results indicate that unital noise is subject to both NIBP and other forms of BP, and that, conversely, HS-contractive non-unital noise may potentially evade NIBP despite experiencing a comparable degree of other BPs. The primary question within the framework of VQA pertains to its efficacy in determining the ground state of a specific Hamiltonian. Consequently, in practical applications where the presence of noise is inevitable, the central concern revolves around identifying the type of noise that is comparatively less detrimental.

To address this question, we determined the final cost function obtained by VQA circuits when subjected to unital and HS-contractive non-unital noise maps. The result displayed in Fig. 3 was achieved through the training of circuits responsible for generating the results illustrated in Figs. 1 and 2, showing the average of the instances for n-qubit Hamiltonians for $n = 3, 5, 7$, and 9 as a function of noise probability p.

The most obvious result from these simulations is that noise has a rapidly increasing, strongly detrimental effect: the final cost function rises rapidly from its optimal value of zero as soon as noise is introduced, whether unital or non-unital. Note that additionally, even at $p = 0$, the final cost is greater than zero for $n \geq 5$ in our simulations, suggesting the effect of BPs.

The difference between unital and non-unital noise is small within the low-noise regime. The non-unital case exhibits a slightly better performance for small p, until reaching a crossover point typically observed within the range of $p = 0.15 - 0.3$ (depending on n). However, a notable divergence in their behavior becomes evident as the probability of noise increases. In the case of depolarizing noise, the cost function flattens out at approximately $p = 0.2$, whereas the cost function continues to rise when subjected to amplitude-damping noise, accelerating around $p = 0.3$. This observation seems consistent with our NILS result, where the cost function concentrates on a fixed point that is noise-independent in the unital case but noise-dependent in the HS-contractive non-unital case [Eq. (48c)]. Indeed,

the final cost function value precisely matches the theoretically predicted $\mathrm{Tr}(H)/d$ in the unital case [Eq. (53)].

Note that the theoretical prediction is concerned with the large L limit, not large p. However, while the simulations in Fig. 3 are for fixed $L = 5$, increasing p at fixed L is tantamount to increasing L at fixed p, as is clear from Eq. (48a), where the factor q^L is responsible for the NILS, and q is the contractivity factor, which is, of course, monotonic in p.

Another measure of noisy circuit quality is trainability: the variance of the cost function gradient should vanish no faster than $\Omega(1/\mathrm{poly}(n))$ [26]. Eq. (90b) implies that this is the case when $L - l = O(\log(n))$; see Appendix I for details. It may appear that this implies an advantage for non-unital noise over unital noise. However, being trainable only implies the potential to be trained, and does not guarantee that the circuit can, in practice, be trained to achieve the global minimum. This is illustrated in Fig. 3, where the final cost does not converge to zero for $p > 0$ due to the NILS.

7 Conclusions

This work expands the study of NIBPs to incorporate arbitrary unital and HS-contractive non-unital noise. Using a generalization of the parameter shift rule that includes noise, we have derived upper bounds for the scaling of the magnitude of the cost function gradient with respect to circuit width n, circuit depth L, and noise strength. In the unital case, we have shown that the onset of an NIBP occurs already for circuits of logarithmic depth (Theorem 2). In contrast, in the HS-contractive non-unital case, we have shown that VQA circuits need not necessarily exhibit an NIBP. This is true, in particular, when a constant number of final layers in a VQA circuit are subject to HS-contractive non-unital noise (Theorem 3).

We found that both unital and HS-contractive non-unital circuits exhibit a phenomenon we call a noise-induced limit set (NILS), whereby the cost function concentrates on a fixed value for circuits of greater than logarithmic depth. In the unital case, this is given by the expectation value of the problem Hamiltonian with respect to the fully mixed state [Eq. (53)], but in the HS-contractive non-unital case, the fixed value is determined by the parameters of the noise map as well [Eq. (48c)].

Our results are validated with numerical simulations for the depolarizing and amplitude-damping maps. Interesting open questions we do not address here are whether HS-contractive ($\|M\| < 1$) or non-HS-contractive ($\|M\| \geq 1$) non-unital noise maps appear more often in practical scenarios, how to characterize them according to a given set of Kraus operators, and what the measure of HS-contractive non-unital maps is in the space of all non-unital maps.

Overall, our work shows that NIBPs present a significant challenge for VQAs, even after mitigation of the standard (noiseless) BP problem. A combination of error suppression, mitigation, and correction methods will be necessary to realize the promise of VQAs, just as is the case for other quantum algorithms running on noisy quantum computers.

Note added. After this work was posted to the arXiv a related work appeared that addresses non-unital maps and barren plateaus [62]. The non-unital maps considered in this work are n-qubit tensor-products maps of n one-qubit non-unital channels. Our results hold for any n-qubit non-unital HS-contractive maps ($\|M\| < 1$), regardless of whether the map is a tensor-product of one-qubit channels.

Acknowledgments

This research was supported by the ARO MURI grant W911NF-22-S-0007. This research was developed with funding from the Defense Advanced Research Projects Agency under Agreement HR00112230006 and Agreement HR001122C0063. We thank Rubén Ibarrondo, Elias Zapusek and Samson Wang for useful comments, and especially Victor Kasatkin for numerous insightful discussions.

References

[1] Preskill, J. Quantum Computing in the NISQ era and beyond. *Quantum*, 2:79, 2018. DOI: 10.22331/q-2018-08-06-79.

[2] Cerezo, M. et al. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–

644, 2021. DOI: 10.1038/s42254-021-00348-9.

[3] Endo, S., Cai, Z., Benjamin, S.C. and Yuan, X. Hybrid quantum-classical algorithms and quantum error mitigation. *Journal of the Physical Society of Japan*, 90, 2021. DOI: 10.7566/JPSJ.90.032001.

[4] McClean, J.R., Romero, J., Babbush, R. and Aspuru-Guzik, A. The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics*, 18(2):023023, 2016. DOI: 10.1088/1367-2630/18/2/023023.

[5] Farhi, E., Goldstone, J. and Gutmann, S. A quantum approximate optimization algorithm. *arXiv preprint arXiv:1411.4028*, 2014. DOI: 10.48550/arXiv.1411.4028.

[6] Moll, N. et al. Quantum optimization using variational algorithms on near-term quantum devices. *Quantum Science and Technology*, 3(3):030503, 2018. DOI: 10.1088/2058-9565/aab822.

[7] Wang, Z., Hadfield, S., Jiang, Z. and Rieffel, E.G. Quantum approximate optimization algorithm for maxcut: A fermionic view. *Phys. Rev. A*, 97:022304, 2018. DOI: 10.1103/PhysRevA.97.022304.

[8] Li, J., Yang, X., Peng, X. and Sun, C.P. Hybrid quantum-classical approach to quantum optimal control. *Physical Review Letters*, 118(15):150503–, 2017. DOI: 10.1103/PhysRevLett.118.150503.

[9] Bravo-Prieto, C., LaRose, R., Cerezo, M., Subasi, Y., Cincio, L. and Coles, P.J. Variational Quantum Linear Solver. *Quantum*, 7:1188, 2023. DOI: 10.22331/q-2023-11-22-1188.

[10] Huang, H.Y., Bharti, K. and Rebentrost, P. Near-term quantum algorithms for linear systems of equations with regression loss functions. *New Journal of Physics*, 23, 2021. DOI: 10.1088/1367-2630/ac325f.

[11] Xu, X., Sun, J., Endo, S., Li, Y., Benjamin, S.C. and Yuan, X. Variational algorithms for linear algebra. *Science Bulletin*, 66(21):2181–2188, 2021. DOI: https://doi.org/10.1016/j.scib.2021.06.023.

[12] Koczor, B., Endo, S., Jones, T., Matsuzaki, Y. and Benjamin, S.C. Variational-state quantum metrology. *New Journal of Physics*, 22, 2020. DOI: 10.1088/1367-2630/ab965e.

[13] Meyer, J.J., Borregaard, J. and Eisert, J. A variational toolbox for quantum multiparameter estimation. *npj Quantum Information*, 7(1):89, 2021. DOI: 10.1038/s41534-021-00425-y.

[14] Khatri, S., LaRose, R., Poremba, A., Cincio, L., Sornborger, A.T. and Coles, P.J. Quantum-assisted quantum compiling. *Quantum*, 3:140, 2019. DOI: 10.22331/q-2019-05-13-140.

[15] Sharma, K., Khatri, S., Cerezo, M. and Coles, P.J. Noise resilience of variational quantum compiling. *New Journal of Physics*, 22, 2020. DOI: 10.1088/1367-2630/ab784c.

[16] Johnson, P.D., Romero, J., Olson, J., Cao, Y. and Aspuru-Guzik, A. Qvector: an algorithm for device-tailored quantum error correction. *arXiv preprint arXiv:1711.02249*, 2017. DOI: 10.48550/arXiv.1711.02249.

[17] Xu, X., Benjamin, S.C. and Yuan, X. Variational circuit compiler for quantum error correction. *Phys. Rev. Applied*, 15:034068, 2021. DOI: 10.1103/PhysRevApplied.15.034068.

[18] Mitarai, K., Negoro, M., Kitagawa, M. and Fujii, K. Quantum circuit learning. *Physical Review A*, 98(3):032309, 2018. DOI: 10.1103/PhysRevA.98.032309.

[19] Farhi, E. and Neven, H. Classification with quantum neural networks on near term processors. *arXiv preprint arXiv:1802.06002*, 2018. DOI: 10.48550/arXiv.1802.06002.

[20] Peruzzo, A. et al. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications*, 5(1):4213, 2014. DOI: 10.1038/ncomms5213.

[21] Kandala, A. et al. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature*, 549:242-246, 2017. DOI: 10.1038/nature23879.

[22] Cerezo, M., Sharma, K., Arrasmith, A. and Coles, P.J. Variational quantum state eigensolver. *npj Quantum Information*, 8(1):113, 2022. DOI: 10.1038/s41534-022-00611-6.

[23] Biamonte, J. Universal variational quantum computation. *Phys. Rev. A*, 103: L030401, 2021. DOI: 10.1103/PhysRevA.103.L030401.

[24] McClean, J.R., Boixo, S., Smelyanskiy, V.N., Babbush, R. and Neven, H. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9(1):

4812, 2018. DOI: 10.1038/s41467-018-07090-4.

[25] Holmes, Z., Sharma, K., Cerezo, M. and Coles, P.J. Connecting ansatz expressibility to gradient magnitudes and barren plateaus. *PRX Quantum*, 3:010313, 2022. DOI: 10.1103/PRXQuantum.3.010313.

[26] Cerezo, M., Sone, A., Volkoff, T., Cincio, L. and Coles, P.J. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature Communications*, 12 (1):1791, 2021. DOI: 10.1038/s41467-021-21728-w.

[27] Ortiz Marrero, C., Kieferová, M. and Wiebe, N. Entanglement-induced barren plateaus. *PRX Quantum*, 2:040316, 2021. DOI: 10.1103/PRXQuantum.2.040316.

[28] Cerezo, M. and Coles, P.J. Higher order derivatives of quantum neural networks with barren plateaus. *Quantum Science and Technology*, 6, 2021. DOI: 10.1088/2058-9565/abf51a.

[29] Arrasmith, A., Cerezo, M., Czarnik, P., Cincio, L. and Coles, P.J. Effect of barren plateaus on gradient-free optimization. *Quantum*, 5:558, 2021. DOI: 10.22331/q-2021-10-05-558.

[30] Cervero Martín, E., Plekhanov, K. and Lubasch, M. Barren plateaus in quantum tensor network optimization. *Quantum*, 7: 974, 2023. DOI: 10.22331/q-2023-04-13-974.

[31] Wang, S. et al. Noise-induced barren plateaus in variational quantum algorithms. *Nature Communications*, 12(1):6961, 2021. DOI: 10.1038/s41467-021-27045-6.

[32] Arrasmith, A., Holmes, Z., Cerezo, M. and Coles, P.J. Equivalence of quantum barren plateaus to cost concentration and narrow gorges. *Quantum Science and Technology*, 7(4):045015, 2022. DOI: 10.1088/2058-9565/ac7d06.

[33] Fefferman, B., Ghosh, S., Gullans, M., Kuroiwa, K. and Sharma, K. Effect of nonunital noise on random-circuit sampling. *PRX Quantum*, 5:030317, 2024. DOI: 10.1103/PRXQuantum.5.030317.

[34] Breuer, H.P. and Petruccione, F. *The Theory of Open Quantum Systems*. Oxford University Press, 2002.

[35] Schuld, M., Bergholm, V., Gogolin, C., Izaac, J. and Killoran, N. Evaluating analytic gradients on quantum hardware. *Physical Review A*, 99, 2019. DOI: 10.1103/PhysRevA.99.032331.

[36] Schumann, M., Wilhelm, F.K. and Ciani, A. Emergence of noise-induced barren plateaus in arbitrary layered noise models. *Quantum Science and Technology*, 9(4):045019, 2024. DOI: 10.1088/2058-9565/ad6285.

[37] Volkoff, T. and Coles, P.J. Large gradients via correlation in random parameterized quantum circuits. *Quantum Science and Technology*, 6, 2021. DOI: 10.1088/2058-9565/abd891.

[38] Grant, E., Wossnig, L., Ostaszewski, M. and Benedetti, M. An initialization strategy for addressing barren plateaus in parametrized quantum circuits. *Quantum*, 3:214, 2019. DOI: 10.22331/q-2019-12-09-214.

[39] Zhang, K., Hsieh, M.H., Liu, L. and Tao, D. Toward trainability of quantum neural networks. *arXiv preprint arXiv:2011.06258*, 2020. DOI: 10.48550/arXiv.2011.06258.

[40] Pesah, A., Cerezo, M., Wang, S., Volkoff, T., Sornborger, A.T. and Coles, P.J. Absence of barren plateaus in quantum convolutional neural networks. *Phys. Rev. X*, 11:041011, 2021. DOI: 10.1103/PhysRevX.11.041011.

[41] Patti, T.L., Najafi, K., Gao, X. and Yelin, S.F. Entanglement devised barren plateau mitigation. *Physical Review Research*, 3(3): 033090, 2021. DOI: 10.1103/PhysRevResearch.3.033090.

[42] Bharti, K. and Haug, T. Iterative quantum-assisted eigensolver. *Phys. Rev. A*, 104:L050401, 2021. DOI: 10.1103/PhysRevA.104.L050401.

[43] Cichy, S., Faehrmann, P.K., Khatri, S. and Eisert, J. Perturbative gadgets for gate-based quantum computing: Nonrecursive constructions without subspace restrictions. *Phys. Rev. A*, 109:052624, 2024. DOI: 10.1103/PhysRevA.109.052624.

[44] Wiersema, R., Zhou, C., Carrasquilla, J.F. and Kim, Y.B. Measurement-induced entanglement phase transitions in variational quantum circuits. *SciPost Phys.*, 14:147, 2023. DOI: 10.21468/SciPostPhys.14.6.147.

[45] Mele, A.A., Mbeng, G.B., Santoro, G.E., Collura, M. and Torta, P. Avoiding barren plateaus via transferability of smooth solutions in a hamiltonian variational ansatz.

Phys. Rev. A, 106:L060401, 2022. DOI: 10.1103/PhysRevA.106.L060401.

[46] Liu, L., Song, T., Sun, Z. and Lei, J. Quantum generative adversarial networks based on rényi divergences. *Physica A: Statistical Mechanics and its Applications*, 607:128169, 2022. DOI: https://doi.org/10.1016/j.physa.2022.128169.

[47] Rosenberg, E., Ginsparg, P. and McMahon, P.L. Experimental error mitigation using linear rescaling for variational quantum eigensolving with up to 20 qubits. *Quantum Science and Technology*, 7:015024, 2022. DOI: 10.1088/2058-9565/ac3b37.

[48] Czarnik, P., Arrasmith, A., Coles, P.J. and Cincio, L. Error mitigation with Clifford quantum-circuit data. *Quantum*, 5:592, 2021. DOI: 10.22331/q-2021-11-26-592.

[49] Wang, S., Czarnik, P., Arrasmith, A., Cerezo, M., Cincio, L. and Coles, P.J. Can Error Mitigation Improve Trainability of Noisy Variational Quantum Algorithms? *Quantum*, 8:1287, 2024. DOI: 10.22331/q-2024-03-14-1287.

[50] Liu, J., Wilde, F., Mele, A.A., Jiang, L. and Eisert, J. Stochastic noise can be helpful for variational quantum algorithms. *arXiv preprint arXiv:2210.06723*, 2023. DOI: 10.48550/arXiv.2210.06723.

[51] Lee, J., Huggins, W.J., Head-Gordon, M. and Whaley, K.B. Generalized unitary coupled cluster wave functions for quantum computation. *Journal of Chemical Theory and Computation*, 15(1):311–324, 2019. DOI: 10.1021/acs.jctc.8b01004.

[52] Cao, Y. et al. Quantum chemistry in the age of quantum computing. *Chemical Reviews*, 119(19):10856–10915, 2019. DOI: 10.1021/acs.chemrev.8b00803.

[53] Gell-Mann, M. Symmetries of baryons and mesons. *Phys. Rev.*, 125:1067–1084, 1962. DOI: 10.1103/PhysRev.125.1067.

[54] Stover, C. Generalized Gell-Mann Matrix. url: https://mathworld.wolfram.com/GeneralizedGell-MannMatrix.html

[55] Kraus, K. *States, Effects and Operations*. Fundamental Notions of Quantum Theory. Springer Berlin, Heidelberg, 1983.

[56] Nielsen, M.A. and Chuang, I.L. *Quantum computation and quantum information*. Cambridge University Press, 2010.

[57] Pérez-García, D., Wolf, M.M., Petz, D. and Ruskai, M.B. Contractivity of positive and trace-preserving maps under lp norms. *Journal of Mathematical Physics*, 47(8):083506, 2006. DOI: 10.1063/1.2218675.

[58] Why are two "random" vectors in $\mathbb{R}^n$ approximately orthogonal for large n? url: https://mathoverflow.net/questions/248466/why-are-two-random-vectors-in-mathbb-rn-approximately-orthogonal-for-large.

[59] Lidar, D.A. Lecture notes on the theory of open quantum systems. *arXiv preprint arXiv:1902.00967*, 2020. DOI: 10.48550/arXiv.1902.00967.

[60] Treinish, M. et al. Qiskit/qiskit: Qiskit 0.38.0, 2022. DOI: 10.5281/zenodo.7080365.

[61] Spall, J.C. An overview of the simultaneous perturbation method for efficient optimization. *Technical report, Johns Hopkins APL technical digest*, 19(4):482-492, 1998. url: https://secwww.jhuapl.edu/techdigest/Content/techdigest/pdf/V19-N04/19-04-Spall.pdf

[62] Mele, A.A. et al. Noise-induced shallow circuits and absence of barren plateaus. *arXiv preprint arXiv:2403.13927*, 2024. DOI: 10.48550/arXiv.2403.13927.

[63] R. Bhatia. *Matrix Analysis*. Number 169 in Graduate Texts in Mathematics. Springer-Verlag, New York, 1997.

[64] Baumgartner, B. An inequality for the trace of matrix products, using absolute values. *arXiv preprint arXiv:1106.6189*, 2011. DOI: 10.48550/arXiv.1106.6189.

[65] Kasatkin, V., Gu, L. and Lidar, D.A. Which differential equations correspond to the lindblad equation? *Physical Review Research*, 5 (4):043163, 2023. DOI: 10.1103/PhysRevResearch.5.043163.

[66] Byrd, M.S. and Khaneja, N. Characterization of the positivity of the density matrix in terms of the coherence vector representation. *Phys. Rev. A*, 68:062322, 2003. DOI: 10.1103/PhysRevA.68.062322.

[67] Arora, S. Lecture 11: High dimensional geometry, curse of dimensionality, dimension reduction. url: https://www.cs.princeton.edu/courses/archive/fall13/cos521/lecnotes/lec11.pdf

A Proof of Eq. (12)

Using the purity condition $P \equiv \mathrm{Tr}\rho^2 \leq 1$ in Eq. (10), we also have

$$1 \geq P = \mathrm{Tr}\left[\left(\frac{1}{d}I + \boldsymbol{F}\cdot\boldsymbol{v}\right)^2\right] \tag{94a}$$

$$= \frac{1}{d} + \sum_{i,j=1}^{M} \mathrm{Tr}(F_iF_j)v_iv_j = \frac{1}{d} + \|\boldsymbol{v}\|^2 \ , \tag{94b}$$

i.e., $\|\boldsymbol{v}\| = \sqrt{P - 1/d}$, and Eq. (12) follows.

B Proof of Eqs. (13) and (14)

Using $\rho' = \mathcal{N}(\rho) = \sum_\alpha K_\alpha \rho K_\alpha^\dagger$ and the expansion $\rho = \frac{1}{d}I + \sum_i v_i F_i$, we have the following series of implications:

$$\frac{1}{d}I + \sum_i v_i' F_i = \sum_\alpha K_\alpha (\frac{1}{d}I + \sum_i v_i F_i) K_\alpha^\dagger \tag{95a}$$

$$\sum_i v_i' F_i = \frac{1}{d}\sum_\alpha K_\alpha K_\alpha^\dagger + \sum_{\alpha i} v_i K_\alpha F_i K_\alpha^\dagger - \frac{1}{d}I \tag{95b}$$

$$\sum_i v_i' \mathrm{Tr}(F_jF_i) = \frac{1}{d}\sum_\alpha \mathrm{Tr}(F_j K_\alpha K_\alpha^\dagger) + \sum_{\alpha i} v_i \mathrm{Tr}(F_j K_\alpha F_i K_\alpha^\dagger) \tag{95c}$$

$$v_j' = c_j + \sum_i M_{ji} v_i. \tag{95d}$$

Eqs. (13) and (14) now follows from Eq. (95).

C Proof of Lemma 1

Recall that the p-norm is defined in Eq. (8). The matrix Hölder inequality states that for $1 \leq a, b \leq \infty$ and $\frac{1}{a} + \frac{1}{b} = 1$ [63, 64] :

$$\langle A, B\rangle \leq \|A\|_a \|B\|_b. \tag{96}$$

An important special case for our purposes is $a = b = 2$, i.e.:

$$\langle A, B\rangle \leq \sqrt{\langle A, A\rangle\langle B, B\rangle}, \tag{97}$$

which is just the Cauchy-Schwarz inequality for matrices.

Any linear map $\Psi : \mathcal{B}(\mathcal{H}) \mapsto \mathcal{B}(\mathcal{H})$ on operators $X \in \mathcal{B}(\mathcal{H})$ can be written as $\Psi(X) = \sum_\alpha E_\alpha X E_\alpha'^\dagger$, where $\{E_\alpha, E_\alpha'\} \in \mathcal{B}(\mathcal{H})$. Its Hermitian conjugate

$$\langle \Psi^\dagger(X), Y\rangle = \langle X, \Psi(Y)\rangle, \tag{98}$$

can be written explicitly as $\Psi^\dagger(X) = \sum_\alpha E_\alpha^\dagger X E_\alpha'$ [65].

Therefore, if $\mathcal{N} = \{K_\alpha\}$ is a unital CPTP map, then so is $\mathcal{N}^\dagger$. The reason is that if $\mathcal{N}$ is unital then $\mathcal{N}(X) = \sum_\alpha K_\alpha X K_\alpha^\dagger$ and $\sum_\alpha K_\alpha K_\alpha^\dagger = \sum_\alpha K_\alpha^\dagger K_\alpha = I$. Thus $\mathcal{N}^\dagger(X) = \sum_\alpha K_\alpha^\dagger X K_\alpha$ has Kraus operators $\{L_\alpha = K_\alpha^\dagger\}$, and it immediately follows that $\sum_\alpha L_\alpha^\dagger L_\alpha = \sum_\alpha L_\alpha L_\alpha^\dagger = I$, i.e., also $\mathcal{N}^\dagger$ is a unital CPTP map. We can now prove Lemma 1.

Proof. Consider $\rho \in \mathcal{B}_+(\mathcal{H})$ and let $P = \mathrm{Tr}(\rho^2) = \langle \rho, \rho\rangle$ denote its purity. The purity P' of $\rho^{(1)} = \mathcal{N}(\rho) \equiv \mathcal{N}^{(0)}(\rho)$ can be written as:

$$P^{(1)} = \langle \rho^{(1)}, \rho^{(1)}\rangle = \langle \mathcal{N}(\rho), \mathcal{N}(\rho)\rangle = \langle \rho, \mathcal{N}^\dagger[\mathcal{N}(\rho)]\rangle \tag{99a}$$

$$= \langle \rho, \mathcal{N}^{(1)}(\rho)\rangle \ , \tag{99b}$$

where in the third equality we used Eq. (98). Here, $\mathcal{N}^{(1)} \equiv \mathcal{N}^{(0)\dagger} \circ \mathcal{N}^{(0)}$ is the composition of two CPTP maps, so $\mathcal{N}^{(1)}$ is itself a CPTP map. Thus $\rho^{(2)} = \mathcal{N}^{(1)}(\rho)$ is a quantum state [i.e., $\rho^{(2)} \in \mathcal{B}_+(\mathcal{H})$].

Define $\forall n \geq 1$ a sequence of quantum maps $\mathcal{N}^{(n+1)} \equiv \mathcal{N}^{(n)\dagger} \circ \mathcal{N}^{(n)}$, purities $P^{(n)} = \langle \rho^{(n)}, \rho^{(n)} \rangle$, and states $\rho^{(n+1)} = \mathcal{N}^{(n)}(\rho)$. Then, using the Cauchy-Schwarz inequality for $n \geq 1$:

$$P^{(n)} = \langle \mathcal{N}^{(n-1)}(\rho), \mathcal{N}^{(n-1)}(\rho) \rangle = \langle \rho, \mathcal{N}^{(n)}(\rho) \rangle \tag{100a}$$

$$\leq \langle \rho, \rho \rangle^{1/2} \langle \rho^{(n+1)}, \rho^{(n+1)} \rangle^{1/2} \tag{100b}$$

$$= P^{1/2} (P^{(n+1)})^{1/2} , \tag{100c}$$

Expanding this recursion, we obtain:

$$P^{(1)} \leq P^{1/2} P^{1/4} \cdots P^{1/2^n} (P^{(n+1)})^{1/2^n} . \tag{101}$$

The purity is lower bounded by that of the fully mixed state I/d, where $d = \dim \mathcal{H}$: $P(I/d) = \mathrm{Tr}[(I/d)^2] = 1/d$. Therefore $\forall n, d$ we have $1/d \leq P^{(n+1)} \leq 1$ and hence $\lim_{n\to\infty} (P^{(n+1)})^{1/2^n} = 1$. Thus, upon taking the limit $n \to \infty$ of Eq. (101) we obtain:

$$P' = P^{(1)} \leq P^{\sum_{n=1}^{\infty} 2^{-n}} = P . \tag{102}$$

Equality in Eq. (100b) holds for all ρ iff $\mathcal{N}^{(n)}(\rho) = \rho \; \forall n$, i.e., $\mathcal{N}^{(n)} = \mathcal{I}$, which means that in particular, after setting $n = 1$, $\mathcal{N}^{(0)\dagger} \circ \mathcal{N}^{(0)} = \mathcal{I}$, so that by definition $\mathcal{N}^{(0)} = \mathcal{N}$ must be a unitary superoperator: $\mathcal{N}(\rho) = \mathcal{U}(\rho) = U \rho U^\dagger$, where U is unitary. □

D Contractivity in the sense of [57]

In Section 2.5, we define the contractivity of a map in terms of the Hilbert-Schmidt norm. This is different from [57], whose contractivity definition and results we briefly summarize here.

A map $\mathcal{N}$ between metric spaces $\mathcal{A}$ and $\mathcal{B}$ is strictly contractive iff there exists $r < 1$ s.t. $\forall A, B \in \mathcal{A}$ we have $d_{\mathcal{B}}(\mathcal{N}(A), \mathcal{N}(B)) \leq r d_{\mathcal{A}}(A, B)$, where d is a distance function. If the above definition is satisfied with $r = 1$ then the map $\mathcal{N}$ is called "non-expansive".

Let $\mathcal{M}_n$ be the space of $n \times n$ matrices, $\|\mathcal{N}\|_{p-p} = \sup_{A \in \mathcal{M}_n} \|\mathcal{N}(A)\|_p / \|A\|_p$ the induced p-norm, and $\|A\|_p$ (as usual) the Schatten p-norm of A. A positive trace-preserving map $\mathcal{N} : \mathcal{M}_n \to \mathcal{M}_{n'}$ is contractive when $\|\mathcal{N}\|_{p-p} < 1$. For a non-unital map with $n = n'$, $\mathcal{N}$ is always non-contractive, i.e., $\|\mathcal{N}\|_{p-p} > 1$ [57]. The crucial difference from our case (and the main reason that this result does not contradict ours) is that $\mathcal{M}_n$ is allowed to contain the 0 matrix (in addition, $\mathcal{N}$ need not be completely positive), which is, of course, different from the space of valid quantum states. Indeed, the proof of the non-contractivity of non-unital maps is essentially to take $A = 0$ and $B = I$; since $\mathcal{N}$ is trace-preserving, $\mathrm{Tr}(\mathcal{N}(B)) = \mathrm{Tr}(B) = n$ and if $\mathcal{N}(B) \neq I$ its norm must be larger than B's.

E Proofs of Lemma 4 and Lemma 5

E.1 Proof of Lemma 4

Proof. To prove Eq. (17a) note that, by definition, non-unital maps do not preserve I, i.e., they do not preserve the maximally mixed state given by the coherence vector $\boldsymbol{v} = \mathbf{0}$. The transformation $\mathbf{0} \to M\mathbf{0} + \boldsymbol{c} = \boldsymbol{c}$ must be non-zero. Hence, $\boldsymbol{c} \neq \mathbf{0}$.

To prove Eq. (17b), note that

$$\|\boldsymbol{v}'\| = \|M\boldsymbol{v} + \boldsymbol{c}\| \leq \sqrt{1 - 1/d} \;\; \forall \boldsymbol{v} \text{ s.t. } \|\boldsymbol{v}\| \leq \sqrt{1 - 1/d}, \tag{103}$$

where we used Eq. (12). This must hold in particular for the maximally mixed state, i.e., when $\boldsymbol{v} = \mathbf{0}$. Hence, $\|\boldsymbol{c}\| \leq 1/\sqrt{1 - 1/d}$.

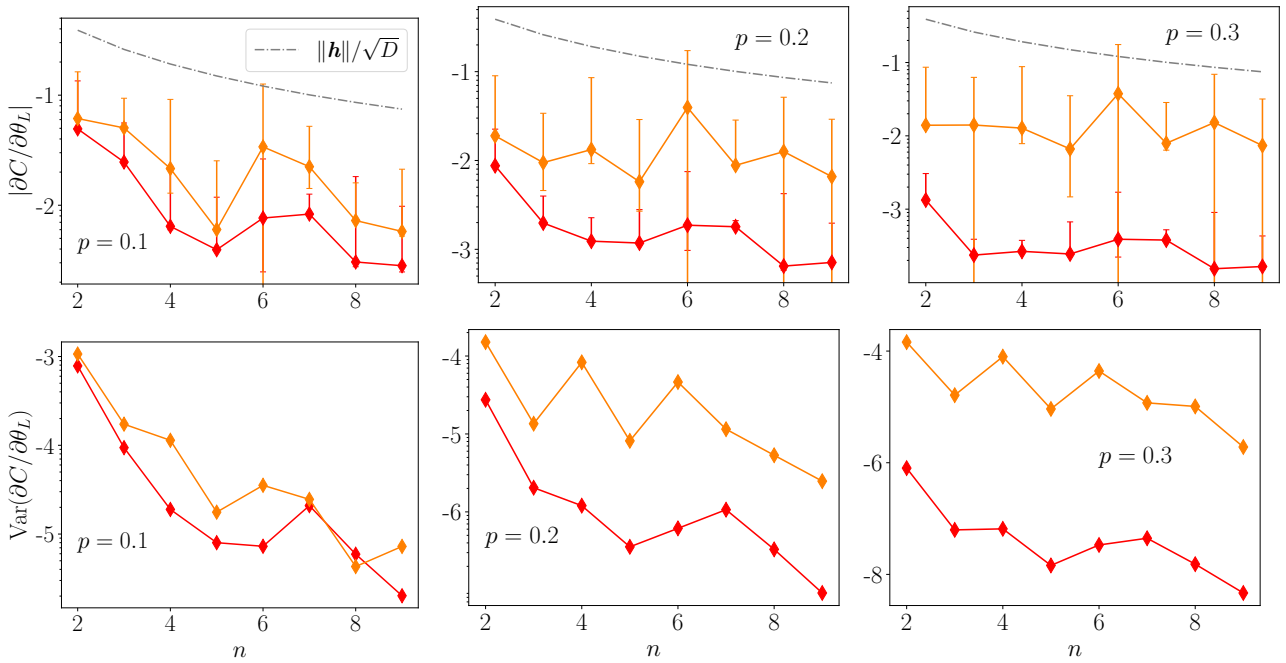

Figure 4: Magnitude and variance of the gradient of the cost functions for depolarizing (red) and amplitude-damping (orange) maps as a function of the number of qubits for noise probabilities $p = 0.1$, 0.2, and 0.3 in the left, middle, and right plots, respectively. Error bars in the top row represent the range of the values.

Since $\mathcal{N}$ is HS-contractive, Eq. (17c) follows directly from Lemma 3.

To prove Eq. (17d), recall the operator norm of M is its maximum singular value:

$$\sigma_{\max}(M) = \|M\| = \sup_{\boldsymbol{v}\neq\boldsymbol{0}} \frac{\|M\boldsymbol{v}\|}{\|\boldsymbol{v}\|}. \tag{104}$$

We use the definition in Eq. (104) and write $1 > \|M\| \geq \|M\boldsymbol{v}\|/\|\boldsymbol{v}\|, \forall \boldsymbol{v} \neq \boldsymbol{0}$, which implies that $\|M\boldsymbol{v}\| < \|\boldsymbol{v}\|, \forall \boldsymbol{v} \neq \boldsymbol{0}$. □

E.2 Proof of Lemma 5

To prove that a single-qubit non-unital map is always HS-contractive, we will show that it satisfies $\|M\| < 1$ and use Lemma 3 to complete the proof.

Proof. Let $\boldsymbol{v}_M$ be a coherence vector satisfying $\|M\| = \|M\boldsymbol{v}_M\|/\|\boldsymbol{v}_M\|$. Without loss of generality we can assume $\boldsymbol{v}_M$ corresponds to a pure state: if it does not, we can normalize it so that $\|\boldsymbol{v}_M\| = 1/\sqrt{2}$. We will prove the claim by contradiction and assume that $\|M\| \geq 1$, which implies $\|M\boldsymbol{v}_M\| = \|M\|\|\boldsymbol{v}_M\| \geq 1/\sqrt{2}$. Without loss of generality, let $(M\boldsymbol{v}_M)\cdot\boldsymbol{c} \geq 0$ (replace $\boldsymbol{v}_M$ by $-\boldsymbol{v}_M$ otherwise). This implies that $\|M\boldsymbol{v}_M + \boldsymbol{c}\| > \|M\boldsymbol{v}_M\| \geq 1/\sqrt{2}$. However, $M\boldsymbol{v}_M + \boldsymbol{c}$ is a valid quantum state which has $\|M\boldsymbol{v}_M + \boldsymbol{c}\| \leq 1/\sqrt{2}$. Hence, by contradiction, $\|M\| < 1$.

Using Lemma 3, $\|M\| < 1$ implies that the channel is HS-contractive. Hence, any single-qubit non-unital map is always HS-contractive. □

Note that while for $d = 2$ positivity is captured entirely by the condition $\|\boldsymbol{v}\| \leq 1/\sqrt{2}$, a condition on $\|\boldsymbol{v}\|$ alone is insufficient to ensure positivity for $d > 2$; additional constraints must be satisfied (see, e.g., Ref. [66, Eqs. (23) & (24)]). Since $\|\boldsymbol{v}_M\| \leq \sqrt{1 - 1/d}$ is not the only condition for $\boldsymbol{v}_M$ to represent a valid state, a coherence vector $\mathbf{v}_M$ such that $\|\boldsymbol{v}_M\| = \sqrt{1 - 1/d}$ could yield an invalid state that is non-positive. This invalidates the proof for general d. Hence, the proof above holds only for $d = 2$.

F Proof of Eq. (34)

Using Eq. (20c), the noisy cost function is:

$$C'(\boldsymbol{\theta}) = \mathrm{Tr}[\tilde{H}\mathcal{V}(\theta_\mu)(\tilde{\rho})] \tag{105a}$$

$$= p_j \mathrm{Tr}[\tilde{H}\mathcal{U}(\theta_\mu)(\tilde{\rho})] + \sum_{k\neq j} p_k \mathrm{Tr}[\tilde{H}\mathcal{V}'_k(\theta_\mu)(\tilde{\rho})]. \tag{105b}$$

Using Eq. (22), we now have:

$$\frac{\partial C'(\boldsymbol{\theta})}{\partial \theta_\mu} = p_j \frac{\partial C(\boldsymbol{\theta})}{\partial \theta_\mu} - \frac{i}{2}\sum_{k\neq j} p_k \mathrm{Tr}[\tilde{H}\mathcal{V}'(\theta_\mu)([P_k,\tilde{\rho}])] \tag{106a}$$

$$= p_j \frac{\partial C(\boldsymbol{\theta})}{\partial \theta_\mu} + \frac{1}{2}\sum_{k\neq j} p_k \mathrm{Tr}[\tilde{H}\mathcal{V}'(\theta_\mu)\times \tag{106b}$$

$$\left(\mathcal{U}_k\left(\frac{\pi}{2}\right)(\tilde{\rho}) - \mathcal{U}_k\left(-\frac{\pi}{2}\right)(\tilde{\rho})\right)]$$

$$= p_j \frac{\partial C(\boldsymbol{\theta})}{\partial \theta_\mu} + \frac{1}{2}\sum_{k\neq j} p_k \mathrm{Tr}(\tilde{H}\tilde{\xi}_k), \tag{106c}$$

where $\tilde{\xi}_k = \tilde{\rho}_k^+ - \tilde{\rho}_k^-$ and $\tilde{\rho}_k^\pm = \mathcal{V}'(\theta_\mu)\mathcal{U}_k\left(\pm\frac{\pi}{2}\right)(\tilde{\rho})$. Thus,

$$\left|\frac{\partial C'(\boldsymbol{\theta})}{\partial \theta_\mu}\right| \leq p_{j(\mu)}\left|\frac{\partial C(\boldsymbol{\theta})}{\partial \theta_\mu}\right| + \frac{1}{2}\sum_{k\neq j} p_{k(\mu)}|\tilde{\boldsymbol{w}}_{k(\mu)}\cdot\boldsymbol{h}|. \tag{107}$$

This directly yields the bound on the gradient given in Eq. (34).

G Proof of Eq. (50)

Proof. Let us reproduce Eq. (49b) for convenience:

$$\boldsymbol{d}_L = p_1^{L-1}\Theta_1\boldsymbol{c}_1 + p_2^{L-2}\Theta_2\boldsymbol{c}_2 + \cdots + p_{L-1}\Theta_{L-1}\boldsymbol{c}_{L-1} + \boldsymbol{c}_L.$$

Rotating a vector does not change its norm, i.e., $\|\Theta_i\boldsymbol{c}_i\| = \|\boldsymbol{c}_i\|$. Recall that $\|\boldsymbol{c}_i\| \leq 1/\sqrt{1-1/d}$ from Lemma 4. Thus, using $p \equiv \max_i p_i$ and $0 \leq p_i < 1, \forall i$:

$$\|\boldsymbol{d}_L\| = \|p_1^{L-1}\Theta_1\boldsymbol{c}_1 + \cdots + p_{L-1}\Theta_{L-1}\boldsymbol{c}_{L-1} + \boldsymbol{c}_L\| \tag{108a}$$

$$\leq \|p_1^{L-1}\Theta_1\boldsymbol{c}_1\| + \cdots + \|p_{L-1}\Theta_{L-1}\boldsymbol{c}_{L-1}\| + \|\boldsymbol{c}_L\| \tag{108b}$$

$$\leq \|p^{L-1}\Theta_1\boldsymbol{c}_1\| + \cdots + \|p\Theta_{L-1}\boldsymbol{c}_{L-1}\| + \|\boldsymbol{c}_L\| \tag{108c}$$

$$\leq (p^{L-1} + \cdots + p + 1)\max_l \|\boldsymbol{c}_l\| \tag{108d}$$

$$\leq (p^{L-1} + \cdots + p + 1)\frac{1}{\sqrt{1-1/d}} \tag{108e}$$

$$= \frac{1-p^L}{1-p}\frac{1}{\sqrt{1-1/d}}. \tag{108f}$$

Using the triangle inequality, we have

$$|\boldsymbol{d}_L\cdot\boldsymbol{h}| \leq \|\boldsymbol{d}_L\|\|\boldsymbol{h}\| \leq \frac{1-p^L}{1-p}\frac{\|\boldsymbol{h}\|}{\sqrt{1-1/d}} \leq \frac{1}{1-p}\frac{\|\boldsymbol{h}\|}{\sqrt{1-1/d}}, \tag{109}$$

which is Eq. (50). □

H Dependence on circuit width

In Eq. (74), we considered the overlap of two randomly chosen D-dimensional vectors, which suggests that the gradient of the cost function scales as $1/\sqrt{D}$. This phenomenon was initially discussed in Ref. [24] and is the original (noise-free) barren plateau (BP). To rederive it, we follow the approach of Ref. [67].

Consider two normalized D-dimensional vectors $\boldsymbol{v}$ and $\boldsymbol{h}$, and choose each component of $\boldsymbol{h}$ uniformly from the surface of a normalized D-Ball, i.e., $\|\boldsymbol{v}\| = \|\boldsymbol{h}\| = 1$. Without loss of generality, we can construct $\boldsymbol{h}$ such that its elements h_i are chosen randomly from $\{-1, 1\}/\sqrt{D}$, for $i \in [1, D]$. The expectation value of the inner product of $\boldsymbol{v}$ and $\boldsymbol{h}$ is $\mathbb{E}[\boldsymbol{v} \cdot \boldsymbol{h}] = \mathbb{E}[\sum_i v_i h_i] = 0$. The variance of their inner product is $\sigma^2[\boldsymbol{v} \cdot \boldsymbol{h}] = \mathbb{E}[\sum_{i,j} v_i h_i v_j h_j] - \mathbb{E}[\boldsymbol{v} \cdot \boldsymbol{h}]^2 = \sum_{i,j} v_i v_j \mathbb{E}[h_i h_j] = \sum_i v_i^2/D = 1/D$. The second to last equality uses $h_i^2 = 1/D$ and $\mathbb{E}[h_i h_j] = 0$ for $i \neq j$.

The Chernoff bound states that $\Pr(|X| > \epsilon) < \exp(-\frac{\epsilon^2}{\sigma[X]^2})$. Applying this bound to Eq. (74), we have $\Pr\left(2\left|\frac{\partial C(\boldsymbol{\theta})}{\partial \theta_{Lm}}\right| > \epsilon\right) = \Pr(|\boldsymbol{v} \cdot \boldsymbol{h}| > \epsilon) < e^{-D\epsilon^2}$. Substituting $\epsilon = 1/\sqrt{D}$, we obtain $\Pr(|\boldsymbol{v} \cdot \boldsymbol{h}| > 1/\sqrt{D}) < 1/e$, ensuring that it is likely that $|\boldsymbol{v} \cdot \boldsymbol{h}|$ stays below $1/\sqrt{D}$.

Note that since our n-qubit VQA Hamiltonian is 2-local, the effective dimension D of $\boldsymbol{h}$ is $\sum_{k=1}^{2} \binom{n}{k} = (n^2 + n)/2$ as discussed below Eq. (74). Thus, this result can be interpreted as stating that the cost function gradient scale as the inverse of the number of qubits (circuit width).

Next, we examine whether this dependence on circuit width can be observed in numerical simulations of the same type as discussed in Section 6. We again employ a set of 50 randomly chosen n-qubit Hamiltonians, with $2 \leq n \leq 9$. As stated in Algorithm 1, we set $\|H\|_2 = 1$; this ensures that $\|\boldsymbol{h}\| \leq 1$, i.e., does not grow with the effective dimension D of $\boldsymbol{h}$. Fig. 4 shows the result of the simulation. The two types of noise we simulated exhibit similar patterns, with amplitude-damping having a larger gradient magnitude and variance. Our simulation results exhibit a discernible trend along the dashed-dotted gray line denoted as $\|\boldsymbol{h}\|/\sqrt{D}$. The magnitude of the gradient (top row) appears to approach the scaling $1/\sqrt{(n^2+n)/2}$ as the noise probability increases. We emphasize that this does not constitute an upper bound for the magnitude; rather, it represents an expected value obtained through averaging over a large number of randomly generated vectors.

I Trainability of HS-contractive non-unital circuits

We will refer to a circuit as trainable when the variance of its cost function gradient vanishes no faster than $\Omega(1/\text{poly}(n))$ [26].

Using Eq. (90b) and letting $q = \min_{i \in [l,L]} \sigma_{\min}(M_i) > 0$, we write

$$\|\tilde{\boldsymbol{v}}_\mu^L\| \geq q^{L-l} d_l, \tag{110}$$

where $\mu = (l, m)$, d_l is a constant [recall the argument between Eqs. (87b) and (88)], and $0 < q < 1$ due to M_i being HS-contractive. We dropped the term $2p^L$ in Eq. (90b) since it becomes negligible in the large L limit. We can rewrite Eq. (59) as

$$|\partial_\mu C| = \frac{1}{2}|\tilde{\boldsymbol{v}}_\mu^L \cdot \boldsymbol{h}| \tag{111a}$$

$$= \frac{1}{2}\|\tilde{\boldsymbol{v}}_\mu^L\|\|\boldsymbol{h}\||\cos(\theta)| \tag{111b}$$

$$\geq q^{L-l}\epsilon. \tag{111c}$$

Here $\epsilon = d_l\|\boldsymbol{h}\||\cos(\theta)|$ and $\cos(\theta) = (\tilde{\boldsymbol{v}}_\mu^L \cdot \boldsymbol{h})/\|\tilde{\boldsymbol{v}}_\mu^L\|\|\boldsymbol{h}\|$ is the BP factor unrelated to noise, i.e., the overlap between two *normalized*, D-dimensional, random vectors that stays below $1/\sqrt{D}$ as argued in Appendix H. The difference is that the norm of $\tilde{\boldsymbol{v}}_\mu^L$ now scales down as q^{L-l}, so there is an additional noise-induced effect.

To account for the variance of the cost function gradient, we reinterpret Eq. (111) as a statement about an ensemble of random quantum circuits following the ansatz of the form in Eq. (1), so that $|\partial_\mu C|$ becomes a random variable. Then, recalling Chebyshev's inequality

$$\Pr(|\partial_\mu C| \geq \delta) \leq \frac{\mathrm{Var}[\partial_\mu C]}{\delta^2}, \tag{112}$$

we can use Eq. (111) to write

$$\mathrm{Var}[\partial_\mu C] \geq q^{2(L-l)}\epsilon^2. \tag{113}$$

Therefore, to satisfy the trainability condition $\mathrm{Var}[\partial_\mu C] \geq \Omega(1/\mathrm{poly}(n))$ we require

$$L - l = O(\log(n)). \tag{114}$$

Theorem 3 states that a circuit for which the last $O(1)$ layers are HS-contractive non-unital does not suffer from an NIBP. This is true also for the last $O(\log(n))$ layers since Eq. (90b) is still $1/\mathrm{poly}(n)$-small in this case. This implies that the cost function within the last $O(\log(n))$ HS-contractive non-unital layers is trainable, in qualitative agreement with Ref. [62], who used a different layer-independent constant.

J Example of non-unital channel outside Theorem 3

We expand on the example of a non-unital channel with $\sigma_{\max}(M) > 0$ and $\sigma_{\min}(M) = 0$, which was mentioned below the proof of Theorem 3. This kind of channel is expected to avoid NIBPs using a Levy's lemma-type argument.

Consider a composite of a bit flip (BF) channel followed by an amplitude damping (AD) channel. The Kraus operators of an AD channel are given in Eq. (92), while those of a BF channel are

$$K_0 = \sqrt{p}I, \quad K_1 = \sqrt{1-p}X, \tag{115}$$

where $1-p$ is the probability that a bit flip occurs. We can also find, using Eq. (14), that $\boldsymbol{c} = \boldsymbol{0}$ (as is true for all unital channels) and $M = \mathrm{diag}(1, 2p-1, 2p-1)$. When $p = 1/2$, we have $\sigma_{\max}(M) = 1$ and $\sigma_{\min}(M) = 0$.

Recall that a composition $\Phi(\rho) = \Psi_2(\Psi_1(\rho))$ of two CP maps $\Psi_1(\rho) = \sum_\alpha J_\alpha \rho J_\alpha^\dagger$ and $\Psi_2(\rho) = \sum_\beta K_\beta \rho K_\beta^\dagger$ is:

$$\Phi(\rho) = \sum_\gamma L_\gamma \rho L_\gamma^\dagger, \tag{116}$$

where $\gamma = (\alpha, \beta)$ and $L_\gamma = K_\beta J_\alpha$ are the Kraus operators of the composite channel.

Using Ψ_1 as a BF channel with $p = 1/2$ and Ψ_2 as an AD channel in Eq. (92), we can construct L_γ as

$$\begin{aligned} L_{(0,0)} &= \begin{pmatrix} 1/\sqrt{2} & 0 \\ 0 & \sqrt{(1-p)/2} \end{pmatrix} \quad L_{(1,0)} = \begin{pmatrix} 0 & \sqrt{p/2} \\ 0 & 0 \end{pmatrix} \\ L_{(0,1)} &= \begin{pmatrix} 0 & 1/\sqrt{2} \\ \sqrt{(1-p)/2} & 0 \end{pmatrix} \quad L_{(1,1)} = \begin{pmatrix} \sqrt{p/2} & 0 \\ 0 & 0 \end{pmatrix}. \end{aligned} \tag{117}$$

This set of Kraus operators corresponds to $M = \mathrm{diag}(\sqrt{1-p}, 0, 0)$ and $\boldsymbol{c} = (0, 0, p)$. We have composed a non-unital channel with $\sigma_{\max}(M) = \sqrt{1-p}$ and $\sigma_{\min}(M) = 0$.

Alternatively, we can derive this using the transformation of the coherence vector $\boldsymbol{v}$, by a BF channel with $p = 1/2$ followed by an AD channel:

$$\begin{aligned} \boldsymbol{v} &\to M_{\mathrm{AD}} M_{\mathrm{BF}}(\boldsymbol{v}) + \boldsymbol{c}_{\mathrm{AD}} \\ &= M\boldsymbol{v} + \boldsymbol{c}, \end{aligned} \tag{118}$$

where $M = M_{\mathrm{AD}} M_{\mathrm{BF}}$ and $\boldsymbol{c} = \boldsymbol{c}_{\mathrm{AD}}$ as expected.

Using Z in place of X in Eq. (115) to first construct a dephasing channel with $p = 1/2$ would also yield $\sigma_{\min}(M) = 0$, as would other unital channels at special values of their parameters.