

Optimal resource cost for error mitigation

Ryuji Takagi *

Nanyang Quantum Hub, School of Physical and Mathematical Sciences,

Nanyang Technological University, 637371 Singapore;

Center for Theoretical Physics and Department of Physics,

Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA;

and QunaSys Inc., Aqua Hakusan Building 9F, 1-13-7 Hakusan, Bunkyo, Tokyo 113-0001, Japan

 (Received 10 July 2020; revised 8 August 2021; accepted 10 August 2021; published 20 August 2021)

One of the central problems for near-term quantum devices is to understand their ultimate potential and limitations. We address this problem in terms of quantum error mitigation by introducing a framework taking into account the full expressibility of near-term devices, in which the optimal resource cost for the probabilistic error cancellation method can be formalized. We provide a general methodology for evaluating the optimal cost by connecting it to a resource-theoretic quantifier defined with respect to the noisy operations that devices can implement. We employ our methods to estimate the optimal cost in mitigating a general class of noise, where we obtain an achievable cost that has a generic advantage over previous evaluations, as well as a fundamental lower bound applicable to a broad class of noisy implementable operations. We improve our bounds for several noise models, where we give the exact optimal costs for the depolarizing and dephasing noise, precisely characterizing the overhead cost while offering an operational meaning to the resource measure in terms of error mitigation. Our result particularly implies that the heuristic approach presented by Temme *et al.* [K. Temme, S. Bravyi, and J. M. Gambetta, *Phys. Rev. Lett.* **119**, 180509 (2017)] is optimal even in our extended framework, putting fundamental limitations on the advantage provided by the extra degrees of freedom inherent in near-term devices for this noise model.

DOI: 10.1103/PhysRevResearch.3.033178

I. INTRODUCTION

Recent technological developments push us toward the realization of quantum information processing in a fully controlled manner, and a near-term cornerstone is to make use of noisy intermediate-scale quantum (NISQ) devices, which focus on manipulating tens to hundreds of qubits without a full error correction [1–3]. However, whether these near-term devices can provide advantages in useful applications is still elusive. In particular, rigorous theoretical analysis on the ultimate potential and limitations of NISQ devices has been largely missing.

In this work, we address this problem in terms of the capability to fight against noise. To deal with the critical noise effect without implementing an error-correcting code that is out of reach for the current technology, various error mitigation protocols have been proposed [4–9]. Among them, the probabilistic error cancellation method [4,8,10–15] stands as a promising candidate, as it can construct an unbiased estimator that faithfully estimates the expectation value of

an observable under a known error model. Probabilistic error cancellation is an application of a more general technique known as quasiprobability sampling [16–21], and the relevant resource cost is the sampling overhead—the number of samples necessary to ensure a certain accuracy—characterized by how much negative portion the quasiprobability has. Thus, the capability of a given noisy device to mitigate errors with this method can be characterized by the minimum negativity in the quasiprobability, providing the optimal resource cost at which the probabilistic error cancellation can be run.

However, the “optimal” cost is not well-defined by itself. Since quasiprobability decomposition expresses a given gate as a linear combination of noisy operations implementable on a noisy device, the optimal cost can highly depend on the choice of implementable operations. Many previous studies—not only on error mitigation but on NISQ algorithms in general—chose Pauli and Clifford gates as building blocks for the implementable operations [2]. As for the probabilistic error cancellation, Clifford operations combined with a target unitary gate followed by a noise channel were heuristically considered for depolarizing and amplitude damping noise [4], being extended to a complete set of Clifford-based operations applicable to a wider class of noise models [8]. However, the necessity of considering the Pauli and Clifford gates has been rarely asked. It surely makes sense to give them special status in the fault-tolerant quantum computation with error-correcting codes where *logical* operations take place in a code space, for which Clifford operations admit simple logical gate

*ryuji.takagi@ntu.edu.sg

constructions [22–24]. On the other hand, for NISQ devices that do not implement error-correcting codes, there is no clear reason for the Clifford gates to be preferred over non-Clifford gates at the level of *physical* operations on unencoded qubits. In this sense, NISQ devices are endowed with extra degrees of freedom, and this flexibility should be fully exploited so that their potential and limitations can be properly gauged.

Here, we introduce a framework that incorporates the full expressibility of noisy near-term devices that do not assume error correction. Our framework formalizes the optimal resource cost for the probabilistic error cancellation with respect to a continuous set of implementable operations, reflecting the flexibility of the NISQ devices. This consideration, however, also raises a demanding theoretical problem: since we need to take into account infinitely many implementable operations, obtaining the optimal resource cost can be extremely challenging. We address this problem by relating the optimal cost to a quantity studied in resource theories [25]. A major pillar of resource theories is the quantification of resources, and previous studies indicated intimate relations between quasiprobability representation and resource quantifiers [26–29]. In particular, a resource quantifier known as *robustness measure* [30] has found several applications in the context of simulating quantum circuits [18,19] and quantum memory [31]. We consider our set of implementable operations as the accessible free resource and find that the optimal mitigation cost can be characterized by the robustness measure defined in our resource-theoretic framework, which can be evaluated by leveraging tools in general convex resource theories [32,33].

Then, we employ our method to obtain universal bounds for optimal error mitigation cost for a general noise model, showing that our framework provides a generic advantage over previous evaluations based on a discrete set of implementable operations, while placing a fundamental lower bound that must be observed by any device whose implementable operations are subsumed by the one introduced in our framework. We also study several specific noise models, for which we find that our bounds can be improved. Notably, our methods provide *exact* optimal costs for depolarizing and dephasing noise channels, precisely characterizing the error mitigation capability of noisy devices while offering an operational meaning to the robustness measure in the context of error mitigation. Our result particularly indicates that the heuristic decomposition for the depolarizing channel given in Ref. [4] is still optimal even in our extended framework, putting fundamental limitations on the enhancement enabled by a continuous set of implementable operations for this specific noise model.

Our results not only provide a systematic way to evaluate the ultimate resource cost for error mitigation crucial for running useful algorithms in practice, but also present an application of ideas in resource theories to address concrete problems [18,34–38], paving the way for a rigorous information-theoretical analysis of noisy near-term devices.

II. PRELIMINARIES

A purpose of quantum computation, in particular for many variational algorithms designed for near-term devices, is to

obtain an expectation value $\langle A \rangle_{\text{ideal}} = \text{Tr}[A\rho_f]$, where A is some observable and $\rho_f = \mathcal{U}_L \circ \dots \circ \mathcal{U}_1(\rho_i)$ is a final quantum state with input state ρ_i fed into a quantum circuit composed of a sequence of unitary gates, $\{\mathcal{U}_i\}_{i=1}^L$. [The curly letter refers to a unitary gate as a quantum channel, i.e., $\mathcal{U}(\rho) := U\rho U^\dagger$]. However, since the application of quantum gates necessarily suffers from noise, the ideal gates $\{\mathcal{U}_i\}_{i=1}^L$ are not directly implementable on noisy devices. Instead, one can consider a set of noisy operations, $\mathcal{I}_{\mathcal{E}}$, for some noise channel \mathcal{E} that is implementable on the device of interest. The idea of the probabilistic error cancellation method is to represent each quantum gate as a linear combination of the noisy implementable operations as $\mathcal{U}_i = \sum_{\alpha} \eta_{i,\alpha} \mathcal{O}_{\alpha}$, $\mathcal{O}_{\alpha} \in \mathcal{I}_{\mathcal{E}}$, where $\eta_{i,\alpha}$ is a (not necessarily positive) real number. Then, for each gate \mathcal{U}_i , we sample an implementable operation \mathcal{O}_{α} at probability $|\eta_{i,\alpha}| / \sum_{\alpha} |\eta_{i,\alpha}|$, prepare a state $\tilde{\rho}_f = \mathcal{O}_{\alpha_L} \circ \dots \circ \mathcal{O}_{\alpha_1}(\rho_i)$ where \mathcal{O}_{α_i} is the implementable operation sampled for \mathcal{U}_i , and measure the observable A . Then, letting $\gamma_i := \sum_{\alpha} |\eta_{i,\alpha}|$, $\gamma_{\text{tot}} := \prod_{i=1}^L \gamma_i$, and $\text{sgn}_{\text{tot}} := \prod_{i=1}^L \text{sgn}(\eta_{i,\alpha_i})$, one can check that this realizes an unbiased estimator of $\langle A \rangle_{\text{ideal}}$ as $\langle A \rangle_{\text{ideal}} = \langle \gamma_{\text{tot}} \text{sgn}_{\text{tot}} \mu(A) \rangle_{\text{samp}}$, where $\mu(A)$ is a random variable for the measurement outcome and $\langle \cdot \rangle_{\text{samp}}$ refers to the expectation value for the sampling average taken for the above procedure.

Although it gives the desired expectation value, canceling the noise comes with a cost: one needs to pay more sampling cost than would be needed to estimate the desired expectation value with a noiseless circuit within the same accuracy. The Hoeffding's inequality [39] ensures that a sufficient number of samples used for estimating the true expectation value with error δ at probability $1 - \varepsilon$ is given by $(2\gamma_{\text{tot}}^2/\delta^2) \ln(2/\varepsilon)$. Thus, having small γ_{tot} is crucial to suppress the sampling overhead and, since γ_{tot} grows exponentially with respect to the number of gates, the problem is reduced to finding a good linear decomposition of each ideal gate \mathcal{U}_i with respect to implementable operations $\mathcal{I}_{\mathcal{E}}$ with small γ_i .

Clearly, the best linear decomposition depends on the choice of $\mathcal{I}_{\mathcal{E}}$, and it has been heuristically chosen on a case-by-case basis. For instance, for the depolarizing noise model $\mathcal{D}_{d,\epsilon}$ where d is the dimension of the system and ϵ is the noise strength, $\mathcal{I}_{\mathcal{D}_{d,\epsilon}} = \{\mathcal{D}_{d,\epsilon} \circ \mathcal{P} \circ \mathcal{U}\}$ where \mathcal{P} is a Pauli operator was considered, while for the single-qubit amplitude damping channel \mathcal{A}_{ϵ} , a set of implementable operations that works for the linear decomposition was found to be $\mathcal{I}_{\mathcal{A}_{\epsilon}} = \{\mathcal{A}_{\epsilon} \circ \mathcal{U}, \mathcal{A}_{\epsilon} \circ \mathcal{Z}^{1/2} \circ \mathcal{U}, \mathcal{A}_{\epsilon} \circ \mathcal{Z}^{-1/2} \circ \mathcal{U}, \mathcal{P}_{|0\rangle}\}$ where $\mathcal{Z}^{1/2}(\cdot) = Z^{1/2} \cdot Z^{1/2\dagger}$ with $Z^{1/2} := \text{diag}(1, i)$ the phase gate and $\mathcal{P}_{|0\rangle}$ the preparation of state $|0\rangle$ [4]. Later, this idea was extended to a Clifford-based universal basis set that works for any noise model with a sufficiently small noise strength [8].

III. FRAMEWORK

Although the above sets of operations can realize *some* decomposition, there is no guarantee that these choices of $\mathcal{I}_{\mathcal{E}}$ achieve the smallest overhead γ_i among other possible choices of the set of implementable operations. To assess the ultimate potential and limitations of the devices' capabilities having access to a continuous set of physical operations, we need to give more freedom to noisy devices with their programmability, i.e., the set of implementable operations under the noiseless

condition. Motivated by this observation, we introduce the set of implementable operations as

$$\mathcal{I}_{\mathcal{E}}(d) = \{\mathcal{E} \circ \Lambda \mid \Lambda \in \tilde{\mathcal{P}}(d)\} \quad (1)$$

for a given noise channel $\mathcal{E} \in \mathcal{T}(d)$, where $\mathcal{T}(d)$ is the set of completely positive trace-preserving (CPTP) maps with input and output systems being d -dimensional quantum systems, and $\tilde{\mathcal{P}}(d)$ is the set of programmable operations defined as

$$\tilde{\mathcal{P}}(d) = \left\{ \sum_i p_i \mathcal{V}_i \mid \mathcal{V}_i \in \mathcal{T}_u(d) \cup \mathcal{S}(d) \right\}, \quad (2)$$

where $\{p_i\}$ is a probability distribution, $\mathcal{T}_u(d)$ is the set of unitary channels on d -dimensional systems, and $\mathcal{S}(d) = \{\mathcal{P}_{|\psi\rangle} \mid |\psi\rangle \in \mathcal{H}_d\}$ with \mathcal{H}_d being the d -dimensional Hilbert space is the set of state preparation channels. The operations in $\mathcal{I}_{\mathcal{E}}(d)$ are implementable on the devices that can program any unitary gate and state preparation under the effect of noise \mathcal{E} , which is reasonable for small d such as $d = 2$ (single qubit) and $d = 4$ (two qubits), and even if a given device is not powerful enough to realize the above operations, our results serve as its ultimate bounds. Note that although different operations may come with different noise channels in general, here we take a fixed noise channel after operations with the same size, which is a standard assumption in the quantitative analysis of noise effects [4,24,40,41]; extending our formalism to accommodate different noise channels is left for future work.

IV. RESOURCE COST AS ROBUSTNESS

Our goal is to find a decomposition for a given ideal unitary gate $\mathcal{U} \in \mathcal{T}_u(d)$ with respect to the set of implementable operations (1) with the minimum absolute sum of the coefficients. Namely, the optimal overhead constant $\gamma_{\text{opt}}(\mathcal{U})$ is written as

$$\gamma_{\text{opt}}(\mathcal{U}) = \min \left\{ \sum_{\alpha} |\eta_{\alpha}| \mid \mathcal{U} = \sum_{\alpha} \eta_{\alpha} \mathcal{O}_{\alpha}, \mathcal{O}_{\alpha} \in \mathcal{I}_{\mathcal{E}}(d) \right\}, \quad (3)$$

where we assume $\gamma_{\text{opt}}(\mathcal{U}) < \infty$. As noted in Ref. [4], Eq. (3) becomes a linear program when $\mathcal{I}_{\mathcal{E}}(d)$ is a discrete set. However, because we aim to exploit the full expressibility of the device and consider the continuous set $\mathcal{I}_{\mathcal{E}}(d)$ given in (1), Eq. (3) is no longer simple linear programming; if one obtains some valid decomposition of \mathcal{U} , it is hard to see whether there exists another decomposition that gives a smaller overhead constant.

Here, we provide a general strategy to approach this problem. First, we introduce the following quantity:

$$R_{\mathcal{I}_{\mathcal{E}}}(\mathcal{U}) := \min \left\{ s \geq 0 \mid \frac{\mathcal{U} + s\Xi}{1+s} \in \mathcal{I}_{\mathcal{E}}, \Xi \in \mathcal{I}_{\mathcal{E}} \right\}. \quad (4)$$

This type of quantities defined for quantum states have been used to quantify the amount of quantum resources (e.g., entanglement) contained in the given state and is known as the (standard) robustness measure in resource theories. In particular, the quantity in (4) can be considered as the robustness measure defined for quantum operations in the context of the resource theory of channels [33,42,43], where $\mathcal{I}_{\mathcal{E}}$ serves as the

set of free channels. Then, one can show that (4) is equivalent to (3) with the relation

$$\gamma_{\text{opt}}(\mathcal{U}) = 2R_{\mathcal{I}_{\mathcal{E}}}(\mathcal{U}) + 1, \quad (5)$$

which we explain in Appendix A. (See also related arguments considered for other settings in Refs. [18,31,44]). Since $\mathcal{I}_{\mathcal{E}}(d)$ is a convex set, i.e., $\mathcal{O}_1, \mathcal{O}_2 \in \mathcal{I}_{\mathcal{E}}(d) \Rightarrow p\mathcal{O}_1 + (1-p)\mathcal{O}_2 \in \mathcal{I}_{\mathcal{E}}(d)$, $0 \leq p \leq 1$, we can bring up ideas and tools developed in general convex resource theories [32,33] and apply them to our resource measure in (4). In particular, we obtain the following dual form of the robustness (details are given in Appendix B):

$$R_{\mathcal{I}_{\mathcal{E}}}(\mathcal{U}) = \max \{ \text{Tr}[YJ_{\mathcal{U}}] - 1 \mid 0 \leq \text{Tr}[YJ_{\Xi}] \leq 1, \forall \Xi \in \mathcal{I}_{\mathcal{E}}, Y \in \mathbb{H} \}, \quad (6)$$

where $J_{\Lambda} := \text{id} \otimes \Lambda(d \cdot \Phi_d)$ with $\Phi_d = \frac{1}{d} \sum_{i,j=0}^{d-1} |ii\rangle\langle jj|$ is the Choi matrix of channel Λ and \mathbb{H} denotes the set of Hermitian operators. The above dual form is known to provide operational meanings to robustness measures and it particularly indicates that the robustness is physically observable [31,33]. Furthermore, these two expressions provide useful bounds for the optimal resource cost as

$$2\text{Tr}[YJ_{\mathcal{U}}] - 1 \leq \gamma_{\text{opt}}(\mathcal{U}) \leq 2s + 1, \quad (7)$$

where $Y \in \mathbb{H}$ and $s \geq 0$ are any Hermitian operator and real number satisfying the condition in (4) and (6). In addition, we show in Appendix C that if we find a decomposition $\mathcal{U} = \sum_{\alpha} \eta_{\alpha} \mathcal{O}_{\alpha} = \sum_{\alpha} \eta_{\alpha} \mathcal{E} \circ \mathcal{V}_{\alpha}$, which gives an upper bound in (7), then $Y = d^{-2} J_{\mathcal{E}^{-1} \circ \mathcal{U}}$ with $\mathcal{E}^{-1} := \sum_{\alpha} \eta_{\alpha} \mathcal{V}_{\alpha} \circ \mathcal{U}^{\dagger}$ satisfies the condition in (6) and provides $2\text{Tr}[\Phi_d \text{id} \otimes \mathcal{E}^{-1}(\Phi_d)] - 1$ as a candidate for a good lower bound.

V. EVALUATION OF RESOURCE COST

The generality of our method allows us to obtain bounds for the optimal resource cost for general noise channels. To see this, let us consider the noise channels of the form $\mathcal{E} = (1 - \epsilon)\text{id} + \epsilon_+ \Lambda - \epsilon_- \Xi$, where $\epsilon, \epsilon_{\pm} \geq 0$, $\Lambda, \Xi \in \tilde{\mathcal{P}}$, which represents a large class of channels and indeed any CPTP map on up to two-qubit systems. In Appendix D, we show the latter claim by introducing a universal set of basis operations for CPTP maps that solely consists of programmable operations in $\tilde{\mathcal{P}}$, which can be of independent interest. Although we are usually interested in the cases of small error ϵ , here we do not impose this assumption. This takes into account the fact that the decomposition is not unique, and the following theorem provides bounds for the optimal cost for any decomposition of this form. To state the result, let $I_{ij} = \{\vec{k} \in \{0, 1\}^i \mid \text{wt}(\vec{k}) = j\}$ denote the set of i -bit strings which have j 1's. Then, for $\vec{k} \in I_{ij}$, we define $(A^j, B^{i-j})_{\vec{k}}$ to be an operation that applies A for j times and B for $i - j$ times with a pattern specified by a binary string \vec{k} . For instance, if $\vec{k} = (0, 1, 1)$, then $(A^2, B^1)_{\vec{k}} = B \circ A \circ A$. Then, we get the following bounds that provide a systematic estimation of the optimal cost (the proof is given in Appendix E):

Theorem 1. For $\mathcal{E} = (1 - \epsilon)\text{id} + \epsilon_+ \Lambda - \epsilon_- \Xi$ where $\mathcal{E} \in \mathcal{T}(d)$, $\epsilon, \epsilon_{\pm} \geq 0$, and $\Lambda, \Xi \in \tilde{\mathcal{P}}(d)$, if $1 - \epsilon > \epsilon_+ + \epsilon_-$, then

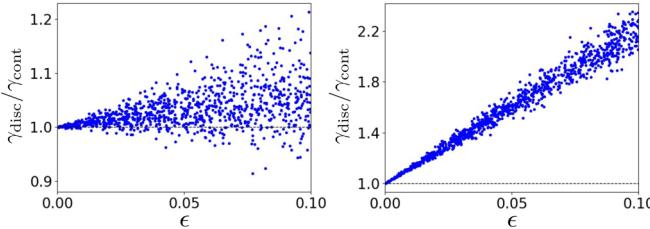


FIG. 1. The ratio of the resource cost obtained by the systematic decomposition with discrete basis operations in Ref. [8] to the achievable cost obtained in Theorem 1 with respect to the noise strength, where the error channel is chosen as $(1 - \epsilon)\text{id} + \epsilon\mathcal{V}$ with \mathcal{V} being a Haar random unitary acting on single-qubit systems (left) and two-qubit systems (right). The black dotted line corresponds to $\gamma_{\text{disc}}/\gamma_{\text{cont}} = 1$. The number of samples used for each figure is 10^3 .

for any $\mathcal{U} \in \mathcal{T}_u(d)$,

$$2 \sum_{i=0}^{\infty} \sum_{j=0}^i \frac{t_{ij}(-\epsilon_+)^j \epsilon_-^{i-j}}{(1 - \epsilon)^{i+1}} - 1 \leq \gamma_{\text{opt}}(\mathcal{U}) \leq \frac{1}{1 - 2\epsilon_+}, \quad (8)$$

where $t_{ij} := \sum_{\vec{k} \in I_{ij}} \text{Tr}[\Phi_d \text{id} \otimes (\Lambda^j, \Xi^{i-j})_{\vec{k}}(\Phi_d)]$.

The upper bound, an achievable cost (see also Appendix F), becomes especially insightful when the given error channel is not expressed by Pauli or Clifford operations. For instance, consider the generalized dephasing channel $\mathcal{F}_{\hat{n},\epsilon}(\rho) := (1 - \epsilon)\rho + \epsilon e^{i\hat{n} \cdot \hat{\sigma}\pi/2} \rho e^{-i\hat{n} \cdot \hat{\sigma}\pi/2}$, where $\hat{n} \cdot \hat{\sigma} = n_x X + n_y Y + n_z Z$ and \hat{n} is the unit vector that determines the rotation axis. As we show in Appendix G, the best systematic decomposition with the Clifford-based basis in Ref. [8] gives $\gamma = \{1 + (\sqrt{2} - 1)\epsilon\}/(1 - 2\epsilon)$ for $\hat{n} = [\cos(\pi/8), 0, \sin(\pi/8)]$, leading to a larger cost than what is achievable in our framework. This advantage comes from our flexible choice of decomposition basis taking advantage of our large set of implementable operations. In fact, we find that the advantage is generic. We compare two costs γ_{disc} and γ_{cont} , where γ_{disc} is the optimal cost for the discrete basis in Ref. [8] and γ_{cont} is the achievable cost $(1 - 2\epsilon_+)^{-1}$ from Theorem 1 realized by our continuous basis, where we set $\mathcal{U} = \text{id}$ for both cases. Figure 1 plots $\gamma_{\text{disc}}/\gamma_{\text{cont}}$ for randomly sampled single-qubit and two-qubit noise models of the form $\mathcal{E} = (1 - \epsilon)\text{id} + \epsilon\mathcal{V}$, where \mathcal{V} is a Haar random unitary. We find that γ_{cont} is smaller than γ_{disc} in many cases for single-qubit error channels and all the sampled cases for two-qubit error channels. The advantage for two-qubit error is particularly significant—the improvement can become around the factor of 2 for the realistic noise strength, which can drastically reduce the total sampling cost that grows exponentially with the number of gates. Since two-qubit noise will be the most demanding one in real experiments, our result may greatly ease the experimental challenges. At the same time, the advantage seen in Fig. 1 confirms the necessity of considering the extended class of implementable operations introduced here to properly assess the ultimate capability of error mitigation.

The lower bound, on the other hand, corresponds to the fundamental limitations on the error mitigation performance. It serves as a universal necessary cost even if we employ the continuous set of implementable operations. Notably,

this bound readily applies to any subset of implementable operations contained in $\mathcal{I}_{\mathcal{E}}$, encompassing a broad class of noisy devices of interest. Although it might look daunting to evaluate the expression for the lower bound, it can be analytically calculated for many cases, as we discuss in Appendix E.

Theorem 1 can also give insights into correlated noise models. If the noise of interest has correlation among a number of qubits, the effective noise channel gets large and a quasiprobability decomposition becomes numerically intractable even if the description of the noise is given and a discrete set of implementable operations is used. On the other hand, Theorem 1 immediately gives an effective evaluation of the optimal cost as long as the noise channel is provided in a certain form.

Interestingly, the bounds in (8) can be improved by focusing on specific noise models. We can even obtain the *exact* optimal costs for depolarizing and dephasing noise, providing a precise characterization of the devices' capability under these noise models (the proof is given in Appendix H). Together with (5), it also gives an operational meaning to the robustness measure in terms of quantum error mitigation.

Theorem 2. Let $\mathcal{D}_{d,\epsilon}$ be a depolarizing channel acting on d -dimensional systems defined by $\mathcal{D}_{d,\epsilon}(\rho) := (1 - \epsilon)\rho + \epsilon\mathbb{I}/d$. Then, for any unitary gate $\mathcal{U} \in \mathcal{T}_u(d)$ and $0 \leq \epsilon < 1$,

$$\gamma_{\text{opt}}(\mathcal{U}) = \frac{1 + (1 - 2/d^2)\epsilon}{1 - \epsilon}, \quad (9)$$

and the minimum in (3) is achieved at $\eta_0 = 1 + (d^2 - 1)\epsilon/(d^2(1 - \epsilon))$, $\mathcal{O}_0 = \mathcal{D}_{d,\epsilon} \circ \mathcal{U}$ and $\eta_i = -\epsilon/(d^2(1 - \epsilon))$, $\mathcal{O}_i = \mathcal{D}_{d,\epsilon} \circ \mathcal{P}_i \circ \mathcal{U}$, for $i = 1, \dots, d^2 - 1$ where \mathcal{P}_i is the i th Pauli channel.

Also, let $\mathcal{F}_{\epsilon} := (1 - \epsilon)\text{id} + \epsilon\mathcal{Z}$ be the qubit dephasing channel. Then, for any unitary gate $\mathcal{U} \in \mathcal{T}_u(2)$ and $0 \leq \epsilon < 1/2$,

$$\gamma_{\text{opt}}(\mathcal{U}) = \frac{1}{1 - 2\epsilon}, \quad (10)$$

and the minimum in (3) is achieved at $\eta_0 = (1 - \epsilon)/(1 - 2\epsilon)$, $\mathcal{O}_0 = \mathcal{F}_{\epsilon} \circ \mathcal{U}$ and $\eta_1 = -\epsilon/(1 - 2\epsilon)$, $\mathcal{O}_1 = \mathcal{F}_{\epsilon} \circ \mathcal{Z} \circ \mathcal{U}$.

This result shows that for these noise models, the optimal cost can be achieved by the Pauli-based basis operations [4,8], and the continuous degrees of freedom do not help reduce the cost. This particularly implies that the heuristic linear decomposition for the depolarizing noise considered in Ref. [4] remains optimal even in our extended framework, putting fundamental restrictions on the error mitigation feasibility.

Our method also provides a tighter bound for the amplitude damping noise (the proof is given in Appendix I).

Theorem 3. Let \mathcal{A}_{ϵ} be the qubit amplitude damping channel with Kraus operators $A_0 = |0\rangle\langle 0| + \sqrt{1 - \epsilon}|1\rangle\langle 1|$, $A_1 = \sqrt{\epsilon}|0\rangle\langle 1|$. Then, for any unitary channel $\mathcal{U} \in \mathcal{T}_u(2)$ and $0 \leq \epsilon < 1$,

$$\frac{\sqrt{1 - \epsilon} + \epsilon/2}{1 - \epsilon} \leq \gamma_{\text{opt}}(\mathcal{U}) \leq \frac{1 + \epsilon}{1 - \epsilon}. \quad (11)$$

The upper bound in (11) can be achieved by a decomposition $\eta_0 = (1 + \sqrt{1 - \epsilon})/(2(1 - \epsilon))$, $\mathcal{O}_0 = \mathcal{A}_{\epsilon} \circ \mathcal{U}$, $\eta_1 = (1 - \sqrt{1 - \epsilon})/(2(1 - \epsilon))$, $\mathcal{O}_1 = \mathcal{A}_{\epsilon} \circ \mathcal{Z} \circ \mathcal{U}$, and $\eta_2 = -\epsilon/(1 - \epsilon)$, $\mathcal{O}_2 = \mathcal{A}_{\epsilon} \circ \mathcal{P}_{(0)}$, and the lower bound is obtained by finding a good witness operator Y in (6) or, alternatively, directly evaluating the lower bound in (8).

VI. CONCLUSIONS

We investigated the ultimate potential and limitations of near-term noisy devices in terms of quantum error mitigation. We pointed out the wide programmability equipped with NISQ devices and formalized their implementable operations to properly assess their full capability. We provided a general methodology to evaluate the optimal error mitigation cost with the probabilistic error cancellation method by establishing a connection to the robustness measure that naturally emerges as a resource quantifier in our framework. We applied our method to a general noise model and obtained universal bounds for the optimal mitigation cost, finding that our framework, which takes into account the flexible choice of implementable operations, leads to a generic advantage over the strategy based on discrete basis sets. We also obtained the exact optimal costs for depolarizing and dephasing noise, rigorously showing that the cost for depolarizing channels given in Ref. [4] is optimal even in our extended framework, as well as obtained an improved bound for the amplitude damping channel.

Our consideration can be combined with other frameworks such as learning algorithms for unknown error models [13–15] and analog error mitigation [10], and may be extended to a broad class of problems for which the quasiprobability sampling is used, including the measurement of observables in variational algorithms and classical simulation of noisy quantum circuits. The generality of our approach also gives us the freedom to consider various sets of implementable operations, allowing one to tailor the analysis in this work to given physical devices. Another important future work is to extend our results toward a unified information-theoretic account of error mitigation and error correction, which will further clarify the true capability of noisy quantum devices.

ACKNOWLEDGMENTS

The author is grateful to Tennin Yan, Nobuyuki Yoshioka, and Xiao Yuan for fruitful discussions, and in particular to Suguru Endo, Kosuke Mitarai, and Yuya O. Nakagawa for valuable discussions and useful comments on the manuscript. This work was supported by National Research Foundation (NRF) Singapore, under its NRFF Fellow programme (Award No. NRF-NRFF2016-02), the Singapore Ministry of Education Tier 1 Grant No. 2019-T1-002-015, NSF, ARO, IARPA, AFOSR, the Takenaka Scholarship Foundation, and the Cross-ministerial Strategic Innovation Promotion Program (SIP), “Photonics and Quantum Technology for Society5.0.” Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore.

APPENDIX A: OPTIMAL COST AND ROBUSTNESS [EQ. (5)]

Here, we show Eq. (5) for completeness. Let $\Psi, \Xi \in \mathcal{I}_{\mathcal{E}}$ be the operations that achieve $[\mathcal{U} + R_{\mathcal{I}_{\mathcal{E}}}(\mathcal{U})\Xi]/[1 + R_{\mathcal{I}_{\mathcal{E}}}(\mathcal{U})] = \Psi$. Then, one can rewrite $\mathcal{U} = [1 + R_{\mathcal{I}_{\mathcal{E}}}(\mathcal{U})]\Psi - R_{\mathcal{I}_{\mathcal{E}}}(\mathcal{U})\Xi$, which is a valid decomposition of \mathcal{U} . Since γ_{opt} is obtained

by minimizing over all the possible decompositions, we get $\gamma_{\text{opt}}(\mathcal{U}) \leq 1 + R_{\mathcal{I}_{\mathcal{E}}}(\mathcal{U}) + R_{\mathcal{I}_{\mathcal{E}}}(\mathcal{U}) = 2R_{\mathcal{I}_{\mathcal{E}}}(\mathcal{U}) + 1$.

On the other hand, suppose that a decomposition $\mathcal{U} = \sum_{\alpha} \eta_{\alpha} \mathcal{O}_{\alpha}$, $\mathcal{O}_{\alpha} \in \mathcal{I}_{\mathcal{E}}$ achieves $\gamma_{\text{opt}} = \sum_{\alpha} |\eta_{\alpha}|$. Then, by separately writing the terms with non-negative coefficients and negative coefficients, one can write

$$\mathcal{U} = \sum_{\eta_{\alpha} \geq 0} \eta_{\alpha} \mathcal{O}_{\alpha} - \sum_{\eta_{\alpha} < 0} |\eta_{\alpha}| \mathcal{O}_{\alpha} = (1 + s)\Psi - s\Xi, \quad (\text{A1})$$

where $s = \sum_{\eta_{\alpha} < 0} |\eta_{\alpha}|$, and $\Psi = (1 + s)^{-1} \sum_{\eta_{\alpha} \geq 0} \eta_{\alpha} \mathcal{O}_{\alpha}$, $\Xi = s^{-1} \sum_{\eta_{\alpha} < 0} |\eta_{\alpha}| \mathcal{O}_{\alpha}$ are CPTP maps. Note that we also used that \mathcal{U} and \mathcal{O}_{α} are trace preserving and thus $\sum_{\alpha} \eta_{\alpha} = 1$; also, the fact that $\Psi, \Xi \in \mathcal{I}_{\mathcal{E}}$ is ensured by the convexity of $\mathcal{I}_{\mathcal{E}}$. Thus, we get $(\mathcal{U} + s\Xi)/(1 + s) \in \mathcal{I}_{\mathcal{E}}$, $\Xi \in \mathcal{I}_{\mathcal{E}}$, which leads to $R_{\mathcal{I}_{\mathcal{E}}}(\mathcal{U}) \leq s = \sum_{\eta_{\alpha} < 0} |\eta_{\alpha}| = (\gamma_{\text{opt}} - 1)/2$. This shows the other direction of the inequality, $\gamma_{\text{opt}}(\mathcal{U}) \geq 2R_{\mathcal{I}_{\mathcal{E}}}(\mathcal{U}) + 1$.

APPENDIX B: DUAL FORM OF THE ROBUSTNESS [EQ. (6)]

Here, we obtain the dual form of the robustness in (6) for general resource theories of channels (i.e., *any* finite-dimensional convex and closed set of free channels), in which our case is included as a special case. The discussion in this section is based on the argument given for another type of robustness measure (known as generalized robustness) for general resource theories of channels [33].

Let $\mathcal{T}(A, B)$ be the set of quantum channels with input subsystem A and output subsystem B . Given a convex and closed set of channels $\mathcal{O}_{\mathcal{F}} \subseteq \mathcal{T}(A, B)$, we consider the (standard) robustness measure for channel $\Lambda \in \mathcal{T}(A, B)$ with respect to $\mathcal{O}_{\mathcal{F}}$ as

$$R_{\mathcal{O}_{\mathcal{F}}}(\Lambda) := \min \left\{ r \geq 0 \mid \frac{\Lambda + r\Theta}{1+r} \in \mathcal{O}_{\mathcal{F}}, \Theta \in \mathcal{O}_{\mathcal{F}} \right\}, \quad (\text{B1})$$

where $J_{\Lambda} = \text{id} \otimes \Lambda(d \cdot \Phi_d)$ denotes the Choi matrix for a channel Λ and we assume $R_{\mathcal{O}_{\mathcal{F}}}(\Lambda) < \infty$. Let $\mathbb{H}_{A(B)}$ be the subspace consisting of Hermitian operators defined on subsystems $A(B)$, and $\mathcal{O}_{\mathcal{F}}^J \subseteq \mathbb{H}_A \otimes \mathbb{H}_B$ be the set of Choi matrices corresponding to channels in $\mathcal{O}_{\mathcal{F}}$. Then, introducing the variable $\tilde{\Xi} = \Lambda + r\Theta$, (B1) can be rewritten as the following optimization problem:

$$\text{minimize } r, \quad (\text{B2})$$

$$\text{subject to } J_{\tilde{\Xi}} - J_{\Lambda} \in \text{cone}(\mathcal{O}_{\mathcal{F}}^J), \quad (\text{B3})$$

$$J_{\tilde{\Xi}} \in \text{cone}(\mathcal{O}_{\mathcal{F}}^J), \quad (\text{B4})$$

$$\text{Tr}_B[J_{\tilde{\Xi}}] = (1 + r)\mathbb{I}_A, \quad (\text{B5})$$

where $\text{cone}(\mathcal{S}) := \{\lambda S \mid \lambda \geq 0, S \in \mathcal{S}\}$. We write the Lagrangian as

$$\begin{aligned} L(J_{\tilde{\Xi}}, r; X, Y, Z) &= r - \text{Tr}[ZJ_{\tilde{\Xi}}] - \text{Tr}[Y(J_{\tilde{\Xi}} - J_{\Lambda})] \\ &\quad - \text{Tr}[X((1+r)\mathbb{I}_A - \text{Tr}_B J_{\tilde{\Xi}})] \\ &= r(1 - \text{Tr}[X]) - \text{Tr}[(Z + Y - X \otimes \mathbb{I}_B)J_{\tilde{\Xi}}] \\ &\quad + \text{Tr}[YJ_{\Lambda}] - \text{Tr}[X], \end{aligned} \quad (\text{B6})$$

where $Z, Y \in \mathcal{O}_{\mathcal{F}}^J$ and $X \in \mathbb{H}_A$ with $\mathcal{S}^* = \{W \mid \text{Tr}[WC] \geq 0, \forall C \in \text{cone}(\mathcal{S})\}$ being a dual cone of $\text{cone}(\mathcal{S})$. This gives the dual problem

$$\text{maximize } \text{Tr}[YJ_{\Lambda}] - 1, \quad (\text{B7})$$

$$\text{subject to } Y = -Z + X \otimes \mathbb{I} \in \mathcal{O}_{\mathcal{F}}^J, \quad (\text{B8})$$

$$X \in \mathbb{H}_A, \text{Tr}[X] = 1, \quad (\text{B9})$$

$$Z \in \mathcal{O}_{\mathcal{F}}^J. \quad (\text{B10})$$

It can be confirmed that the Slater's condition [45] holds by taking $X = \mathbb{I}_A/d_A$ and $Z = \mathbb{I}_{AB}/(2d_A)$, where d_A is the dimension of A . Thus, the strong duality holds and the dual optimum coincides with the primal optimum. Using (B8)–(B10), we get, for any $\Xi \in \mathcal{O}_{\mathcal{F}}$,

$$\text{Tr}[(-Y + X \otimes \mathbb{I})J_{\Xi}] = -\text{Tr}[YJ_{\Xi}] + \text{Tr}[\Xi(X)] \quad (\text{B11})$$

$$= -\text{Tr}[YJ_{\Xi}] + 1 \geq 0. \quad (\text{B12})$$

Since the objective function does not include X or Z , we reach the following equivalent formulation:

$$\text{maximize } \text{Tr}[YJ_{\Lambda}] - 1, \quad (\text{B13})$$

$$\text{subject to } Y \in \mathcal{O}_{\mathcal{F}}^J \iff \text{Tr}[YJ_{\Xi}] \geq 0, \forall \Xi \in \mathcal{O}_{\mathcal{F}}, \quad (\text{B14})$$

$$\text{Tr}[YJ_{\Xi}] \leq 1, \forall \Xi \in \mathcal{O}_{\mathcal{F}}, \quad (\text{B15})$$

which results in (6) by taking $\mathcal{O}_{\mathcal{F}} = \mathcal{I}_{\mathcal{E}}$.

APPENDIX C: SYSTEMATIC LOWER BOUND FOR EQ. (7)

Here, we present a systematic lower bound that can be constructed by a specific decomposition used for the upper bound. Suppose $\mathcal{U} = \sum_{\alpha} \eta_{\alpha} \mathcal{E} \circ \mathcal{V}_{\alpha}$ for $\mathcal{V}_{\alpha} \in \tilde{\mathcal{P}}$ and define $\mathcal{E}^{-1} := \sum_{\alpha} \eta_{\alpha} \mathcal{V}_{\alpha} \circ \mathcal{U}^{\dagger}$ so that $\mathcal{E} \circ \mathcal{E}^{-1} = \text{id}$. Then, as long as $\mathcal{V}_{\alpha}, \mathcal{E} \in \mathcal{T}(d)$, it also holds that $\mathcal{E}^{-1} \circ \mathcal{E} = \text{id}$. This can be shown by considering a matrix representation of an operation $\Lambda \in \mathcal{T}(d)$ with a $d^2 \times d^2$ matrix $M_{ab,ij}^{\Lambda}$ so that a matrix element of $\sigma := \Lambda(\rho)$ is written as $\sigma_{ab} = \sum_{ij} M_{ab,ij}^{\Lambda} \rho_{ij}$. Then, $\mathcal{E} \circ \mathcal{E}^{-1} = \text{id}$ implies $M^{\mathcal{E}} M^{\mathcal{E}^{-1}} = M^{\text{id}} = \mathbb{I}$. Since the right inverse of a square matrix is also the left inverse, we get $M^{\mathcal{E}^{-1}} M^{\mathcal{E}} = \mathbb{I}$, resulting in $\mathcal{E}^{-1} \circ \mathcal{E} = \text{id}$. Using this, we get

$$\begin{aligned} \frac{1}{d^2} \text{Tr}[J_{\mathcal{E}^{-1} \circ \mathcal{U}} J_{\mathcal{E} \circ \mathcal{V}}] &= \text{Tr}[\text{id} \otimes \mathcal{U}(\Phi_d) \text{id} \otimes \mathcal{E}^{-1} \circ \mathcal{E} \circ \mathcal{V}(\Phi_d)] \\ &= \text{Tr}[\text{id} \otimes \mathcal{U}(\Phi_d) \text{id} \otimes \mathcal{V}(\Phi_d)], \end{aligned} \quad (\text{C1})$$

which ensures $0 \leq d^{-2} \text{Tr}[J_{\mathcal{E}^{-1} \circ \mathcal{U}} J_{\mathcal{E} \circ \mathcal{V}}] \leq 1$ for any $\mathcal{V} \in \tilde{\mathcal{P}}(d)$. Since $J_{\mathcal{E}^{-1} \circ \mathcal{U}}$ is also Hermitian, $Y := d^{-2} J_{\mathcal{E}^{-1} \circ \mathcal{U}}$ is a valid operator satisfying the condition in (6). We can further obtain

$$\begin{aligned} \text{Tr}[YJ_{\mathcal{U}}] &= \text{Tr}[\text{id} \otimes \mathcal{U}(\Phi_d) \text{id} \otimes \mathcal{E}^{-1} \circ \mathcal{U}(\Phi_d)] \\ &= \text{Tr}[\mathcal{U}^T \otimes \text{id}(\Phi_d) \mathcal{U}^T \otimes \mathcal{E}^{-1}(\Phi_d)] \\ &= \text{Tr}[\Phi_d \text{id} \otimes \mathcal{E}^{-1}(\Phi_d)], \end{aligned} \quad (\text{C2})$$

TABLE I. Universal basis presented in Ref. [8] based on Clifford unitaries and projections. Curly letters represent that they are considered as channels, e.g., $\mathcal{H}(\cdot) = H \cdot H$. Note that H, S are the Hadamard gate and the phase gate, and $K = SH$ is the Clifford gate that cycles Pauli operators as $K^{\dagger}XK = Y, K^{\dagger}YK = Z, K^{\dagger}ZK = X$. $\pi_z(\cdot) = (\frac{\mathbb{I}+Z}{2}) \cdot (\frac{\mathbb{I}+Z}{2})$ is a trace nonincreasing projection onto the $|0\rangle$ state.

| | |
|--------------------|---|
| \mathcal{B}_1 | id |
| \mathcal{B}_2 | \mathcal{X} |
| \mathcal{B}_3 | \mathcal{Y} |
| \mathcal{B}_4 | \mathcal{Z} |
| \mathcal{B}_5 | $\mathcal{K}^{\dagger} \circ \mathcal{S}^{\dagger} \circ \mathcal{K}$ |
| \mathcal{B}_6 | $\mathcal{K} \circ \mathcal{S}^{\dagger} \circ \mathcal{K}^{\dagger}$ |
| \mathcal{B}_7 | \mathcal{S}^{\dagger} |
| \mathcal{B}_8 | $\mathcal{K} \mathcal{H} \mathcal{K}^{\dagger}$ |
| \mathcal{B}_9 | \mathcal{H} |
| \mathcal{B}_{10} | $\mathcal{K}^{\dagger} \mathcal{H} \mathcal{K}$ |
| \mathcal{B}_{11} | $\mathcal{K}^{\dagger} \pi_z \mathcal{K}$ |
| \mathcal{B}_{12} | $\mathcal{K} \pi_z \mathcal{K}^{\dagger}$ |
| \mathcal{B}_{13} | π_z |
| \mathcal{B}_{14} | $\mathcal{K}^{\dagger} \circ \pi_z \circ \mathcal{X} \circ \mathcal{K}$ |
| \mathcal{B}_{15} | $\mathcal{K} \circ \pi_z \circ \mathcal{X} \circ \mathcal{K}^{\dagger}$ |
| \mathcal{B}_{16} | $\pi_z \circ \mathcal{X}$ |

where $\mathcal{U}^T(\cdot) = U^T \cdot (U^T)^{\dagger}$ is the unitary transposed with respect to the Schmidt basis of Φ_d , and we also used that $\text{id} \otimes A |\Phi_d\rangle = A^T \otimes \text{id} |\Phi_d\rangle$ for any matrix A . Thus, (7) implies that $2\text{Tr}[\Phi_d \text{id} \otimes \mathcal{E}^{-1}(\Phi_d)] - 1$ serves as a valid lower bound for $\gamma_{\text{opt}}(\mathcal{U})$.

APPENDIX D: BASIS OPERATIONS FOR CPTP MAPS

Although any noise channel can be represented by a CPTP map that possibly involves some external systems, one could consider CP trace nonincreasing maps as effective error channels acting on systems of interest. A trace nonincreasing map on d -dimensional quantum systems is an element of the d^4 -dimensional vector space of linear maps, and thus it can be represented as a linear combination of d^4 linearly independent trace nonincreasing maps. A specific set of such maps $\{\mathcal{B}_i\}_{i=1}^{16}$ (Table I) that serves as a universal basis that can decompose any trace nonincreasing map was introduced in Ref. [8]. It is a set of 16 linearly independent maps, consisting of 10 Clifford unitaries and 6 trace nonincreasing projections. Then, any operation acting on qubit systems can be linearly decomposed with respect to this basis, and its tensor product also serves as a universal basis for multiqubit systems. They also showed that this basis remains linearly independent after suffering from a noise channel, as long as the noise strength is sufficiently small, and used this for the quasiprobability decomposition for probabilistic error cancellation.

Although a universal basis should contain trace nonincreasing maps if the noise model can also be trace nonincreasing in general, in many cases noise models of interest are described as trace-preserving maps. Then, intuitively, one should be able to find a universal basis only consisting of trace-preserving maps for decomposing an arbitrary CPTP map, which is more favorable for probabilistic error cancellation because using trace nonincreasing maps usually incurs a larger overhead

constant due to the larger coefficients in the decomposition needed to complement the non-unit trace (see also the example at the end of this section). In addition, using many projective measurements is not preferred due to the relatively large measurement error rate.

Here, we argue that if we restrict our attention to trace-preserving maps, the number of elements necessary for a universal basis is reduced to $d^4 - d^2 + 1$, which themselves are trace preserving. We then provide a specific set of CPTP maps consisting of unitary and state preparations that serve as a universal basis for CPTP maps on single-qubit and two-qubit systems. To this end, let $\mathcal{O}_{\text{CPTP}}^J(d)$ be the set of Choi matrices for CPTP maps with input and output being d -dimensional quantum systems, and

$$\mathcal{V}(d) := \text{span}[\mathcal{O}_{\text{CPTP}}^J(d)] = \left\{ \sum_i c_i J_{\mathcal{E}_i} \mid c_i \in \mathbb{R}, \mathcal{E}_i \in \mathcal{T}(d) \right\} \quad (\text{D1})$$

be the span of the set of Choi matrices. We can show that $\mathcal{V}(d)$ is equivalent to the subspace

$$\tilde{\mathcal{V}}(d) := \{X \in \mathbb{H}(A, B) \mid \text{Tr}_B X \propto \mathbb{I}_A\}, \quad (\text{D2})$$

where $\mathbb{H}(A, B)$ is the set of Hermitian operators acting on the composite system AB , each of which has local dimension d . This can be seen as follows. Since $\text{Tr}_B J_{\mathcal{E}} = \mathbb{I}_A, \forall \mathcal{E} \in \mathcal{T}(d)$, it is clear that $\mathcal{V}(d) \subseteq \tilde{\mathcal{V}}(d)$. On the other hand, take any $X \in \tilde{\mathcal{V}}(d)$. Then, X can always be decomposed as $X = X_+ - X_-$, where $X_{\pm} \geq 0$, and $\text{Tr}_B X_+ - \text{Tr}_B X_- = c \mathbb{I}_A$ for some constant $c \in \mathbb{R}$. Define

$$\begin{aligned} \tilde{X}_+(\lambda) &:= X_+ + (\lambda \mathbb{I}_A - \text{Tr}_B X_+) \otimes \mathbb{I}_B/d, \\ \tilde{X}_-(\lambda) &:= X_- + [(\lambda - c) \mathbb{I}_A - \text{Tr}_B X_-] \otimes \mathbb{I}_B/d \\ &= X_- + (\lambda \mathbb{I}_A - \text{Tr}_B X_+) \otimes \mathbb{I}_B/d. \end{aligned} \quad (\text{D3})$$

One can check that $X = X_+ - X_- = \tilde{X}_+(\lambda) - \tilde{X}_-(\lambda)$ and $\text{Tr}_B \tilde{X}_+(\lambda) = \lambda \mathbb{I}_A, \text{Tr}_B \tilde{X}_-(\lambda) = (\lambda - c) \mathbb{I}_A$ for any λ . Moreover, by taking large enough λ , we can always ensure that $\tilde{X}_{\pm}(\lambda) \geq 0$. This means that there always exist two CPTP maps $\mathcal{E}_1, \mathcal{E}_2$ and real numbers c_1, c_2 such that $X = c_1 J_{\mathcal{E}_1} + c_2 J_{\mathcal{E}_2}$. This shows $\tilde{\mathcal{V}}(d) \subseteq \mathcal{V}(d)$, resulting in $\mathcal{V}(d) = \tilde{\mathcal{V}}(d)$. $\tilde{\mathcal{V}}(d)$ is a subspace of $\mathbb{H}(A, B)$ with $\dim \mathbb{H}(A, B) = d^4$ and, because of the constraints $\tilde{\mathcal{V}}(d)$ has, the dimension of $\tilde{\mathcal{V}}(d)$ is reduced to $d^4 - d^2 + 1$. Since $\mathcal{V}(d) = \tilde{\mathcal{V}}(d)$, we have $\dim \mathcal{V}(d) = \dim \tilde{\mathcal{V}}(d) = d^4 - d^2 + 1$. Thus, $d^4 - d^2 + 1$ linearly independent CPTP maps are necessary and sufficient to linearly decompose an arbitrary CPTP map.

Up to two-qubit systems, such sets of CPTP maps can be explicitly constructed. For single-qubit systems, we replace $\mathcal{B}_{11}, \mathcal{B}_{12}, \mathcal{B}_{13}$ with trace-preserving state preparation channels and remove $\mathcal{B}_{14}, \mathcal{B}_{15}, \mathcal{B}_{16}$, leading to a set of linearly independent CPTP maps $\{\tilde{\mathcal{B}}_i\}_{i=1}^{13}$ (Table II). For multiqubit systems, simply tensoring $\{\tilde{\mathcal{B}}_i\}_{i=1}^{13}$ is not sufficient to construct a universal basis in general because $13^k < 16^k - 4^k + 1$ for $k \geq 2$. For two-qubit systems, one needs to find $16^2 - 4^2 + 1 - 13^2 = 72$ additional independent operations to construct

TABLE II. Universal basis for CPTP maps acting on single-qubit systems. $\mathcal{P}_{|\psi\rangle}$ is a channel that prepares a state $|\psi\rangle$. $|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$ and $|+y\rangle = \frac{1}{\sqrt{2}}(|0\rangle + i|1\rangle)$ are $+1$ eigenstates of X and Y .

| | |
|----------------------------|---|
| $\tilde{\mathcal{B}}_1$ | $\text{id} = \mathcal{B}_1$ |
| $\tilde{\mathcal{B}}_2$ | $\mathcal{X} = \mathcal{B}_2$ |
| $\tilde{\mathcal{B}}_3$ | $\mathcal{Y} = \mathcal{B}_3$ |
| $\tilde{\mathcal{B}}_4$ | $\mathcal{Z} = \mathcal{B}_4$ |
| $\tilde{\mathcal{B}}_5$ | $\mathcal{K}^{\dagger} \circ \mathcal{S}^{\dagger} \circ \mathcal{K} = \mathcal{B}_5$ |
| $\tilde{\mathcal{B}}_6$ | $\mathcal{K} \circ \mathcal{S}^{\dagger} \circ \mathcal{K}^{\dagger} = \mathcal{B}_6$ |
| $\tilde{\mathcal{B}}_7$ | $\mathcal{S}^{\dagger} = \mathcal{B}_7$ |
| $\tilde{\mathcal{B}}_8$ | $\mathcal{K} \mathcal{H} \mathcal{K}^{\dagger} = \mathcal{B}_8$ |
| $\tilde{\mathcal{B}}_9$ | $\mathcal{H} = \mathcal{B}_9$ |
| $\tilde{\mathcal{B}}_{10}$ | $\mathcal{K}^{\dagger} \mathcal{H} \mathcal{K} = \mathcal{B}_{10}$ |
| $\tilde{\mathcal{B}}_{11}$ | $\mathcal{P}_{ +\rangle}$ |
| $\tilde{\mathcal{B}}_{12}$ | $\mathcal{P}_{ +y\rangle}$ |
| $\tilde{\mathcal{B}}_{13}$ | $\mathcal{P}_{ 0\rangle}$ |

a complete basis, and they can be explicitly found as in Table III. An explicit construction beyond two-qubit systems is left for future work.

Note that since all the operations in Tables II and III are unitary or state preparations, they are elements of programmable operations $\tilde{\mathcal{P}}$ in our framework. Thus, for any CPTP map \mathcal{E} on single-qubit and two-qubit systems, there always exists some decomposition $\mathcal{E} = (1 - \epsilon)\text{id} + \epsilon_+ \Lambda - \epsilon_- \Xi$, $\Lambda, \Xi \in \tilde{\mathcal{P}}$ up to unitary.

Using $\{\tilde{\mathcal{B}}_i\}$ instead of $\{\mathcal{B}_i\}$ can save cost for error mitigation. For instance, one can check that the optimal cost $\gamma = \sum_{\alpha} |\eta_{\alpha}|$ to mitigate the amplitude damping channel with $\{\mathcal{B}_i\}$ results in $(1 + 2\epsilon)/(1 - \epsilon)$, whereas the optimal cost with $\{\tilde{\mathcal{B}}_i\}$ is $(1 + \epsilon)/(1 - \epsilon)$ as in (II).

TABLE III. Universal basis for CPTP maps acting on two-qubit systems. $\mathcal{C}\mathcal{X}$, $\mathcal{C}\mathcal{S}$, $\mathcal{C}\mathcal{H}$, $\mathcal{C}\mathcal{H}\mathcal{X}$ are channel versions of CNOT, controlled-phase, controlled-Hadamard, and NOT controlled with ± 1 eigenstates of the Hadamard gate. \mathcal{S}_W and $i\mathcal{S}_W$ are channel versions of SWAP and iSWAP ($= |00\rangle\langle 00| + i|10\rangle\langle 01| + i|01\rangle\langle 10| + |11\rangle\langle 11|$) gates. The subscripts refer to the subsystems that the operations act on. “ \mathcal{U} + conjugation with \mathcal{V} ” refers to sandwiching \mathcal{U} with \mathcal{V} and its conjugation \mathcal{V}^{\dagger} as $\mathcal{V}^{\dagger} \circ \mathcal{U} \circ \mathcal{V}$. Then, “ \mathcal{U} + conjugation with $\mathcal{K}_{1,2}, \mathcal{K}_{1,2}^{\dagger}$ ” collects all of the nine possible conjugations with $\text{id}_{12}, \mathcal{K}_1, \mathcal{K}_2, \mathcal{K}_1^{\dagger}, \mathcal{K}_2^{\dagger}, \mathcal{K}_1 \otimes \mathcal{K}_2, \mathcal{K}_1 \otimes \mathcal{K}_2^{\dagger}, \mathcal{K}_1^{\dagger} \otimes \mathcal{K}_2, \mathcal{K}_1^{\dagger} \otimes \mathcal{K}_2^{\dagger}$ for \mathcal{U} .

| | |
|---|--|
| $\tilde{\mathcal{B}}_{1-}\tilde{\mathcal{B}}_{169}$ | $\{\mathcal{B}_i\}_{i=1}^{13} \otimes \{\mathcal{B}_i\}_{i=1}^{13}$ |
| $\tilde{\mathcal{B}}_{170}-\tilde{\mathcal{B}}_{178}$ | $\mathcal{C}\mathcal{X} + \text{conjugation with } \mathcal{K}_{1,2}, \mathcal{K}_{1,2}^{\dagger}$ |
| $\tilde{\mathcal{B}}_{179}-\tilde{\mathcal{B}}_{187}$ | $\mathcal{X}_1 \circ \mathcal{C}\mathcal{X} \circ \mathcal{X}_1 + \text{conjugation with } \mathcal{K}_{1,2}, \mathcal{K}_{1,2}^{\dagger}$ |
| $\tilde{\mathcal{B}}_{188}-\tilde{\mathcal{B}}_{196}$ | $\mathcal{C}\mathcal{S} + \text{conjugation with } \mathcal{K}_{1,2}, \mathcal{K}_{1,2}^{\dagger}$ |
| $\tilde{\mathcal{B}}_{197}-\tilde{\mathcal{B}}_{205}$ | $\mathcal{C}\mathcal{H} + \text{conjugation with } \mathcal{K}_{1,2}, \mathcal{K}_{1,2}^{\dagger}$ |
| $\tilde{\mathcal{B}}_{206}-\tilde{\mathcal{B}}_{214}$ | $\mathcal{C}\mathcal{H}\mathcal{X} + \text{conjugation with } \mathcal{K}_{1,2}, \mathcal{K}_{1,2}^{\dagger}$ |
| $\tilde{\mathcal{B}}_{215}-\tilde{\mathcal{B}}_{223}$ | $\mathcal{C}\mathcal{X} \circ \mathcal{H}_1 + \text{conjugation with } \mathcal{K}_{1,2}, \mathcal{K}_{1,2}^{\dagger}$ |
| $\tilde{\mathcal{B}}_{224}-\tilde{\mathcal{B}}_{226}$ | $\mathcal{S}_W + \text{conjugation with } \mathcal{K}_2, \mathcal{K}_2^{\dagger}$ |
| $\tilde{\mathcal{B}}_{227}-\tilde{\mathcal{B}}_{232}$ | $i\mathcal{S}_W + \text{conjugation with } \mathcal{K}_{1,2}, \mathcal{K}_2^{\dagger}$ |
| $\tilde{\mathcal{B}}_{233}-\tilde{\mathcal{B}}_{241}$ | $\mathcal{S}_W \circ \mathcal{H}_1 + \text{conjugation with } \mathcal{K}_{1,2}, \mathcal{K}_{1,2}^{\dagger}$ |

APPENDIX E: PROOF OF THEOREM 1

Proof. We first obtain an upper bound. Let us define

$$\Psi_e = \frac{1}{N_e} \sum_{i=0}^{\infty} \sum_{j:\text{even}} \sum_{\vec{k} \in I_{ij}} \eta_{ij\vec{k}} (\Lambda^j, \Xi^{i-j})_{\vec{k}}, \quad (\text{E1})$$

$$\Psi_o = \frac{1}{N_o} \sum_{i=1}^{\infty} \sum_{j:\text{odd}} \sum_{\vec{k} \in I_{ij}} |\eta_{ij\vec{k}}| (\Lambda^j, \Xi^{i-j})_{\vec{k}}, \quad (\text{E2})$$

where

$$\eta_{ij\vec{k}} = \frac{(-\epsilon_+)^j \epsilon_-^{i-j}}{(1-\epsilon)^{i+1}}, \quad \forall \vec{k} \in I_{ij}, \quad j = 0, \dots, i, \quad (\text{E3})$$

and $N_{e,o}$ are normalization constants,

$$\begin{aligned} N_e &= \sum_{i=0}^{\infty} \sum_{j:\text{even}} \sum_{\vec{k} \in I_{ij}} \eta_{ij\vec{k}} \\ &= \sum_{i=0}^{\infty} \sum_{j:\text{even}} \binom{i}{j} \frac{\epsilon_+^j \epsilon_-^{i-j}}{(1-\epsilon)^{i+1}} \\ &= \sum_{i=0}^{\infty} \frac{1}{2} \frac{(\epsilon_+ + \epsilon_-)^i + (-\epsilon_+ + \epsilon_-)^i}{(1-\epsilon)^{i+1}} \\ &= \frac{1}{2[1-\epsilon - (\epsilon_+ + \epsilon_-)]} + \frac{1}{2[1-\epsilon - (-\epsilon_+ + \epsilon_-)]}, \end{aligned} \quad (\text{E4})$$

$$\begin{aligned} N_o &= \sum_{i=1}^{\infty} \sum_{j:\text{odd}} \sum_{\vec{k} \in I_{ij}} |\eta_{ij\vec{k}}| \\ &= \sum_{i=1}^{\infty} \sum_{j:\text{odd}} \binom{i}{j} \frac{\epsilon_+^j \epsilon_-^{i-j}}{(1-\epsilon)^{i+1}} \\ &= \sum_{i=1}^{\infty} \frac{1}{2} \frac{(\epsilon_+ + \epsilon_-)^i - (-\epsilon_+ + \epsilon_-)^i}{(1-\epsilon)^{i+1}} \\ &= \frac{1}{2[1-\epsilon - (\epsilon_+ + \epsilon_-)]} - \frac{1}{2[1-\epsilon - (-\epsilon_+ + \epsilon_-)]}, \end{aligned} \quad (\text{E5})$$

where we changed the order of the infinite sum because $\sum_{i=0}^{\infty} \left| \frac{(\epsilon_+ + \epsilon_-)^i}{(1-\epsilon)^{i+1}} \right| < \infty$ and $\sum_{i=0}^{\infty} \left| \frac{(-\epsilon_+ + \epsilon_-)^i}{(1-\epsilon)^{i+1}} \right| < \infty$. Noting that $\tilde{\mathcal{P}}$ is closed under concatenation, i.e., $\mathcal{V}_1, \mathcal{V}_2 \in \tilde{\mathcal{P}} \Rightarrow \mathcal{V}_1 \circ \mathcal{V}_2 \in \tilde{\mathcal{P}}$, as well as the convexity of $\tilde{\mathcal{P}}$, we have $\Psi_e, \Psi_o \in \tilde{\mathcal{P}}$. Then,

$$\begin{aligned} \mathcal{E} \circ [N_e \Psi_e - N_o \Psi_o] &= [(1-\epsilon)\text{id} + \epsilon_+ \Lambda - \epsilon_- \Xi] \circ \\ &\quad \lim_{n \rightarrow \infty} \sum_{i=0}^n \sum_{j=0}^i \sum_{\vec{k} \in I_{ij}} \left[\frac{(-\epsilon_+)^j \epsilon_-^{i-j}}{(1-\epsilon)^{i+1}} (\Lambda^j, \Xi^{i-j})_{\vec{k}} \right] \\ &= \text{id} + \lim_{n \rightarrow \infty} \sum_{j=0}^{n+1} \sum_{\vec{k} \in I_{n+1,j}} \frac{(-\epsilon_+)^j \epsilon_-^{n+1-j}}{(1-\epsilon)^{n+1}} (\Lambda^j, \Xi^{n+1-j})_{\vec{k}} \\ &= \text{id}, \end{aligned} \quad (\text{E6})$$

where the last equality is because

$$\begin{aligned} &\left\| \sum_{j=0}^{n+1} \sum_{\vec{k} \in I_{n+1,j}} \frac{(-\epsilon_+)^j \epsilon_-^{n+1-j}}{(1-\epsilon)^{n+1}} (\Lambda^j, \Xi^{i-j})_{\vec{k}} \right\|_{\diamond} \\ &\leq d^2 \frac{\sum_{j=0}^{n+1} \binom{n+1}{j} \epsilon_+^j \epsilon_-^{n+1-j}}{(1-\epsilon)^{n+1}} \\ &= d^2 \left(\frac{\epsilon_+ + \epsilon_-}{1-\epsilon} \right)^{n+1} \rightarrow 0 \quad (n \rightarrow \infty), \end{aligned} \quad (\text{E7})$$

where d is the dimension of the system. Thus, we have $\mathcal{U} = N_e \mathcal{E} \circ \Psi_e \circ \mathcal{U} - N_o \mathcal{E} \circ \Psi_o \circ \mathcal{U}$, and since $\Psi_e \circ \mathcal{U}, \Psi_o \circ \mathcal{U} \in \tilde{\mathcal{P}}$, we get $\gamma_{\text{opt}}(\mathcal{U}) \leq N_e + N_o = \frac{1}{1-\epsilon - (\epsilon_+ + \epsilon_-)} = \frac{1}{1-2\epsilon_+}$.

Next, we obtain the lower bound using the dual form of the robustness (6). Take the following witness:

$$\begin{aligned} Y &= \frac{1}{d^2} J_{\mathcal{E}^{-1\dagger} \circ \mathcal{U}} \\ &= \frac{1}{d} \sum_{i=0}^{\infty} \sum_{j=0}^i \sum_{\vec{k} \in I_{ij}} \frac{(-\epsilon_+)^j \epsilon_-^{i-j}}{(1-\epsilon)^{i+1}} \text{id} \otimes [(\Lambda^{\dagger j}, \Xi^{\dagger i-j})_{\vec{k}} \circ \mathcal{U}](\Phi_d). \end{aligned} \quad (\text{E8})$$

One can check that this is a well-defined bounded operator by

$$\begin{aligned} \|Y\|_{\infty} &\leq \frac{1}{d} \sum_{i=0}^{\infty} \sum_{j=0}^i \sum_{\vec{k} \in I_{ij}} \frac{\epsilon_+^j \epsilon_-^{i-j}}{(1-\epsilon)^{i+1}} \\ &\quad \times \|\text{id} \otimes (\Lambda^{\dagger j}, \Xi^{\dagger i-j})_{\vec{k}} \circ \mathcal{U}(\Phi_d)\|_{\infty} \\ &\leq \frac{1}{d} \sum_{i=0}^{\infty} \sum_{j=0}^i \sum_{\vec{k} \in I_{ij}} \frac{\epsilon_+^j \epsilon_-^{i-j}}{(1-\epsilon)^{i+1}} \max_{i,j,\vec{k}} \|(\Lambda^{\dagger j}, \Xi^{\dagger i-j})_{\vec{k}}\|_{\infty} \\ &= \frac{1}{d} \frac{1}{1-\epsilon - (\epsilon_+ + \epsilon_-)} \max_{i,j,\vec{k}} \|(\Lambda^j, \Xi^{i-j})_{\vec{k}}\|_{\infty} \\ &= \frac{1}{d} \frac{1}{1-\epsilon - (\epsilon_+ + \epsilon_-)}, \end{aligned} \quad (\text{E9})$$

where, in the first inequality, we used the triangle inequality of the operator norm; in the second inequality, we used the property of the completely bounded infinite norm $\|(\text{id} \otimes \cdot)(X)\|_{\infty} \leq \|\cdot\|_{\infty} := \sup_X \|(\text{id} \otimes \cdot)(X)\|_{\infty}$; and in the first equality, we used the dual property between the completely bounded infinite norm and diamond norm [46]. Thus, Y is a valid bounded Hermitian operator.

As argued in Appendix C, this choice of Y satisfies the condition in (6), and defining

$$t_{ij} := \sum_{\vec{k} \in I_{ij}} \text{Tr}[\Phi_d \text{id} \otimes (\Lambda^j, \Xi^{i-j})_{\vec{k}}(\Phi_d)], \quad (\text{E10})$$

we get

$$\text{Tr}[Y J_{\mathcal{U}}] = \text{Tr}[\Phi_d \text{id} \otimes \mathcal{E}^{-1}(\Phi_d)] = \sum_{i=0}^{\infty} \sum_{j=0}^i \frac{t_{ij} (-\epsilon_+)^j \epsilon_-^{i-j}}{(1-\epsilon)^{i+1}}. \quad (\text{E11})$$

Using (7) results in the lower bound of the statement. ■

Although it might look daunting to evaluate t_{ij} , it can actually be exactly obtained for many cases. An important observation is that if at least one state preparation is involved, t_{ij} collapses to a constant. For instance, for the amplitude damping channel \mathcal{A}_δ where $\epsilon = \frac{1+\delta-\sqrt{1-\delta}}{2}$, $\epsilon_+ = \delta$, $\epsilon_- = \frac{\sqrt{1-\delta}-(1-\delta)}{2}$ and $\Lambda = \mathcal{P}_{|0\rangle}$, $\Xi = \mathcal{Z}$, we get

$$t_{ij} = \begin{cases} \frac{1}{4} \binom{i}{j} & (j \neq 0) \\ 1 & (j = 0, i: \text{even}) \\ 0 & (j = 0, i: \text{odd}), \end{cases} \quad (\text{E12})$$

which gives

$$\begin{aligned} & \sum_{i=0}^{\infty} \sum_{j=0}^i \frac{t_{ij}(-\epsilon_+)^j \epsilon_-^{i-j}}{(1-\epsilon)^{i+1}} \\ &= \sum_{i=0}^{\infty} \sum_{j=1}^i \frac{1}{4} \binom{i}{j} \frac{(-\epsilon_+)^j \epsilon_-^{i-j}}{(1-\epsilon)^{i+1}} + \sum_{i:\text{even}} \frac{\epsilon_-^i}{(1-\epsilon)^{i+1}} \\ &= \sum_{i=0}^{\infty} \frac{1}{4} \left(\frac{(-\epsilon_+ + \epsilon_-)^i}{(1-\epsilon)^{i+1}} - \frac{\epsilon_-^i}{(1-\epsilon)^{i+1}} \right) + \sum_{i:\text{even}} \frac{\epsilon_-^i}{(1-\epsilon)^{i+1}} \\ &= \frac{1}{4} \left(1 - \frac{1}{1-\epsilon-\epsilon_-} \right) + \frac{1-\epsilon}{(1-\epsilon-\epsilon_-)(1-\epsilon+\epsilon_-)} \\ &= -\frac{1}{4} \frac{\epsilon+\epsilon_-}{1-\epsilon-\epsilon_-} + \frac{1-\epsilon}{(1-\epsilon-\epsilon_-)(1-\epsilon+\epsilon_-)}. \quad (\text{E13}) \end{aligned}$$

Substituting $\epsilon = \frac{1+\delta-\sqrt{1-\delta}}{2}$, $\epsilon_+ = \delta$, and $\epsilon_- = \frac{\sqrt{1-\delta}-(1-\delta)}{2}$, we get $\gamma_{\text{opt}}(\mathcal{U}) \geq \frac{\sqrt{1-\delta}+\delta/2}{1-\delta}$, which reproduces the lower bound in (11).

APPENDIX F: EFFICIENT SAMPLING FOR GENERAL NOISE CHANNELS

As can be seen in the proof of Theorem 1, the upper bound can be achieved by a decomposition composed of operations applying Λ and Ξ in different patterns. Although it involves an infinite series of operations, an efficient sampling from the probability distribution originating from the infinite series is possible, which allows one to run the probabilistic error cancellation at this cost in practice.

To run the probabilistic error cancellation at the cost $1/(1-2\epsilon_+)$ for an error channel $\mathcal{E} = (1-\epsilon)\text{id} + \epsilon_+\Lambda - \epsilon_-\Xi$, one needs to sample each operation to be applied at the right probability. Specifically, the proof of Theorem 1 implies that the operation $(\Lambda^j, \Xi^{i-j})_{\vec{k}}$ should be applied at probability

$$P_{ijk} := \frac{|\eta_{ijk}|}{\sum_{i=0}^{\infty} \sum_{j=0}^i \sum_{\vec{k} \in I_{ij}} |\eta_{ijk}|} = \frac{1-\epsilon - (\epsilon_+ + \epsilon_-)}{(1-\epsilon)^{i+1}} \epsilon_+^j \epsilon_-^{i-j}. \quad (\text{F1})$$

This sampling can be easily done as follows. First, note that the probability of realizing an element with index i is given by

$$P_i := \sum_{j=0}^i \sum_{\vec{k} \in I_{ij}} P_{ijk} = \left(1 - \frac{\epsilon_+ + \epsilon_-}{1-\epsilon} \right) \left(\frac{\epsilon_+ + \epsilon_-}{1-\epsilon} \right)^i, \quad (\text{F2})$$

which is the probability of observing heads for i times in sequence followed by a tail at the $i+1$ th trial when flipping a

biased coin with $p_{\text{head}} = (\epsilon_+ + \epsilon_-)/(1-\epsilon)$. Also, note that

$$\frac{P_{ij}}{P_i} := \frac{\sum_{\vec{k} \in I_{ij}} P_{ijk}}{P_i} = \binom{i}{j} \left(\frac{\epsilon_+}{\epsilon_+ + \epsilon_-} \right)^j \left(\frac{\epsilon_-}{\epsilon_+ + \epsilon_-} \right)^{i-j}, \quad (\text{F3})$$

which is a binomial distribution, and

$$\frac{P_{ijk}}{P_{ij}} = \frac{P_{ijk'}}{P_{ij}}, \quad \forall \vec{k}, \vec{k}' \in I_{ij}. \quad (\text{F4})$$

Thus, the sampling procedure is summarized as follows:

(1) Flip a biased coin with $p_{\text{head}} = (\epsilon_+ + \epsilon_-)/(1-\epsilon)$ until a tail is observed. Suppose the tail was observed at the $i+1$ th trial.

(2) Flip another biased coin with $p_{\text{head}} = \epsilon_+/(\epsilon_+ + \epsilon_-)$ for i times. Suppose heads were observed for j times.

(3) Randomly choose \vec{k} out of $\binom{i}{j}$ possible vectors in I_{ij} and apply $(\Lambda^j, \Xi^{i-j})_{\vec{k}}$.

APPENDIX G: COST FOR THE GENERALIZED DEPHASING CHANNEL WITH THE UNIVERSAL BASIS IN REF. [8]

Consider the generalized dephasing channel $\mathcal{F}_{\hat{n},\epsilon}(\rho) := (1-\epsilon)\rho + \epsilon e^{i\hat{n} \cdot \hat{\sigma}\pi/2} \rho e^{-i\hat{n} \cdot \hat{\sigma}\pi/2}$, where $\hat{n} \cdot \hat{\sigma} = n_x X + n_y Y + n_z Z$ and \hat{n} is the unit vector that determines the rotation axis. As an example, take $\hat{n} = [\cos(\pi/8), 0, \sin(\pi/8)]$, which is the π rotation with respect to the axis halfway between the X rotation and the Hadamard rotation.

Let us consider an optimal decomposition $\text{id} = \sum_i \eta_i \mathcal{F}_{\hat{n},\epsilon} \circ \mathcal{B}_i$, which is equivalent to decomposing the inverse map $\mathcal{F}_{\hat{n},\epsilon}^{-1} = \sum_i \eta_i \mathcal{B}_i$, using the universal basis in Ref. [8] (see Table I in Appendix D). One can check that

$$\begin{aligned} \text{id} = \mathcal{F}_{\hat{n},\epsilon} \circ & \left[\frac{1-\epsilon}{1-2\epsilon} \text{id} + \frac{(\sqrt{2}-1)\epsilon}{2(1-2\epsilon)} \mathcal{Z} - \frac{\epsilon}{2(1-2\epsilon)} \mathcal{X} \right. \\ & \left. - \frac{\sqrt{2}\epsilon}{2(1-2\epsilon)} \mathcal{H} \right], \quad (\text{G1}) \end{aligned}$$

with $\gamma = \{1 + (\sqrt{2}-1)\epsilon\}/(1-2\epsilon)$, is optimal with this basis (as well as the basis in Table II in Appendix D). This is greater than the upper bound in (8) and this gap comes from the fact that the upper bound in (8) is obtained by adaptively choosing an appropriate basis using the known information about the noise channel $\mathcal{F}_{\hat{n},\epsilon}$.

APPENDIX H: PROOF OF THEOREM 2

Proof. We first show the upper bound using (4). To this end, we consider a specific decomposition of the identity map,

$$\text{id} = \left(1 + \frac{(d^2-1)\epsilon}{d^2(1-\epsilon)} \right) \mathcal{D}_{d,\epsilon} \circ \text{id} - \frac{\epsilon}{d^2(1-\epsilon)} \sum_{i=1}^{d^2-1} \mathcal{D}_{d,\epsilon} \circ \mathcal{P}_i. \quad (\text{H1})$$

By applying \mathcal{U} to both sides from the right and noting that $\mathcal{P}_i \circ \mathcal{U} \in \mathcal{T}_u$, we get $\gamma_{\text{opt}}(\mathcal{U}) \leq \{1 + (1-2/d^2)\epsilon\}/(1-\epsilon)$ by (3). On the other hand, consider the following $Y \in \mathbb{H}$ providing

the systematic lower bound:

$$Y = d^{-2} J_{\mathcal{D}_{d,\epsilon}^{-1\dagger} \circ \mathcal{U}} = \frac{1}{d} \frac{1}{1-\epsilon} \text{id} \otimes \mathcal{U}(\Phi_d) - \frac{1}{d} \frac{\epsilon}{d^2(1-\epsilon)} \mathbb{I}. \quad (\text{H2})$$

A straightforward computation together with (7) leads to

$$\begin{aligned} \gamma_{\text{opt}}(\mathcal{U}) &\geq 2\text{Tr}[\Phi_d \text{id} \otimes \mathcal{D}_{d,\epsilon}^{-1}(\Phi_d)] - 1 \\ &= \{1 + (1 - 2/d^2)\epsilon\}/(1 - \epsilon), \end{aligned} \quad (\text{H3})$$

concluding the proof.

The case for the dephasing noise can be shown similarly. We first get an upper bound by providing a specific decomposition. Namely, consider the following decomposition:

$$\text{id} = \frac{1-\epsilon}{1-2\epsilon} \mathcal{F}_\epsilon \circ \text{id} - \frac{\epsilon}{1-2\epsilon} \mathcal{F}_\epsilon \circ \mathcal{Z}, \quad (\text{H4})$$

which can be explicitly checked as

$$\begin{aligned} ((1-\epsilon)\text{id} + \epsilon\mathcal{Z}) \circ &\left(\frac{1-\epsilon}{1-2\epsilon} \text{id} - \frac{\epsilon}{1-2\epsilon} \mathcal{Z} \right) \\ &= \left(\frac{(1-\epsilon)^2}{1-2\epsilon} - \frac{\epsilon^2}{1-2\epsilon} \right) \text{id} + \left(-\frac{\epsilon(1-\epsilon)}{1-2\epsilon} + \frac{\epsilon(1-\epsilon)}{1-2\epsilon} \right) \mathcal{Z} \\ &= \text{id}. \end{aligned} \quad (\text{H5})$$

By applying \mathcal{U} to both sides of (H4) from the right and using (3), we get $\gamma_{\text{opt}}(\mathcal{U}) \leq \frac{1}{1-2\epsilon}$.

Next, we obtain a lower bound using (7) and the dual form of the robustness (6). Consider the following $Y \in \mathbb{H}$:

$$\begin{aligned} Y &= \frac{1}{4} J_{\mathcal{F}_\epsilon^{-1\dagger} \circ \mathcal{U}} \\ &= \frac{1}{2} \frac{1-\epsilon}{1-2\epsilon} \text{id} \otimes \mathcal{U}(\Phi_2) - \frac{1}{2} \frac{\epsilon}{1-2\epsilon} [\text{id} \otimes (\mathcal{Z} \circ \mathcal{U})](\Phi_2). \end{aligned} \quad (\text{H6})$$

As argued in Appendix C, this choice of Y satisfies the condition in (6). A lower bound of the robustness can be computed using this Y as

$$\begin{aligned} \text{Tr}[YJ_{\mathcal{U}}] - 1 &= \text{Tr}[\Phi_2 \text{id} \otimes \mathcal{F}_\epsilon^{-1}(\Phi_2)] - 1 \\ &= \frac{1-\epsilon}{1-2\epsilon} - \frac{\epsilon}{1-2\epsilon} \text{Tr}[\Phi_2 \text{id} \otimes \mathcal{Z}(\Phi_2)] - 1 \\ &= \frac{\epsilon}{1-2\epsilon}. \end{aligned} \quad (\text{H7})$$

Thus, the dual form of the robustness (6) implies $R_{\mathcal{I}_{\mathcal{F}_\epsilon}}(\mathcal{U}) \geq \frac{\epsilon}{1-2\epsilon}$, and using (7), we get $\gamma_{\text{opt}}(\mathcal{U}) \geq \frac{1}{1-2\epsilon}$. Combining it with the matching upper bound shown above concludes the proof of the statement. ■

APPENDIX I: PROOF OF THEOREM 3

Proof. Consider the following decomposition:

$$\begin{aligned} \text{id} &= \frac{1+\sqrt{1-\epsilon}}{2(1-\epsilon)} \mathcal{A}_\epsilon \circ \text{id} + \frac{1-\sqrt{1-\epsilon}}{2(1-\epsilon)} \mathcal{A}_\epsilon \circ \mathcal{Z} \\ &\quad - \frac{\epsilon}{1-\epsilon} \mathcal{A}_\epsilon \circ \mathcal{P}_{|0\rangle}, \end{aligned} \quad (\text{I1})$$

which proves $\gamma_{\text{opt}}(\mathcal{U}) \leq \frac{1+\epsilon}{1-\epsilon}$.

To obtain a lower bound using the dual form (6), consider the following operator:

$$\begin{aligned} Y &= \frac{1}{4} J_{\mathcal{A}_\epsilon^{-1\dagger} \circ \mathcal{U}} \\ &= \frac{1}{2} \frac{1+\sqrt{1-\epsilon}}{2(1-\epsilon)} \text{id} \otimes \mathcal{U}(\Phi_2) \\ &\quad + \frac{1}{2} \frac{1-\sqrt{1-\epsilon}}{2(1-\epsilon)} [\text{id} \otimes (\mathcal{Z} \circ \mathcal{U})](\Phi_2) \\ &\quad - \frac{1}{4} \frac{\epsilon}{1-\epsilon} \mathcal{U}^T(|0\rangle\langle 0|) \otimes \mathbb{I}. \end{aligned} \quad (\text{I2})$$

As argued in Appendix C, this choice of Y brings us to

$$0 \leq \text{Tr}[YJ_{\mathcal{A}_\epsilon \circ \mathcal{V}}] = \text{Tr}[\text{id} \otimes \mathcal{U}(\Phi_2) \text{id} \otimes \mathcal{V}(\Phi_2)] \leq 1, \quad \forall \mathcal{V} \in \tilde{\mathcal{P}}, \quad (\text{I3})$$

ensuring the condition in (6). Finally, using the form of (I2), we get a lower bound of the robustness,

$$\begin{aligned} \text{Tr}[YJ_{\mathcal{U}}] - 1 &= \text{Tr}[\Phi_2 \text{id} \otimes \mathcal{A}_\epsilon^{-1}(\Phi_2)] - 1 \\ &= \frac{\sqrt{1-\epsilon} - 1 + 3\epsilon/2}{2(1-\epsilon)}. \end{aligned} \quad (\text{I4})$$

Using (7), we obtain $\gamma_{\text{opt}}(\mathcal{U}) \geq \frac{\sqrt{1-\epsilon} + \epsilon/2}{1-\epsilon}$, which concludes the proof. ■

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

[2] S. McArdle, S. Endo, A. Aspuru-Guzik, S. C. Benjamin, and X. Yuan, Quantum computational chemistry, *Rev. Mod. Phys.* **92**, 015003 (2020).

[3] F. Arute *et al.*, Quantum supremacy using a programmable superconducting processor, *Nature* **574**, 505 (2019).

[4] K. Temme, S. Bravyi, and J. M. Gambetta, Error Mitigation for Short-Depth Quantum Circuits, *Phys. Rev. Lett.* **119**, 180509 (2017).

[5] Y. Li and S. C. Benjamin, Efficient Variational Quantum Simulator Incorporating Active Error Minimization, *Phys. Rev. X* **7**, 021050 (2017).

[6] J. R. McClean, M. E. Kimchi-Schwartz, J. Carter, and W. A. de Jong, Hybrid quantum-classical hierarchy for mitigation of decoherence and determination of excited states, *Phys. Rev. A* **95**, 042308 (2017).

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

[8] S. Endo, S. C. Benjamin, and Y. Li, Practical Quantum Error Mitigation for Near-Future Applications, *Phys. Rev. X* **8**, 031027 (2018).

[9] S. McArdle, X. Yuan, and S. Benjamin, Error-Mitigated Digital Quantum Simulation, *Phys. Rev. Lett.* **122**, 180501 (2019).

[10] J. Sun, X. Yuan, T. Tsunoda, V. Vedral, S. C. Benjamin, and S. Endo, Mitigating Realistic Noise in Practical Noisy Intermediate-Scale Quantum Devices, *Phys. Rev. Appl.* **15**, 034026 (2021).

[11] C. Song, J. Cui, H. Wang, J. Hao, H. Feng, and Y. Li, Quantum computation with universal error mitigation on a superconducting quantum processor, *Sci. Adv.* **5**, eaaw5686 (2019).

[12] S. Zhang, Y. Lu, K. Zhang, W. Chen, Y. Li, J.-N. Zhang, and K. Kim, Error-mitigated quantum gates exceeding physical fidelities in a trapped-ion system, *Nat. Commun.* **11**, 587 (2020).

[13] P. Czarnik, A. Arrasmith, P. J. Coles, and L. Cincio, Error mitigation with Clifford quantum-circuit data, [arXiv:2005.10189](https://arxiv.org/abs/2005.10189).

[14] A. Strikis, D. Qin, Y. Chen, S. C. Benjamin, and Y. Li, Learning-based quantum error mitigation, [arXiv:2005.07601](https://arxiv.org/abs/2005.07601).

[15] A. Zlokapa and A. Gheorghiu, A deep learning model for noise prediction on near-term quantum devices, [arXiv:2005.10811](https://arxiv.org/abs/2005.10811).

[16] M. Troyer and U.-J. Wiese, Computational Complexity and Fundamental Limitations to Fermionic Quantum Monte Carlo Simulations, *Phys. Rev. Lett.* **94**, 170201 (2005).

[17] H. Pashayan, J. J. Wallman, and S. D. Bartlett, Estimating Outcome Probabilities of Quantum Circuits Using Quasiprobabilities, *Phys. Rev. Lett.* **115**, 070501 (2015).

[18] M. Howard and E. Campbell, Application of a Resource Theory for Magic States to Fault-Tolerant Quantum Computing, *Phys. Rev. Lett.* **118**, 090501 (2017).

[19] J. R. Seddon, B. Regula, H. Pashayan, Y. Ouyang, and E. T. Campbell, Quantifying quantum speedups: Improved classical simulation from tighter magic monotones, *PRX Quantum* **2**, 010345 (2021).

[20] F. Buscemi, M. Dall’Arno, M. Ozawa, and V. Vedral, Direct observation of any two-point quantum correlation function, [arXiv:1312.4240](https://arxiv.org/abs/1312.4240).

[21] F. Buscemi, M. Dall’Arno, M. Ozawa, and V. Vedral, Universal optimal quantum correlator, *Intl. J. Quantum Inf.* **12**, 1560002 (2014).

[22] A. M. Steane, Error Correcting Codes in Quantum Theory, *Phys. Rev. Lett.* **77**, 793 (1996).

[23] P. W. Shor, Scheme for reducing decoherence in quantum computer memory, *Phys. Rev. A* **52**, R2493 (1995).

[24] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, Surface codes: Towards practical large-scale quantum computation, *Phys. Rev. A* **86**, 032324 (2012).

[25] E. Chitambar and G. Gour, Quantum resource theories, *Rev. Mod. Phys.* **91**, 025001 (2019).

[26] G. Vidal and R. F. Werner, Computable measure of entanglement, *Phys. Rev. A* **65**, 032314 (2002).

[27] R. Takagi and Q. Zhuang, Convex resource theory of non-Gaussianity, *Phys. Rev. A* **97**, 062337 (2018).

[28] F. Albarelli, M. G. Genoni, M. G. A. Paris, and A. Ferraro, Resource theory of quantum non-Gaussianity and Wigner negativity, *Phys. Rev. A* **98**, 052350 (2018).

[29] K. C. Tan, S. Choi, and H. Jeong, Negativity of Quasiprobability Distributions as a Measure of Nonclassicality, *Phys. Rev. Lett.* **124**, 110404 (2020).

[30] G. Vidal and R. Tarrach, Robustness of entanglement, *Phys. Rev. A* **59**, 141 (1999).

[31] X. Yuan, Y. Liu, Q. Zhao, B. Regula, J. Thompson, and M. Gu, Universal and operational benchmarking of quantum memories, *npj Quantum Inf.* **7**, 108 (2021).

[32] R. Takagi, B. Regula, K. Bu, Z.-W. Liu, and G. Adesso, Operational Advantage of Quantum Resources in Subchannel Discrimination, *Phys. Rev. Lett.* **122**, 140402 (2019).

[33] R. Takagi and B. Regula, General Resource Theories in Quantum Mechanics and Beyond: Operational Characterization via Discrimination Tasks, *Phys. Rev. X* **9**, 031053 (2019).

[34] N. Y. Halpern and D. T. Limmer, Fundamental limitations on photoisomerization from thermodynamic resource theories, *Phys. Rev. A* **101**, 042116 (2020).

[35] R. Takagi, Skew information from an operational view via resource theory of asymmetry, *Sci. Rep.* **9**, 14562 (2019).

[36] K. Fang and Z.-W. Liu, No-Go Theorems for Quantum Resource Purification, *Phys. Rev. Lett.* **125**, 060405 (2020).

[37] R. Takagi, K. Wang, and M. Hayashi, Application of the Resource Theory of Channels to Communication Scenarios, *Phys. Rev. Lett.* **124**, 120502 (2020).

[38] S. Zhou, Z.-W. Liu, and L. Jiang, New perspectives on covariant quantum error correction, *Quantum* **5**, 521 (2020).

[39] W. Hoeffding, Probability inequalities for sums of bounded random variables, *J. Am. Stat. Assoc.* **58**, 13 (1963).

[40] C. Chamberland, T. Jochym-O’Connor, and R. Laflamme, Thresholds for Universal Concatenated Quantum Codes, *Phys. Rev. Lett.* **117**, 010501 (2016).

[41] R. Takagi, T. J. Yoder, and I. L. Chuang, Error rates and resource overheads of encoded three-qubit gates, *Phys. Rev. A* **96**, 042302 (2017).

[42] Y. Liu and X. Yuan, Operational resource theory of quantum channels, *Phys. Rev. Research* **2**, 012035(R) (2020).

[43] Z.-W. Liu and A. Winter, Resource theories of quantum channels and the universal role of resource erasure, [arXiv:1904.04201](https://arxiv.org/abs/1904.04201).

[44] B. Regula, Convex geometry of quantum resource quantification, *J. Phys. A* **51**, 045303 (2017).

[45] S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, New York, 2004).

[46] J. Watrous, Semidefinite Programs for Completely Bounded Norms, *Theory Comput.* **5**, 217 (2009).