



**HAL**  
open science

# Exponentially tighter bounds on limitations of quantum error mitigation

Yihui Quek, Daniel Stilck França, Sumeet Khatri, Johannes Jakob Meyer,  
Jens Eisert

► **To cite this version:**

Yihui Quek, Daniel Stilck França, Sumeet Khatri, Johannes Jakob Meyer, Jens Eisert. Exponentially tighter bounds on limitations of quantum error mitigation. 2022. hal-03903322

**HAL Id: hal-03903322**

**<https://hal.science/hal-03903322>**

Preprint submitted on 16 Dec 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Exponentially tighter bounds on limitations of quantum error mitigation

Yihui Quek,<sup>1</sup> Daniel Stilck França,<sup>2,3,1</sup> Sumeet Khatri,<sup>1</sup> Johannes Jakob Meyer,<sup>1</sup> and Jens Eisert<sup>1,4</sup>

<sup>1</sup>*Dahlem Center for Complex Quantum Systems,  
Freie Universität Berlin, 14195 Berlin, Germany*

<sup>2</sup>*Department of Mathematical Sciences, University of Copenhagen, 2100 København, Denmark*

<sup>3</sup>*Univ Lyon, Inria, ENS Lyon, UCBL, LIP, F-69342, Lyon Cedex 07, France.*

<sup>4</sup>*Helmholtz-Zentrum Berlin für Materialien und Energie, 14109 Berlin, Germany*

(Dated: October 24, 2022)

Quantum error mitigation has been proposed as a means to combat unwanted and unavoidable errors in near-term quantum computing by classically post-processing outcomes of multiple quantum circuits. It does so in a fashion that requires no or few additional quantum resources, in contrast to fault-tolerant schemes that come along with heavy overheads. Error mitigation leads to noise reduction in small schemes of quantum computation. In this work, however, we identify strong limitations to the degree to which quantum noise can be effectively ‘undone’ for larger system sizes. We start out by presenting a formal framework that rigorously encapsulates large classes of meaningful and practically applied schemes for quantum error mitigation, including virtual distillation, Clifford data regression, zero-noise extrapolation and probabilistic error cancellation. With the framework in place, our technical contribution is to construct families of random circuits that are highly sensitive to noise, in the sense that even at  $\log \log(n)$  depth, a whisker beyond constant, quantum noise is seen to super-exponentially rapidly scramble their output into the maximally-mixed state.

Our results exponentially tighten arguments that have been used in the literature for error mitigation, but they go beyond that: with modifications, our arguments can be applied to kernel estimation for quantum machine learning or to compute the depth at which barren plateaus emerge, implying that the scrambling due to noise kicks in at exponentially smaller depths than thought in prior works. Finally, our results also say that a noisy device must be sampled exponentially-many times (in the number of gates in the light-cone of the observable) to estimate an expectation value of an observable. There are classical algorithms that exhibit the same scaling in complexity. While improvements in quantum hardware will push noise levels down, if error mitigation is used, ultimately this can only lead to an exponential time algorithm with a better exponent when compared with classical algorithms, putting up a strong obstruction to the hope for exponential quantum speedups in this setting.

## CONTENTS

I. Motivation and overview	2
A. Our results	3
B. Consequences of our results	7
1. Loss of quantum advantage at $\log \log n$ depth	7
2. Extending previous distinguishability bounds for unital noise to non-unital noise	7
3. A construction of random quantum circuits whose distance to the ground state decays exponentially in both depth and system size	7
4. Consequences for near-term applications	8
5. No exponential quantum speedups to estimate expectation values on noisy devices	8
C. Related work	8
D. Organization of this work	9
II. Error mitigation: The lay of the land	10
A. Preliminaries: Relative entropy and derived quantities	10
B. Relation to practical error mitigation protocols	11
C. Error mitigation setting	13
D. Relationship between notions of error mitigation	15
1. Weak error mitigation implies strong error mitigation only with exponentially-many observables	16
III. Conceptual background and contributions	19
A. Preliminaries	19
1. Pauli operators and their properties	19

2. Depolarizing noise	20
3. Unitary 2-designs and Clifford unitaries	21
B. Appetizer: mitigating depolarizing noise requires exponential-in-circuit depth samples	21
C. The input-state <i>aware</i> case	25
IV. Technical background and contributions	27
A. Technical background	27
B. Bound for multiple layers of depolarizing noise	30
C. Putting it all together	30
D. Strengthening our results	31
1. Geometrically local circuits	31
2. Smaller depths and light-cone size dependence	32
V. Beyond unital noise	33
A. Purity and overlap change after one noisy gate	33
B. Convergence of global random circuits	35
C. Bounds on the probability of successful virtual distillation	37
VI. Outlook	38
VII. Acknowledgments	38
References	39

## I. MOTIVATION AND OVERVIEW

Quantum computers promise to efficiently solve some computational tasks out of reach of classical supercomputers. As early as in the 1980s [1, 2], it was suspected that quantum devices may have computational capabilities that go substantially beyond those of classical Turing machines and hence of classical computers. Shor’s algorithm, presented in the mid 1990s, confirmed this suspicion by presenting an efficient quantum algorithm for factoring for which no efficient classical algorithm is known [3]. Since then, the quantum computer has been a hugely inspiring theoretical idea. It also served as a guiding principle in devising actual quantum devices. Soon, it became clear that unwanted interactions with an environment and hence the concomitant decoherence would be the major threat against realizing quantum computers as actual physical devices.

Early fears that decoherence could not be overcome in principle, fortunately, could be proven wrong. The field of quantum error correction presented ideas that show that even if one cannot read out logical quantum information along the way without necessarily perturbing the very same quantum information, one can still correct for errors, in fact arbitrary unknown errors [4, 5]. This key insight triggered a development that led to the blueprint of what is called the fault-tolerant quantum computer [6, 7], a (so far still fictitious) device that allows for arbitrary errors and can still maintain an arbitrarily long and complex quantum computation. That said, known schemes for fault tolerance come along with demanding, possibly prohibitive overheads. In the most popular of fault tolerant schemes, one would think of realizing surface codes, while computation would be performed by so-called lattice surgery and magic state distillation. This means that logical qubits are encoded in a number of physical qubits that is vastly larger [7, 8]. For the quantum devices that have been experimentally developed in recent years—extremely impressive devices after all the time of quantum computers being primarily objects of theoretical thoughts—such prescriptions still seem by far out of scope.

For this reason, techniques of *quantum error mitigation* have been developed [9–16], methods that would largely undo quantum errors by classical means, with no or little overheads in physical hardware. This is a compelling idea. Refs. [9, 10] on the one hand specifically discuss situations in which a quantum circuit is operated at the minimum possible quantum noise level, repeating the sequence having deliberately increased the physical error rate. Based on the available data, one conceivably can interpolate to the estimate of the zero-error value, presuming that the error sources had scaled proportionately. Ref. [17] on the other hand have proposed to use a learning-based approach where data from the quantum device is amended with known outputs of quantum circuits that are efficiently simulatable.

*But how far can this go?* To what extent can one effectively “undo” quantum noise? Identifying the potential and limitations of this idea seems of pivotal importance to assess the capabilities of near-term quantum computing without quantum error correction. This is no detail: What is at stake is no less than the question whether noise can be effectively canceled in near-term quantum devices. Hence, the elephant in the room is the question to what extent there is scope for quantum computing before the advent of fault tolerant quantum computers.

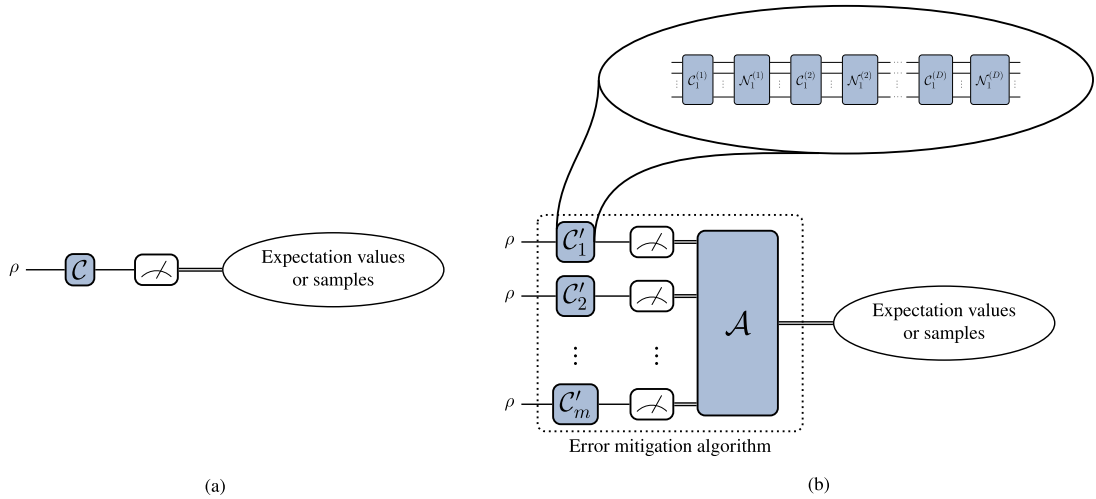


FIG. 1. (a) Idealization of near-term quantum algorithms without quantum noise: Most such algorithms work by running an  $n$ -qubit quantum circuit  $\mathcal{C}$ , measuring it and outputting expectation values of specified observables or samples. (b) The model of error mitigation used in this work.  $\mathcal{C}'_i$  is the  $i$ -th run of  $\mathcal{C}'$ , the noisy version of  $\mathcal{C}$ . We model the noise acting on  $\mathcal{C}$  by interleaving its layers with layers of a given noise channel. In general, our model encompasses those error mitigation protocols that modify  $\mathcal{C}$  on each iteration. In section II B, we will show that this model applies to practical error mitigation protocols such as *virtual distillation* [20, 21], *Clifford data regression* (CDR) [17], *zero-noise extrapolation* (ZNE) [10] and *probabilistic error cancellation* (PEC) [10, 22, 23]. We will study how  $m$ , the number of noisy circuit runs, scales with  $n$  and  $D$ .

In this work, we identify striking and compelling limitations to such prescriptions, by resorting to notions of statistical learning theory. While ideas and methods of quantum error mitigation are surely useful for a class of small noisy quantum devices, we find fundamental limitations of quantum error mitigation by proving lower bounds on its worst-case complexity. Since this work is aimed at providing *rigorous bounds*, we have to be precise what we exactly mean by a quantum error mitigation protocol. We basically discuss two types of schemes. In the first type, the error mitigation algorithm is provided a classical description of a noiseless circuit and has access to several specimens of the noisy circuit's output state, for some fixed input state. On this output, arbitrary measurements can be performed. The goal of the quantum error mitigation algorithm is to output a collection of expectation values of the noiseless circuit's output state. We refer to this family of algorithms as *expectation-value-error mitigation* or *weak error mitigation*. This approach is particularly important in variational quantum circuits [9, 18], an important family of circuits for near-term quantum devices. In other situations, it is desired to output a sample from the computational basis of the anticipated clean output state. We refer to this family of methods as *sample-error mitigation* or *strong error mitigation*. For both families of methods, we prove rigorous limitations. While the language and framework used is that of mathematical physics, to come up with precise statements, it should be clear that it captures practically minded algorithms as they are actually being used in today's laboratories around the world.

Although our results show stark limitations of error mitigation when the depth goes slightly above constant in the worst case, it also invites the question of how to design unitary circuits in such a way that they are more resilient to noise to avoid such behavior. Furthermore, it also motivates the study of whether other models of quantum computation, say dissipative preparation [19], could offer good and more robust alternatives to the circuit model in the NISQ era. Another interesting question that is raised by our work is what is the most effective way to obtain sample error-mitigation protocols, that is, obtaining samples from the noiseless circuit given access to a noisy device.

## A. Our results

To establish the framework, we start off by defining what constitutes 'error mitigation' in the rest of this work. In the literature, the term 'error mitigation' has been used to describe protocols that reduce the unwanted effect of noise on a quantum circuit by measuring it and classically post-processing the results. At its core, an error mitigation algorithm is one that converts noisy quantum states to a classical representation of the noiseless quantum state. The appropriate representation varies: error mitigation is often appended after a quantum algorithm, to revert its outputs to their noiseless version, so the error mitigation should output the desired output of the quantum 'outer loop' into which it is inserted; see Fig. 1 for a schematic depiction.

In slightly more detail, let us consider the following problem, where  $\sigma$  and  $\sigma'$  denote the noiseless and noisy circuit's

output state respectively and  $\mathcal{C}$  and  $\mathcal{C}'$  denote the noiseless circuit and noisy circuit we wish to run, respectively:

**Problem 1** (Error mitigation, informal definition). *Given as input:*

- Classical descriptions of a noiseless circuit  $\mathcal{C}$ , some noise map that transforms  $\mathcal{C}$  to  $\mathcal{C}'$  and a set  $\mathcal{M} = \{O_i\}$  of observables.
- Copies of  $\sigma'$ , and the ability to measure those copies.

*Output either:*

- Estimates of the expectation values  $\text{Tr}(O_i\sigma)$  for each  $O_i \in \mathcal{M}$ . We call this *weak error mitigation*; denote the number of copies of  $\sigma'$  needed as  $m_{EM,weak}$ .
- Samples from a probability distribution which is approximately that associated with measuring  $\sigma$  in the computational basis, i.e.,  $x \sim \text{Tr}(|x\rangle\langle x|\sigma)$ . We call this *strong error mitigation*; denote the number of copies of  $\sigma'$  needed as  $m_{EM,strong}$ .

We refer the reader to Definitions 2 and 3 for rigorous definitions of error mitigation algorithms, and to Section IIB that shows that lower bounds against this model translate into lower bounds on well-known protocols such as zero-noise extrapolation and probabilistic error cancellation—that, for instance, run  $\mathcal{C}$  with varying levels of noise, or apply simple quantum post-processing to  $\mathcal{C}$ , or adapt their classical/quantum post-processing steps based on the output of intermediate measurements.

In this work, we take the perspective that error mitigation can be cast a *statistical inference* problem on quantum states—where ‘statistical inference’ is an umbrella term for problems that deal with identifying underlying parameters of an underlying distribution based on (possibly corrupted) samples. In this view, the underlying distribution in error mitigation comes from the *noiseless* state, and the samples are those from measuring the *noisy* state. Moreover, ‘good’ error mitigation should act as an effective denoiser, allowing one to distinguish one state from another, even if the states can only be accessed by measuring their noisy versions. This motivates us to consider the following problem:

**Problem 2** (Noisy state discrimination (informal)). *Given as input:*

- Classical descriptions of a set of  $n$ -qubit states  $S = \{\rho_0, \rho_1, \dots, \rho_N\}$ , a noiseless circuit  $\mathcal{C}$ , and some noise map that transforms  $\mathcal{C}$  to  $\mathcal{C}'$ .
- $m_{dist}$  copies of an unknown  $\mathcal{C}'(\rho_i)$  for  $i \in \{0, \dots, N\}$  and the ability to measure those copies.

*Output:*  $\hat{i} \in \{0, \dots, N\}$  such that  $\hat{i} = i$  (“success”) with high probability.

A central observation in this work is that noisy state discrimination for a particular choice of  $S$  and  $\mathcal{C}$  reduces to weak error mitigation. Therefore if it is known that *at least* a certain number of copies of the noisy state are needed for successful discrimination for that choice of  $S$  and  $\mathcal{C}$  (solving Problem 2), then *at least* the same number of copies are needed for successful weak error mitigation (solving Problem 1). We use the *generalized Fano method for multiple hypothesis testing* to provide the aforementioned information-theoretic lower bound for noisy state discrimination. The implication for error mitigation is encapsulated in the following central lemma:

**Lemma 1** (Central lemma). *Let  $P_i$  be the distribution over the outputs of an algorithm  $\mathcal{A}$  for noisy state discrimination (Problem 2) when the unknown state is  $\mathcal{C}'(\rho_i)$ . Let us define  $\alpha$  as an upper-bound to the following quantity for  $N \geq 2$ :*

$$\frac{1}{N+1} \sum_{k=0}^N D(P_k \| P_N) \leq \log(N) \cdot \alpha, \quad (1)$$

where  $D(\cdot \| \cdot)$  is the (classical) relative entropy. Then  $\mathcal{A}$  succeeds with probability at most  $1 - \alpha$ . Moreover, there is a set of instances of Problem 2, such that an algorithm for weak error mitigation (Problem 1) suffices to solve these instances. This implies that

$$m_{EM,weak} \geq m_{dist} \quad (2)$$

This lemma implies an immediate lower bound on the number of copies  $m_{dist}$  required to ensure a constant probability of success in Problem 2: Namely, the number required for  $\alpha$  (generically a function of the number  $n$  of qubits, the depth  $D$  of the circuit, and  $m_{dist}$ ) to be a constant.

*Proof Sketch.* The bound on the probability of success originates from Ref. [24] (Corollary 2.6). We refer the reader to Lemma 8 for an extended statement. We now prove the reduction, i.e., Inequality (2), by exhibiting the aforementioned set of instances. In the setting of noisy state discrimination (Problem 2), let the set  $S$  consist of  $\rho_N := \mathbb{I}/2^n$ , the maximally-mixed state; and for  $x \in \{0, 1\}^n$ ,  $\rho_x := |x\rangle\langle x|$ . Consider then solving Problem 1 (weak EM) with the observables  $\mathcal{M} = \{\mathcal{C}^\dagger(Z_i)\}_{i \in [n]}$ , and note that such a solution consists in estimates of the expectation values on the noiseless state,  $\text{Tr}(Z_i \rho_x)$ . If the unknown state had been  $\sigma' = \mathcal{C}'(\rho_x)$  for  $x \in \{0, 1\}^n$ , then  $\text{Tr}(\rho_x Z_i) = 2x_i - 1$ ; while if the unknown state had been  $\sigma' = \mathcal{C}'(\rho_N)$ ,  $\text{Tr}(\rho_N Z_i) = 0 \quad \forall i \in [n]$ . The output of Problem 1 thus completely identifies the label of the state,  $i$ , thus solving Problem 2.  $\square$

In words, the quantity  $(N+1)^{-1} \sum_{k=0}^N D(P_k \| P_N)$  can be understood as a sort of ‘average distance’ between the noisy states  $\mathcal{C}'(\rho_k)$  and the maximally-mixed state (one application of data-processing suffices to see this). This quantity is central to our techniques. The novelty of our results hinges on controlling this quantity precisely, using properties of quantum relative entropies, unital and non-unital quantum noise and Clifford 2-designs to construct  $\mathcal{C}'$  that yields small values of  $\alpha$ .

With this perspective, we are able to establish fundamental information-theoretic limits on error mitigation. Specifically, we are able to answer the following questions:

**How many noisy copies of  $\sigma'$  are required for mitigating depolarizing noise?** One might expect that this number—the *sample complexity* of error mitigation—should scale in the complexity of the circuit,  $\mathcal{C}$  and the amount of noise affecting it. Indeed, existing error mitigation protocols require a number of samples that scales exponentially in the number of gates in the light-cone of the observable of interest, with the exponent depending on the noise levels [11]. Thus, it is natural to ask if such a scaling is unavoidable or one could hope for better protocols.

We start off by considering the complexity of mitigating circuits affected by local depolarizing noise and arbitrary unitary circuits. That is, denoting the noiseless circuit by  $\mathcal{C}$ , the noise channel by  $\mathcal{N}$ , and the noisy output state of the circuit by  $\sigma'_{\mathcal{C}, \mathcal{N}}$ , we obtain the following statement, which has many basic features of our main results:

**Theorem 1** (Appetizer: mitigating depolarizing noise). Let  $\mathcal{A}$  be a weak error mitigation algorithm that mitigates the errors in a  $D$ -layer quantum circuit  $\mathcal{C}$  affected by local depolarizing noise  $\mathcal{N}$  of parameter  $p$ .  $\mathcal{A}$  requires at least  $p^{-\Omega(D)}$  copies of  $\sigma'_{\mathcal{C}, \mathcal{N}}$ .

We refer readers to Theorem 8 for the full statement. Our proof uses the relation to noisy state distinguishability and various properties of the quantum relative entropy to compute  $\alpha$  in Eq. (1). The idea of studying the rate of convergence of a circuit’s output state to the maximally-mixed state is not new: fast convergence is disastrous for error mitigation [25–28], kernel estimation [29] in quantum machine learning, as well as the depth at which quantum advantage from sampling noisy random quantum circuits [30] or variational quantum algorithms [27, 31, 32] can be obtained. However, these results *a la* Theorem 1 share the common feature that they are only able to show an exponential-in-*depth* contraction of the circuit’s output state towards the maximally-mixed state and only show that the distance to the maximally mixed state goes below constant order at logarithmic depth. But we are now able to show that they are not tight in general.

**Theorem 2** (Mitigating depolarizing noise requires exponential-in- $n, D$  samples). Let  $\mathcal{A}$  be a weak error mitigation algorithm that mitigates the errors in an  $n$ -qubit,  $D$ -layer quantum circuit  $\mathcal{C}$  affected by local depolarizing noise  $\mathcal{N}$  of parameter  $p$ . For some parameter  $s > 0$  and depths  $D \geq \Omega(\log^2(n/s))$ ,  $\mathcal{A}$  requires as input at least  $s^{-1} p^{-\Omega(nD/s)}$  copies of  $\sigma'_{\mathcal{C}, \mathcal{N}}$  in the worst case.

We refer readers to Theorem 10 for the full statement. To prove this bound, this time we engineer a family of circuits  $\mathcal{C}$  that converge very fast to the maximally mixed state under noise. This is our technical contribution. By picking, e.g., the parameter  $s = \Omega(n/\log^2(n))$ , we see that even at depths  $D = \text{poly log log}(n)$ , error mitigation requires a number of samples that is super-polynomial in  $n$ . Setting  $s = \mathcal{O}(1)$ , we see that our construction shows that at poly-logarithmic depth already an exponential number of samples is required to perform error mitigation. The circuit construction used to prove this result builds on a construction by Ref. [33] that uses 2-qubit Clifford gates to obtain a 2-design in short depth. While this construction requires all-to-all connectivity, we show that for circuits on a  $d$  dimensional lattice the same effects kick in at depths  $\mathcal{O}(n^{1/d})$ , which is the minimal depth required to ensure that all qubits have a lightcone proportional to the system’s size. Thus, these constructions showcase that, in general, error mitigation protocols will indeed require a number of samples that grows exponentially with the number of gates in the light-cone of observables.

**How many noisy copies of  $\sigma'$  are required for mitigating non-unital noise?** Our previous bounds relied critically on the structure of depolarizing noise—namely, the fact that a depolarizing noise channel can be expressed

as a linear combination of Pauli noise channels. Moving beyond the toy model of depolarizing noise, we are also able to show sample complexity bounds for circuits affected by non-unital noise in another simplified setting. We show that also for non-unital noise, whenever a family of circuits is highly entangling, which we model through the assumption that it forms a 2-design, mitigating local noise typically requires a number of samples that is exponential both in the number of channel applications and number of qubits.

**Theorem 3** (Mitigating non-unital noise also requires exponential-in- $n$ ,  $D$  samples). *Let  $\mathcal{A}$  be a weak error mitigation algorithm that mitigates the errors in an  $n$ -qubit,  $\ell$ -layer quantum circuit  $\mathcal{C}$  affected by non-unital product noise  $\mathcal{N} = \otimes_{i=1}^n \mathcal{N}_i$ , such that the noisy circuit takes the form  $T_D = \bigcirc_{t=1}^{\ell} \mathcal{N} \circ \mathcal{U}_t$ , where  $\mathcal{U}_t$  are drawn independently from a 2-design. Then, in expectation,  $\mathcal{A}$  requires as input at least  $c^{-\Omega(n\ell)}$  copies of  $\sigma'_{\mathcal{C},\mathcal{N}}$  for some constant  $c$  that depends on the quantum channel  $\mathcal{N}$ .*

The reader may refer to Theorem 13 for the full statement. The proof of this statement is based on computing the expected value of the overlap of two quantum states after we apply a unitary stemming from a 2-design followed by a noisy channel. From that, we obtain a formula for the expected overlap after  $D$  uses that only depends on the initial overlap and constants that depend on the noisy channel. From these computations we can then infer the expected Hilbert Schmidt distance between two inputs of the noisy circuits and use it to upper-bound the trace distance.

While our model of only applying local noise after a (global) unitary is simplified, to our knowledge this is the first result to study error mitigation under non-unital noise beyond the setting of the variational algorithms studied in Ref. [27]. With Theorem 3, we expect our results under depolarizing noise to also transfer to the non-unital case: as long as the underlying noisy circuit is highly entangling and errors are not corrected, the cost of error mitigation is exponential in the depth and the number of qubits.

Theorems 1, 2, 3 were proven in the setting of weak error mitigation, assuming the error mitigation algorithm does not use knowledge of the input state to the circuit. While this is the setting most often used in practice, it does not cover all proposals in the literature. We now ask if our results can be extended to close variants of this setting which are also practically relevant.

**How does fixing the input state change the picture?** Our definition of error mitigation enables us to derive strong sample complexity lower bounds using tools from statistical inference, but it comes at the cost of requiring that error mitigation does not use any information about the input state to the noisy circuit—hence that it is *input-state agnostic*. Does lifting this restriction—thereby including *input-state aware* mitigation protocols—change the picture significantly? A first observation is that now, we can no longer rule out the possibility of successful error mitigation with sub-exponential worst-case sample complexity—simply because classical simulation with no resort to the noisy quantum device at all is a valid error mitigation algorithm that takes zero samples. However, we prove that an error mitigation algorithm that makes use of a noisy quantum device must use exponentially many copies of the noisy state to produce an output meaningfully different from a procedure which does not invoke the quantum device:

**Theorem 4** (Resource cost of successful error mitigation). *A successful weak error mitigation algorithm  $\mathcal{A}$  must use  $m = c^{\mathcal{O}(nD)}$  copies of the noisy output state  $\sigma'$  in the worst case or there there exists an equivalent algorithm  $\mathcal{A}'$  with purely classical inputs such that the output of  $\mathcal{A}'$  is indistinguishable from the output of  $\mathcal{A}$ .*

For the full statement, see Theorem 9. In this way, we show that our constructions apply both to the input state-agnostic and input state-aware setting, with, however, different implications.

**Does strong error mitigation imply weak error mitigation, or vice versa?** We also study the relationship between the two types of error mitigation algorithms we have defined (weak and strong). We ask if the output of one is sufficient to obtain the output of the other. One direction of this question is easy to answer—classically, samples of a probability distribution suffice to estimate expectation values of bounded functions. The quantum generalization of this is that for local observables, strong error mitigation implies weak error mitigation. However, in some cases, weak error mitigation of noisy circuits does not quite achieve the algorithmic goal: for instance in solving hard combinatorial optimization problems, at this point almost a canonical proposed application of noisy quantum devices, a class that includes the famous *quantum approximate optimization algorithm* (QAOA) [18, 34, 35]. In that case, one is not only necessarily interested in the optimal value of the cost function (which one could obtain through weak error mitigation), but also in an assignment that achieves that value (which could be obtained through strong error mitigation). This then motivates the question if it is possible to obtain strong error mitigation (i.e., samples) from weak error mitigation (i.e., expectation values). We give a partial negative answer that rules out certain types of protocols:

**Theorem 5** ((Informal statement) Exponentially-many observables (in the same eigenbasis) are needed to output samples from the same basis). *Suppose we have a weak error mitigation algorithm that, given a set of observables  $\mathcal{M} = \{O_i\}_i$  in the same eigenbasis, outputs their expectation value estimates  $\{\hat{o}_i\}$ . In general, at least  $\exp(n)$ -many distinct  $\hat{o}_i$ s must be queried by any algorithm that takes as input the  $\{\hat{o}_i\}$  and outputs samples from their eigenbasis.*

We refer readers to Theorem 7 for the full statement. The proof of this statement takes the perspective of expectation value estimates as *statistical queries* [36–38], and leverages existing results on optimality guarantees of hypothesis selection [39] and on the statistical query hardness of learning a class of distributions known as PARITIES.

## B. Consequences of our results

### 1. Loss of quantum advantage at $\log \log n$ depth

In the course of proving Theorem 2, we prove the existence of very rapidly mixing circuits: circuits whose output states converge exponentially fast, in both number of qubits ( $n$ ) and circuit depth ( $D$ ), to the maximally-mixed state. The idea of studying the rate of convergence (according to various distance measures) of a circuit’s output state to the maximally-mixed state is not new: fast convergence is disastrous for error mitigation [25–28], kernel estimation [29] in quantum machine learning, as well as the depth at which quantum advantage from sampling noisy random quantum circuits [30] or variational quantum algorithms [27, 31, 32] can be obtained. However, these results share the common feature that they are only able to show an exponential-in-*depth* contraction of the circuit’s output state towards the maximally-mixed state and only show that the distance to the maximally mixed state goes below constant order at logarithmic depth. Furthermore, one can show that such results are tight for trivial circuits consisting solely of the identity gate.

We point out that the applications mentioned above are all intended to be run on *noisy intermediate-scale quantum* (NISQ) processors. The quantum circuits employed in these applications are shallow, which means that their depth scales like  $d = O(\log(n))$ , implying that their claimed rate of convergence is inverse polynomial in  $n$ . Our results thus imply an exponentially-faster rate of convergence than previous thought. In fact, a variation of our main result shows that even at  $\log \log(n)$  depth, increasing the number of qubits already brings about a superexponential drop in distinguishability from the maximally-mixed state, with all the above-mentioned attendant implications.

The consequence of this is two-fold: Firstly, we open up a new research direction: how does the amount of entanglement generated by a quantum circuit relate to noise sensitivity? Indeed our circuit construction comprises highly-entangling gates, showcasing the intuition that such circuits are exponentially more sensitive to noise than are general circuits. Secondly, and more pragmatically, we now have quantum circuits that converge to the uniform distribution superpolynomially fast even at  $\log \log(n)$  depth, implying that for all the above applications (robustness to errors, kernel estimation, random circuit sampling, variational quantum optimization) quantum advantage is lost exponentially earlier than originally anticipated.

### 2. Extending previous distinguishability bounds for unital noise to non-unital noise

In addition, all of the above references can only show exponential convergence under the assumption that the quantum circuit is affected by unital noise. In our work, we go one step further and establish results for a toy model of circuits affected by non-unital noise interspersed by global 2-designs in Section V. We show that even when the noise is non-unital, we are still able to obtain a decay in distinguishability between different inputs that is exponential in both the depth and the number of qubits. Such an extension is important because many physically relevant noise models, such as amplitude damping, are not unital. Our toy model illustrates that as long as the circuit is entangling enough the decay in distinguishability will be exponential in number of qubits.

### 3. A construction of random quantum circuits whose distance to the ground state decays exponentially in both depth and system size

Additionally, in order to prove Theorem 2, we do not just construct a single circuit, but a random *ensemble* of circuits. We are able to prove that, on expectation, their relative entropy from the maximally-mixed state displays the stated exponential decay. A result with a particularly close setting to this one is that of Ref. [30], which has shown that *on expectation*, the total variation distance between a noisy random quantum circuit output distribution  $\mathcal{D}$  and the uniform distribution decays exponentially in depth only, that is

$$\mathbb{E}_{\mathcal{B}}[\|\mathcal{D} - U\|_{TV}] = \Omega(\exp(-CD)) \quad (3)$$

where  $\mathcal{B}$  is an ensemble of random quantum circuits that differs from the one we consider. Indeed, the authors of [30] consider the model of applying uniformly random 2 qubit Clifford gates at each step, whereas we consider more



structured random circuits that mix faster.

Let us quickly give an (over)simplified summary of the proof of Eq. (3). Suppose for simplicity that we are actually considering an ensemble of random Clifford circuits. A constant fraction of Clifford gates acting on 2 qubits are a product of 1 qubit gates. Thus, at each step there is a constant probability  $q > 0$  that the gate we pick to act on qubit 1 is product. As the gates at different layers are independent, at depth  $D$  with probability at least  $q^D$  the first qubit of the system will be in a product state with the rest. And in that case the effect of the noise on that qubit cannot depend on the system's size. Thus, we see that in some sense the argument relies in conditioning on the circuit not being too entangling. However, our examples show that if the circuits are picked from a random ensemble that is highly entangling, then the effect of the noise is much more pronounced and the decay can also depend on the system's size.

#### 4. Consequences for near-term applications

It is hoped that applying currently available noise near-term quantum devices to computational problems can bring forth some sort of quantum advantage before the era of quantum error correction and fault-tolerance is reached. Error mitigation is touted as a means towards this end. However, our results further strengthen already known obstructions for these methods.

When optimizing the energy of a Hamiltonian, the outputs of noisy quantum circuits concentrate strongly [27, 31] which causes any possibility of quantum advantage to be lost when the depth exceeds  $\Omega(p^{-1})$ , an argument that follows a similar route as our work. Our contributions show that this generic bound can be loose in the worst case where a system size dependent error probability  $p = \mathcal{O}((nD)^{-1})$  is required to ensure potentially good solutions. These arguments go hand in hand with the fact that any circuit that aims at preparing the ground state has to have at least logarithmic depth for important classes of local Hamiltonians [27, 40–43]. For highly entangled ground states of non-local Hamiltonians, we expect even larger lower bounds on the depth, as entanglement has to be built between distant sites. Building entanglement, however, is also the key of our construction for a system-size dependent decay to the maximally mixed state.

Together these statements imply that noisy variational ground state preparation is mostly a lose-lose proposition: Either the depth is insufficient to reach a good approximation of the ground state or it is so high that noise takes over.

#### 5. No exponential quantum speedups to estimate expectation values on noisy devices

Having defined the task of estimating the expectation value of a noiseless circuit, the question of successful error mitigation immediately turns into a question about its complexity relative to noisy quantum devices. Our results show that, typically, the number of samples from a noisy device required to estimate the expectation value of the clean circuit will scale exponentially in the number of gates in the light-cone of the observable and therefore rules out the possibility of exponential quantum advantages for this task on noisy devices, as there are classical algorithms that exhibit the same scaling in complexity [44–46]. As such, one can only hope for better exponents in the worst-case complexity.

### C. Related work

As a critical enabler in the IBM road map for near-term quantum computation, error mitigation has been the subject of intense study in the past few years [9–16]. Prior art typically falls into two camps: ‘go’ results, which propose new error mitigation algorithms and (often numerically) study their guarantees, and ‘no-go’ theorems, which demarcate the resource limitations of these algorithms. In Section II B, we will show that our framework encompasses many ‘go’ algorithms. While our main results fall into the latter ‘no-go’ category, they are by no means an island. We will now review other findings in this category. Before mentioning more quantitative arguments identifying limitations of quantum error mitigation, it is worth noting that there is a body of heuristic evidence of instances of quantum error mitigation to perform poorly in certain cases. For example, Ref. [9] discusses limitations of the extrapolation method to the noise-free case being unable to approach the mean of the error-free circuit.

More specifically related to the concrete results of our work, Refs. [25] and [26] also study the sample complexity of weak error mitigation under depolarizing noise. They do not mention strong error mitigation. Similar to us, their bounds are information-theoretic, and arise from considering how noise affects the distinguishability of quantum states. We clarify that their sample complexity bounds do not apply to input state-aware protocols if our metric of

success is approximating the expectation value of a set of observables. For these papers implicitly assume knowledge of the noiseless circuit—and with knowledge of the input state on top of that, an error mitigation algorithm could produce expectation value estimates without taking even a single sample, simply by simulating the circuit on the input state, even if it takes an exponential amount of computation. Thus, we need to be a bit more careful when we define in which sense error mitigation is doomed to fail in the case where the observables, noiseless circuit and input state are known to the algorithm.

Ref. [26] consider two desiderata for the error mitigation algorithm: a metric, similar to our Eq. (27), that requires the algorithm to be probably-approximately correct on *each* observable [47]; and the maximum bias/standard deviation of the estimators. While they also study the relative entropy decay to the maximally-mixed state, their ultimate sample complexity lower bound decays exponentially only in the *depth*  $D$  of the circuit—and polynomially in the number of qubits  $n$ . By contrast, our sample complexity lower bounds (Theorem 10) decay *exponentially* in both  $n$  and  $D$ —indicating that their bounds are exponentially loose. Moreover, we make an incursion into proving our results for non-unital noise models. At the time of writing, Appendix G of Ref. [26], which attempts to do the same, contained an error: Eq. (G6) claims that the channel composed by interleaving the layers of non-unital noise with the layers of the noiseless circuit, contracts towards the fixed point. However, it is unclear how they obtain the stated contraction coefficient, as when the fixed point of the noise is not the identity, one cannot remove the intermediate unitaries to obtain the contraction coefficient.

Ref. [48] studies weak error mitigation from the perspective of Fisher information. More precisely, the authors show that the quantum Fisher information associated with layered noisy circuits decays exponentially with circuit depth, which thus also lower bounds the variance of any unbiased estimator of expectation values. However, we have our reservations about using estimator variance as a metric of resources required, as it gives a very loose bound on the sample complexity to attain a certain estimation error. Take, for instance, the random variable which is  $e^n$  with probability  $e^{-n/2}$ ,  $-e^n$  with probability  $e^{-n/2}$  and 0 otherwise. This zero-mean random variable has exponential variance, but with a sub-exponential number of samples, the empirical mean is overwhelmingly likely to be exactly the true mean. Of course, such pathological examples use the fact that the underlying random variables take exponentially large values, but this is also the case for some error mitigation protocols.

#### D. Organization of this work

This work is organized as follows:

- In Section II, we introduce our error mitigation setting as well as two variants of error mitigation that we call weak and strong respectively. We further argue that this setting encompasses many error-mitigation protocols used in practice, including *virtual distillation* [20, 21], *Clifford data regression* (CDR) [17], *zero-noise extrapolation* (ZNE) [10] and *probabilistic error cancellation* (PEC) [10, 11, 22]. While most theoretical analyses of error mitigation have focussed exclusively on the weak error mitigation setting, here we argue with reference to the practical protocols that the strong setting is equally relevant, and we make a first attempt at relating the two settings. We also show that many existing error mitigation protocols fit into the model that we will work with.
- Section III presents our conceptual contribution. Here we prove the ‘Appetizer’ theorem 8 which studies the limitations of error mitigation under depolarizing noise. While this Appetizer theorem implies a weaker scaling of error mitigation complexity as compared to our main theorems, we elect to start here as the proof already introduces the conceptual ‘outer loop’ that we will use in the rest of our theorems, namely the application of the generalized Fano’s lemma to relate error mitigation to the noisy state discrimination problem.
- In Section IV we present our main theorem (Theorem 10), and fill out the technical ‘inner loop’ of our argument under depolarizing noise. The proof of Theorem 10 is our first technical contribution: the worst-case circuit ensemble that we construct, which has the property that the relative entropy of its average output state to the maximally-mixed state—and hence error mitigation’s sample complexity—scales, in the worst case, exponentially in both  $n$  and  $D$ . We conclude the section by further refining our results so that they hold for local circuits and even lower depths than before.
- Section V contains our second technical contribution: we extend all our previous conclusions to error mitigation under a new class of noise models, non-unital noise, although under the toy model of local non-unital local noise interspersed by global 2–designs. We arrive at the same conclusions in this toy model, the sample complexity for error mitigation is exponential.
- Finally we conclude with an outlook in Section VI.

## II. ERROR MITIGATION: THE LAY OF THE LAND

The aim of an error mitigation procedure is to produce a representation of the output of a noiseless quantum circuit given access to the actual noisy quantum device. Before we formally define different error mitigation tasks, we outline the model for noisy quantum circuits that is used throughout this work. We consider quantum circuits of depth  $D$  acting on  $n$  qubits. In the quantum channel picture, such circuits take the form

$$\mathcal{C} = \mathcal{C}^{(D)} \circ \dots \circ \mathcal{C}^{(2)} \circ \mathcal{C}^{(1)} \quad (4)$$

where each  $\mathcal{C}^{(1)}, \dots, \mathcal{C}^{(D)}$  denotes a layer of unitary quantum gates. We will then assume that in the noisy version, a quantum channel will act after each unitary we implement. That is, for such a circuit, we take its noisy version to be

$$\Phi_{\mathcal{C}, \mathcal{N}^{(D)}, \dots, \mathcal{N}^{(1)}} = \mathcal{N}^{(D)} \circ \mathcal{C}^{(D)} \circ \dots \circ \mathcal{N}^{(2)} \circ \mathcal{C}^{(2)} \circ \mathcal{N}^{(1)} \circ \mathcal{C}^{(1)}, \quad (5)$$

where  $\mathcal{N}^{(1)}, \mathcal{N}^{(2)}, \dots, \mathcal{N}^{(D)}$  are quantum channels that describe the noise.

For simplicity, we will often consider the case in which every layer has an identical noise channel acting on it; in that case we will denote the noisy version of the circuit as  $\Phi_{\mathcal{C}, \mathcal{N}}$ , or in cases when the noise channel  $\mathcal{N}$  can be described by a parameter  $p$ ,  $\Phi_{\mathcal{C}, p}$ . However, our results extend to the case where the noise is not uniform, or not unital. For pedagogical reasons, our results are initially derived for local depolarizing noise and then extended to other noise models.

### A. Preliminaries: Relative entropy and derived quantities

At the heart of many of our technical arguments are notions of relative entropies—measures of distance between two quantum states—and how they contract under noise. We first define distance measures on classical probability distributions. Let  $R, S$  be two probability measures on the same support  $\mathcal{X}$ . It will suffice for our purposes to let  $\mathcal{X}$  be a finite set.

- **Total variation (TV) distance.** This is defined as

$$d_{TV}(R, S) := \sup_{A \subseteq \mathcal{X}} |R(A) - S(A)| = \frac{1}{2} \sum_{x \in \mathcal{X}} |R(x) - S(x)|. \quad (6)$$

- **Kullback-Leibler (KL) divergence, or classical relative entropy.** The classical relative entropy is defined as

$$D(R||S) := \sum_{x \in \mathcal{X}} R(x) \log \frac{R(x)}{S(x)} = \mathbb{E}_{x \sim R} \left[ \log \frac{R(x)}{S(x)} \right], \quad (7)$$

where throughout the manuscript we take log to be base 2.

Now we introduce distance measures on quantum states. The primary such measure we will consider is quantum relative entropy, which can be understood as a quantum generalization of classical *KL divergence*. For this reason, we will use the same notation,  $D(\cdot||\cdot)$ , for both quantum relative entropy and classical relative entropy.

Let  $\rho, \sigma$  be two quantum states in  $D(\mathcal{H}_n)$  (though in general these quantities are defined with  $\sigma$  any positive semi-definite operator). We will use the following divergences:

- **Relative entropy.** If  $\text{supp}(\rho) \subseteq \text{supp}(\sigma)$ , we define the (quantum) relative entropy as

$$D(\rho||\sigma) := \text{Tr}(\rho \log \rho) - \text{Tr}(\rho \log \sigma). \quad (8)$$

- **Petz-Rényi relative entropy.** For a parameter  $\alpha \in (0, 1) \cup (1, \infty)$ ,

$$D_\alpha(\rho||\sigma) := \frac{1}{\alpha - 1} \log \text{Tr} [\rho^\alpha \sigma^{1-\alpha}]. \quad (9)$$

In the limit  $\alpha \rightarrow \infty$ , the Petz-Rényi relative entropy reduces to the **max-relative entropy**,

$$D_\infty(\rho||\sigma) = D_{\max}(\rho||\sigma) := \inf \{ \gamma : \rho \leq 2^\gamma \sigma \} \quad , \quad (10)$$

while in the limit  $\alpha \rightarrow 1$ , it reduces to the quantum relative entropy, i.e.,  $D(\rho\|\sigma)$ .

It will often be illuminating to fix the second argument of the relative entropy to be the maximally mixed state, while putting the state of interest in the first argument of the relative entropy. Relative entropies of this form may even be upper-bounded explicitly in terms of quantities relating to the state of interest, as we now show the following.

**Lemma 2** (Purity controls relative entropy to the maximally mixed state). *For any  $\rho$  on  $n$  qubits,*

$$D\left(\rho\left\|\frac{\mathbb{I}}{2^n}\right.\right) \leq n + \log(\text{Tr}(\rho^2)). \quad (11)$$

*Proof.* The statement follows from the fact that relative entropy is upper-bounded by 2-Rényi relative entropy

$$D_2(\rho\|\sigma) := \log \text{Tr}[\rho^2 \sigma^{-1}]. \quad (12)$$

This can be seen from the fact that the Petz-Rényi relative entropies for  $\alpha \in (0, 1) \cup (1, \infty)$  satisfy an ordering property [49], i.e.,

$$\text{for } \alpha > \beta > 0: \quad D_\alpha(\rho\|\sigma) \geq D_\beta(\rho\|\sigma), \quad (13)$$

and

$$\lim_{\alpha \rightarrow 1} D_\alpha(\rho\|\sigma) = D(\rho\|\sigma). \quad (14)$$

For any  $\rho$  on  $n$  qubits and  $\sigma$  the maximally mixed state, we thus have

$$D\left(\rho\left\|\frac{\mathbb{I}}{2^n}\right.\right) \leq D_2\left(\rho\left\|\frac{\mathbb{I}}{2^n}\right.\right) = \log \text{Tr}\left(\rho^2 \left(\frac{\mathbb{I}}{2^n}\right)^{-1}\right) = n + \log(\text{Tr}(\rho^2)) \quad (15)$$

as stated.  $\square$

## B. Relation to practical error mitigation protocols

In this subsection, we list some broad classes of error mitigation protocols that have been proposed in the literature, briefly explain how they work, and describe to what our extent our results apply to them. In all of the following,  $\omega_{\mathcal{C}, \mathcal{N}}$  refers to the output state of the circuit affected by noise, while  $|\psi\rangle$  refers to the output state of the noiseless circuit.

1. *Virtual distillation* [20, 21]: This protocol takes as input  $k$  copies of the noisy circuit output state  $\omega_{\mathcal{C}, \mathcal{N}}$  outputs  $\omega_{\mathcal{C}, \mathcal{N}}^k / \text{Tr}\{\omega_{\mathcal{C}, \mathcal{N}}^k\}$  (instead of  $|\psi\rangle\langle\psi|$ ) with probability  $\text{Tr}\{\omega_{\mathcal{C}, \mathcal{N}}^k\}$ . The intention of this is to amplify the dominant eigenvector in the eigen-decomposition of  $\omega_{\mathcal{C}, \mathcal{N}}$ , as under some assumptions, this dominant eigenvector has a large overlap with the noiseless output state vector  $|\psi\rangle\langle\psi|$ .

For circuits affected by unital noise, Ref. [27] has shown that the probability of success of virtual distillation decays exponentially both in the number of qubits and depth of the circuit. Here we will show that the same holds for local non-unital noise for circuits that form a 2-design. As we show in Section VC, after we apply a 2-design followed by one layer of product noise, the purity of the system is already exponentially small in expectation. This allows us to conclude that also the success probability of virtual distillation will be exponentially small.

2. *Clifford data regression* (CDR) [17]: One runs classically simulatable circuits, in this case Clifford circuits with finitely many  $T$ s, to generate the training pairs  $(X_i^{\text{noisy}}, X_i^{\text{exact}})_i$ , which are expectation values of the observable on noisy and noiseless circuits. The noisy expectation values are obtained via running the actual circuits; the noiseless expectation values are obtained by classical simulation. One then makes the ansatz that  $X_i^{\text{noisy}}$  and  $X_i^{\text{exact}}$ , for classes of observables or circuits, are related via some functional relation—Ref. [17] uses a linear ansatz—and fits the parameters of the ansatz using the training pairs.

Let  $\mathcal{C}_1, \dots, \mathcal{C}_r$  be the  $r$  distinct Clifford circuits that are run in the process of generating the training data, and suppose each circuit is run  $g$  times and measured after each run. Let **Train** be the algorithm that generates the training data—that is, **Train** takes in

$$\omega_{\mathcal{C}_1, \mathcal{N}}^{\otimes g} \otimes \omega_{\mathcal{C}_2, \mathcal{N}}^{\otimes g} \otimes \dots \otimes \omega_{\mathcal{C}_r, \mathcal{N}}^{\otimes g}, \quad (16)$$

(here we do not specify the input to the circuits used to generate the training data because it is irrelevant), makes a quantum measurement on  $\omega_{\mathcal{C}_1, \mathcal{N}}^{\otimes g}$  to generate  $X_i^{\text{noisy}}$  and classically simulates it to generate  $X_i^{\text{exact}}$ , for all  $i$ . Further, let  $\mathcal{A}$  be the error mitigation algorithm that takes as input  $m$  copies of  $\Phi_{\mathcal{C}, \mathcal{N}}$ , the noisy output state of the actual circuit  $\mathcal{C}$ , and the training data (which is the output of  $\text{Train}$ ), and outputs the expectation values  $\hat{O} \in \mathbb{R}^m$  where  $\hat{O}[i]$  is the estimate for  $\langle \psi | O_i | \psi \rangle$ . We denote the probability distribution of this algorithm over outputs  $\hat{O}$  when the circuit input is  $|k\rangle\langle k|$  as  $P_k$ . Then

$$D(P_k \| P_N) = D \left( \mathcal{A} \left( \Phi_{\mathcal{C}, \mathcal{N}, \rho_k}^{\otimes m} \otimes \text{Train}(\omega_{\mathcal{C}_1, \mathcal{N}}^{\otimes g} \otimes \omega_{\mathcal{C}_2, \mathcal{N}}^{\otimes g} \otimes \dots \otimes \omega_{\mathcal{C}_r, \mathcal{N}}^{\otimes g}) \right) \left\| \mathcal{A} \left( \frac{\mathbb{I}}{2^n}^{\otimes m} \otimes \text{Train}(\omega_{\mathcal{C}_1, \mathcal{N}}^{\otimes g} \otimes \omega_{\mathcal{C}_2, \mathcal{N}}^{\otimes g} \otimes \dots \otimes \omega_{\mathcal{C}_r, \mathcal{N}}^{\otimes g}) \right) \right. \right) . \quad (17)$$

$$\begin{aligned} &\leq D \left( \Phi_{\mathcal{C}, \mathcal{N}, \rho_k}^{\otimes m} \otimes \text{Train}(\omega_{\mathcal{C}_1, \mathcal{N}}^{\otimes g} \otimes \omega_{\mathcal{C}_2, \mathcal{N}}^{\otimes g} \otimes \dots \otimes \omega_{\mathcal{C}_r, \mathcal{N}}^{\otimes g}) \left\| \frac{\mathbb{I}}{2^n}^{\otimes m} \otimes \text{Train}(\omega_{\mathcal{C}_1, \mathcal{N}}^{\otimes g} \otimes \omega_{\mathcal{C}_2, \mathcal{N}}^{\otimes g} \otimes \dots \otimes \omega_{\mathcal{C}_r, \mathcal{N}}^{\otimes g}) \right. \right) \\ &\leq mD \left( \Phi_{\mathcal{C}, \mathcal{N}, \rho_k} \left\| \frac{\mathbb{I}}{2^n} \right. \right) + D \left( \text{Train}(\omega_{\mathcal{C}_1, \mathcal{N}}^{\otimes g} \otimes \omega_{\mathcal{C}_2, \mathcal{N}}^{\otimes g} \otimes \dots \otimes \omega_{\mathcal{C}_r, \mathcal{N}}^{\otimes g}) \left\| \text{Train}(\omega_{\mathcal{C}_1, \mathcal{N}}^{\otimes g} \otimes \omega_{\mathcal{C}_2, \mathcal{N}}^{\otimes g} \otimes \dots \otimes \omega_{\mathcal{C}_r, \mathcal{N}}^{\otimes g}) \right. \right) \\ &\leq mD \left( \Phi_{\mathcal{C}, \mathcal{N}, \rho_k} \left\| \frac{\mathbb{I}}{2^n} \right. \right) , \end{aligned} \quad (18)$$

where the first inequality follows by data processing and the second follows from the tensorization of relative entropy. Plugging Eq. (18) into the Central Lemma (Lemma 1), we then get the desired result.

3. *Zero-noise extrapolation (ZNE)* [10]: Given the circuit  $\mathcal{C}$  with noise strength  $\lambda$ , one runs it at  $n$  different *amplified* noise levels  $c_j \lambda$  ( $c_j > 1$ ), computes expectation values on each such circuit, and combines the noisy estimates into an estimate for the expectation value without noise. Variations on the theme include considering other noise models and interpolation schemes [11, 50, 51].

We comment now on how our results apply to the proposal of Ref. [10]. The extension to other variations of this scheme should be similar. Here, to run a circuit at noise level  $c\lambda$  where  $c \geq 1$ , each gate is run for a time  $cT$ , where  $c = 1$  corresponds to the noise rate in the original circuit. Ref. [10] further prescribes that the final estimate should be computed as

$$\hat{E} = \sum_{j=1}^n \gamma_j \hat{E}(c_j \lambda) \quad , \quad (19)$$

where  $\sum_{j=0}^n \gamma_j = 1$  and  $\sum_{j=0}^n \gamma_j c_j^k = 0$  for  $k = 1, \dots, n$ . If the evolution of the original circuit is

$$\frac{\partial}{\partial t} \rho = -i[K(t), \rho] + \lambda \mathcal{L}(\rho) \quad , \quad (20)$$

where  $K(t)$  is the coherent circuit and  $\mathcal{L}$  represents the noise we want to mitigate, then stretching all gate times by  $c$  has the effect of replacing  $\lambda \leftarrow c\lambda$  for  $c > 1$  in the above equation. In the quantum circuit picture, if we assume that each circuit is acted upon by local depolarizing noise with depolarizing parameter  $p$  after each gate, this has the effect of replacing the parameter  $p \leftarrow p^c$ . Let us assume that  $c$  is the minimum noise level at which the circuit is run. Indeed, it cannot be assumed that we can pick  $c$  arbitrarily small, as in this case error mitigation would not be required. Then, letting  $P_k$  denote the probability distribution over  $\hat{E}$  when the input to the circuit is  $|k\rangle\langle k|$ , we can make a similar replacement in Theorem 10 to obtain that

$$\frac{1}{N+1} \sum_{k=0}^N D(P_k \| P_N) \leq p^{cnD} \quad . \quad (21)$$

The Central Lemma then applies.

4. *Probabilistic error cancellation* [10, 11, 22]: In this protocol, one rewrites a noiseless circuit  $\mathcal{U}$  in terms of a linear combination of noisy circuits  $\mathcal{F}_i$ :  $\mathcal{U} = \sum_i a_i \mathcal{F}_i$  where  $a_i \in \mathbb{R}$  could be negative. Similarly, any desired

expectation value of interest (of the observable  $M$ , say) can be decomposed as the following convex combination:

$$\mathrm{Tr}(M\mathcal{U}(\cdot)) = \sum_i a_i \mathrm{Tr}(M\mathcal{F}_i(\cdot)) = \gamma \sum_i p_i \mathrm{sign}(a_i) \mathrm{Tr}(M\mathcal{F}_i(\cdot)) \quad . \quad (22)$$

Then with probability  $p_i$ ,  $\mathrm{Tr}(M\mathcal{F}_i(\cdot))$  is estimated by running the noisy circuit  $\mathcal{F}_i$   $g$  times. This is repeated  $r$  times, the results are averaged and the estimate is multiplied by  $\gamma$ .

Our analysis applies here too, because we can model the computation of the desired quantity (the right-hand-side of Eq. (22)) with a quantum channel  $\mathcal{M}$ , as follows: Consider the state

$$\omega_k^{\otimes r} = \bigotimes_{j=1}^r \left( \sum_i p_i |i\rangle\langle i|_{A_j} \otimes (\mathcal{F}_i(|k\rangle\langle k|)^{\otimes g})_{B_j} \right) \quad , \quad (23)$$

and the channel  $\mathcal{M}$  which consists in 1) Applying  $\bigotimes_j P_{A_j}$  on this state, where  $P_{A_j}$  is a projector on the classical system  $A_j$ ; 2) On each of the  $r$  systems, measuring  $\mathcal{M}^{\otimes g}$  on  $B_j$  and averaging the result; 3) Classically averaging the results across the  $r$  systems and multiplying by  $\gamma$ .

Then let the output of this channel be denoted by  $P_k$ , for  $k \in \{0, 1\}^n$ , while let  $P_N$  denote  $\mathcal{M}$  acting on the state

$$\sigma^{\otimes r} = \bigotimes_{j=1}^r \left( \sum_i p_i |i\rangle\langle i|_{A_j} \otimes \left( \frac{\mathbb{I}}{2^n} \right)_{B_j} \right) \quad . \quad (24)$$

Then we follow similar steps as before to obtain

$$\begin{aligned} D(P_k \| P_N) &= D(\mathcal{M}(\omega_k^{\otimes r}) \| \mathcal{M}(\sigma^{\otimes r})) \leq r D(\omega_k \| \sigma) \\ &\leq r \sum_i p_i D\left(|i\rangle\langle i| \otimes (\mathcal{F}_i(|k\rangle\langle k|)^{\otimes g}) \left\| |i\rangle\langle i| \otimes \left(\frac{\mathbb{I}}{2^n}\right)\right.\right) \\ &= rg \sum_i p_i D\left(\mathcal{F}_i(|k\rangle\langle k|) \left\| \frac{\mathbb{I}}{2^n}\right.\right) . \end{aligned} \quad (25)$$

The Central Lemma then applies.

5. *Variable noise Clifford data regression* (vnCDR) [52]: This work generalizes ZNE and CDR. Given an input circuit, one constructs as similar as possible classically simulatable circuits, i.e., Clifford + const.  $T$ , and runs them on different noise levels, as in ZNE. Then, again one performs regression on this data to obtain a correction function (which the authors again take to be a simple linear function). Such protocols are input state-aware, but they can be analyzed using a combination of our previous tools and the framework we propose in Section III C to deal with the input state dependence.

### C. Error mitigation setting

The protocols listed in Subsection II B are but a slice of the wealth of error mitigation techniques that have so far been proposed. In this section we aim to unify all of these techniques in introducing the model of error mitigation that we will be using in the rest of this work. As explained in Subsection II B, lower bounds against the model we will introduce imply lower bounds for many of those protocols. We will take the input of an error mitigation protocol to be the following.

**Definition 1** (Resources for error mitigation). *An error mitigation algorithm predicts properties of the noiseless output state of a quantum circuit  $\mathcal{C}(\rho)$  given:*

- *Classical descriptions of  $\mathcal{C}$  and the noise channel  $\mathcal{N}$  acting on  $\mathcal{C}$ .*
- *(Optional) A classical description of the input state to the circuit,  $\rho$ . If the algorithm utilizes this classical description, we call the algorithm input-state aware. If the algorithm doesn't utilize this classical description, we call it input-state agnostic.*

- *The ability to perform collective quantum measurements on multiple copies of the noisy circuit output state  $\Phi_{\mathcal{C},\mathcal{N}}(\rho)$  (see the remarks below for the case of randomized families of circuits).*

We now make several remarks about this definition.

Firstly, our goal is to demarcate the information-theoretic limits of error mitigation. This we do by studying the *sample complexity* of error mitigation in our model—how many copies of the noisy output state are required to achieve the desired error mitigation output—rigorously quantifying how this number scales in the complexity of  $\mathcal{C}$  and the noise  $\mathcal{N}$ . We note that sample complexity lower bounds computational complexity, and so lower bounds for sample complexity are also lower bounds for computational complexity.

Secondly, the practical scope of our model is broader than meets the eye. It encompasses even those protocols that run *multiple* different circuits with varying levels of noise, or take as additional input ‘training data’ which are pairs of (experimental) noisy and (simulated) noiseless expectation values for different circuits. In this case, we can always identify a representative circuit/noise level out of the class of circuits that may be run; it is this circuit whose parameters will determine the complexity of error mitigation. The reader is referred to the discussion in the Subsection II B for more details.

Thirdly, we start in Section III by proving lower bounds against input-state *agnostic* protocols, and then in Section III C we extend the results to input-state *aware* protocols. Input-state *agnostic* protocols are a natural starting point, because they cannot possibly work by simulating the circuit  $\mathcal{C}$ , no matter how shallow that circuit is—simply because they do not know the input to the circuit. In fact, many existing error mitigation protocols in the literature are already covered by the input state-agnostic model, because the classical description of the input state is not a parameter in the protocol.

Fourthly, note that we have allowed for the ability to perform arbitrary collective measurements. This requires access to a quantum memory, a nontrivial quantum resource that might be out of reach. Because of this, most error mitigation protocols in the literature do not require this ability. However, our conclusions hold up *even* against algorithms that have such additional power.

Error mitigation is not an end in itself; typically it is used as the last step in a pipeline to solve some (quantum) computational task. Depending on what kind of task that is, different outputs of the mitigation procedure are required. Arguably, the most popular intended application for near-term quantum computers are *variational quantum algorithms* [18, 53, 54] where a quantum state is prepared through a parametrized quantum circuit and the parameters are iteratively adjusted to optimize a function  $L(\langle O_1 \rangle, \langle O_2 \rangle, \dots)$  of expectation values of operators evaluated on said state. The archetypal variational quantum algorithm is the variational quantum eigensolver [55] where the function

$$L(|\psi\rangle) = \sum_i \langle \psi | O_i | \psi \rangle = \sum_i \langle O_i \rangle \quad (26)$$

is the expectation value of a Hamiltonian  $H = \sum_i O_i$ . In this case, the ground state of  $H$  yields the solution to the optimization problem, and so the optimized parametrized quantum circuit should ideally prepare a state close to the ground state of  $H$ . However, such circuits are noisy and the goal of error mitigation is to correct this. It stands to reason then that an error mitigation protocol should output estimates of the expectation values  $\langle O_i \rangle$  on the state output by the noiseless version of these circuits. We can formally capture this task in the following definition of *weak* error mitigation. In all of the definitions below, let  $|\psi\rangle = \mathcal{C}(\rho)$  be the state output by the noiseless circuit.

**Definition 2** (Weak error mitigation (expectation value error mitigation)). *An  $(\epsilon, \delta)$  weak error mitigation algorithm  $\mathcal{A}$  with resources as in Definition 1 takes as input a classical description of a set of observables  $\mathcal{M} = \{O_1, \dots, O_\ell\}$  satisfying  $\|O_i\| \leq 1$  and outputs estimates  $\hat{o}_i$  of  $o_i = \langle \psi | O_i | \psi \rangle$  such that*

$$\mathbb{P}[\|\hat{o}_i - o_i\| \leq \epsilon \text{ for all } 1 \leq i \leq \ell] \geq 1 - \delta. \quad (27)$$

Here the probability in Eq. (27) is over the randomness of the error mitigation algorithm. This randomness could come from using classical random bits or from making measurements of the quantum states available as input.

The task of computing expectation values is ubiquitous in near-term quantum computing. Most error mitigation algorithms in the literature address the weak error mitigation task, including all protocols listed in Section II B. However, in some applications, knowing expectation values is not sufficient. In these cases, we would like to represent the strongest possible access, on par with access to the quantum computer, which is sampling of the circuit’s output. This leads us to a definition of *strong* error mitigation:

**Definition 3** (Strong error mitigation (sample error mitigation)). *An  $(\epsilon, \delta)$  strong error mitigation algorithm  $\mathcal{A}$  with resources as in Definition 1 outputs a bitstring  $z$  sampled according to a distribution  $z \sim \mu$  such that with probability*

$1 - \delta$ ,

$$d_{\text{TV}}(\mu, D_{|\psi\rangle}) \leq \epsilon, \quad (\text{additive error } \epsilon) \quad (28)$$

or alternatively

$$\frac{D_{|\psi\rangle}(z)}{\mu(z)} \leq \kappa \text{ for all } z \in \{0, 1\}^n, \quad (\text{multiplicative error } \kappa) \quad (29)$$

where  $D_{|\psi\rangle}$  is the distribution arising from a computational basis measurement of  $|\psi\rangle$ .

As the strong error mitigation task is more difficult than the weak error mitigation task—as we will show below, strong error mitigation implies weak error mitigation—and weak error mitigation is usually sufficient for near-term applications, there are fewer methods available that achieve this, an example being virtual distillation [20, 21]. The two notions of error are related— $\kappa$  multiplicative error implies

$$\epsilon = \sqrt{1 - \frac{1}{\kappa}} \quad (30)$$

additive error. To see this, note that the multiplicative error requirement can be re-written as

$$\frac{D_{|\psi\rangle}(z)}{\mu(z)} \leq \kappa \rightarrow D_{\infty}(D_{|\psi\rangle} \parallel \mu) \leq \log(\kappa) \quad . \quad (31)$$

By the monotonicity of relative entropies, we then have

$$D_{KL}(D_{|\psi\rangle} \parallel \mu) = D_1(D_{|\psi\rangle} \parallel \mu) \leq D_{\infty}(D_{|\psi\rangle} \parallel \mu) \leq \log(\kappa) \quad . \quad (32)$$

By the *Bretagnolle-Huber inequality* [56], we can then relate this to total variation distance as

$$d_{\text{TV}}(D_{|\psi\rangle}, \mu) \leq \sqrt{1 - \exp(-D_{KL}(D_{|\psi\rangle} \parallel \mu))} \leq \sqrt{1 - \frac{1}{\kappa}}. \quad (33)$$

On the other hand, it is easy to check that additive error does not imply multiplicative error for any setting of the parameters.

The additive-error requirement for strong error mitigation makes intuitive sense when sampling access to the noiseless quantum state is required and it is not important that the samples generated by the error-mitigated algorithm should come from any particular subset of the support. However, this is not the case for some of the near-term quantum algorithms that one might want to error mitigate. Consider for example variational quantum optimization algorithms like the quantum approximate optimization problem [34], where a diagonal Hamiltonian  $\mathcal{H}$  encodes a hard combinatorial optimization problem. Here, computational basis states of low energy correspond to approximate solutions of the combinatorial optimization problem. If the noiseless state  $|\psi\rangle$  has an overlap of  $O(1/\text{poly}(n))$  with the low-energy subspace of  $\mathcal{H}$ , polynomially many samples from the noiseless distribution  $D_{|\psi\rangle}$  are sufficient to solve the optimization problem with high probability. And then the same would hold for sampling from an error mitigated distribution with multiplicative error  $\kappa = O(1/\text{poly}(n))$ : such distribution must also have an inverse polynomially small weight on the set of low-energy strings. Such important examples motivate our definition of the multiplicative error mitigation.

#### D. Relationship between notions of error mitigation

Having established that a comprehensive study of error mitigation must consider both weak and strong versions, we now ask: how are they related? We are particularly interested in understanding if the output of one kind of error mitigation can be used, in polynomial time, to compute the output of another kind of error mitigation. A first observation is that **strong error mitigation implies weak error mitigation with local observables**. Let  $\mathcal{A}$  be a strong error mitigation algorithm. Definition 3 requires  $\mathcal{A}$  to output a sample from the computational basis of the noiseless circuit's output state. However, if  $O_i$  is a local observable (say a product Pauli observable) and we assume that the cost of strong error mitigation on  $\mathcal{C}$  does not increase significantly by appending a layer of product unitaries to  $\mathcal{C}$ , then  $\mathcal{A}$  can also output samples from the probability distribution associated with measuring  $\mathcal{C}$  in an eigenbasis of  $O_i$ . After applying the strong error mitigation procedure to obtain enough clean samples from the eigenbasis of



$O_i$ , we can estimate the expectation value  $\text{Tr}(O_i\mathcal{C}(\rho))$  empirically to a desired precision, thereby achieving weak error mitigation. One of the main questions asked in this work is whether we can hope for the other direction to hold. That is, whether weak error mitigation protocols can be used to also obtain samples from the noiseless circuit. The remainder of this subsection answers this question.

1. *Weak error mitigation implies strong error mitigation only with exponentially-many observables*

Let us now consider the problem of using the outputs of a weak error mitigation algorithm to obtain a strong error mitigation algorithm. That is, what is the minimum number of expectation values required to produce samples from the noiseless circuit? Here we allow arbitrarily complex post-processing steps because our focus is on sample complexity, or fundamental information-theoretic limits. Restricting our focus to bounded-time (i.e., realistic) computations would only further limit the set of protocols under consideration.

First note that in the limit of *exponentially* many error-mitigated expectation values, obtaining a (potentially inefficient) strong error mitigation algorithm is possible. Indeed, suppose we were able to perform weak error mitigation, outputting estimates of all  $n$ -qubit Paulis

$$\text{Tr}(\mathcal{C}(\rho)P_i) \quad \forall P_i \in \mathcal{Q}_n \quad . \quad (34)$$

up to exponentially small precision. This allows us to perform full tomography on the noiseless output  $\mathcal{C}(\rho)$  [57]—if we were allowed to query exponentially-many such expectation values (since there are exponentially many members of  $\mathcal{Q}_n$ ). However, this procedure would clearly be inefficient and likely more costly than just simulating the circuit classically. Thus, the more interesting question is: **could there be an algorithm that only needs to see *polynomially-many* expectation values from the output of weak error-mitigation to obtain strong error mitigation?**

We proceed to provide a partial negative answer to this question, showing that there cannot exist such an algorithm if the expectation values are all of observables *diagonal in the same eigenbasis*. To do so, we showcase an instance of weak error mitigation that provides an explicit counterexample to the conjecture. First, we will need to introduce the notion of a *statistical query* [58]:

**Definition 4** (Statistical query). *A statistical query is a pair  $(q, \tau)$  with*

- a function  $q : \{0, 1\}^n \times \{0, 1\} \rightarrow \{0, 1\}$ .
- $\tau$ : a tolerance parameter  $\tau \geq 0$ .

Now we are ready to define a statistical query oracle.

**Definition 5** (Statistical query oracle). *Fix an unknown function  $c : \{0, 1\}^n \rightarrow \{0, 1\}$  and a probability distribution  $D$  over the domain  $\{0, 1\}^n$ . The statistical query oracle,  $SQ(q, \tau)$ , when given a statistical query, returns any value in the range as*

$$[\mathbb{E}_{x \sim D}[q(x, c(x))] - \tau, \mathbb{E}_{x \sim D}[q(x, c(x))] + \tau]. \quad (35)$$

We will also need to use the observation that the problem of PARITIES on  $n$  bits (which we now define) can be solved with  $\text{poly}(n)$  samples but requires  $\exp(n)$  statistical queries.

**Definition 6** (PARITIES). *The class of PARITIES is the set of functions  $\{c_s : \{0, 1\}^n \rightarrow \{0, 1\}\}_{s \in \{0, 1\}^n}$  where*

$$c_s(x) = x \cdot s \quad \text{for } s \in \{0, 1\}^n \quad . \quad (36)$$

We may define the associated PARITIES distribution  $P_s : \{0, 1\}^{n+1} \rightarrow [0, 1]$  via

$$P_s(x \bowtie y) = \begin{cases} 2^{-n}, & \text{if } y = x \cdot s \\ 0, & \text{else,} \end{cases} \quad (37)$$

where  $x \in \{0, 1\}^n$ ,  $y \in \{0, 1\}^{n+1}$ , and the symbol  $\bowtie$  means concatenation. That is, the PARITIES distribution on  $n + 1$  bits corresponding to a secret  $n$ -bit-string  $s$ , is the distribution which is supported uniformly on those bit-strings whose last bit is the parity of the subset of the first  $n$  bits that are indexed by the string  $s$ .

There are three important facts about PARITIES distributions that we will use.

1. For every  $s \in \{0, 1\}^n$ , there exists an  $n+1$ -qubit quantum circuit  $\mathcal{C}_s$  that ‘encodes’ the PARITIES distribution  $P_s$ , in the following sense: when initialized on the all-0s state, the output state has support only on the computational basis states whose labels are in the support of the distribution  $P_s$ :

$$\mathcal{C}_s |0\rangle^{\otimes n+1} = \frac{1}{\sqrt{2^n}} \sum_{x \in \{0,1\}^n} |x \bowtie c_s(x)\rangle \quad , \quad (38)$$

for the parity function  $c_s(x) = x \cdot s$ . This is proven in, for example, Ref. [59].

2. PARITIES are not solvable with sub-exponentially-many statistical queries [38] (see Refs. [37, 60] for tutorials):

**Lemma 3** (SQ-hardness of PARITIES [38]). *Any algorithm for learning the class of  $n$ -bit PARITIES to constant accuracy and probability of success and which makes statistical queries of the form  $(\chi, \tau)$ , where  $\tau \geq \tau_0$  for each query, must make  $\Omega(\tau_0^2 \cdot 2^n)$  queries.*

3. We can lower-bound the total variation distance between any two PARITIES distributions as follows:

**Lemma 4.** *For any two  $s, s' \in \{0, 1\}^k$ ,  $d_{TV}(P_s, P_{s'}) \geq 1/2$ .*

*Proof.* Le Cam’s two point method says that the failure probability of any binary hypothesis test  $\Psi : X \rightarrow \{P, Q\}$  on two distributions  $P, Q : X \rightarrow [0, 1]$  each chosen with probability  $1/2$ , is lower bounded as follows:

$$\min_{\Psi} P_{\text{error}}(\Psi) \geq \frac{1 - d_{TV}(P, Q)}{2} \quad . \quad (39)$$

For any two strings  $s, s' \in \{0, 1\}^k$ , let  $P = P_s$  and  $Q = P_{s'}$ . We wish to bound  $d_{TV}(P_s, P_{s'})$ . To do so we will exhibit a hypothesis test on  $P_s$  versus  $P_{s'}$  that has error probability  $1/4$ . Plugging this into Eq. (39) yields the desired lower bound on  $d_{TV}(P_s, P_{s'})$ . A simple test is then as follows: given a string  $x \bowtie c$  drawn from either  $P_s$  or  $P_{s'}$ , where  $c$  is one bit, we compute the parity of  $x$  with  $s$  and  $s'$ . In Case 1, we have  $x \cdot s = x \cdot s'$ . Then we output  $s$  or  $s'$  uniformly at random. This succeeds with probability  $1/2$ . Otherwise (call this Case 2)  $x \cdot s \neq x \cdot s'$ . Then we output whichever string  $s$  or  $s'$  yields the right parity  $c$ . This succeeds with probability  $1$  in distinguishing the two distributions. So the probability of error of the test we have described is simply

$$P_{\text{error}} = P(\text{Case 1})1/2 \quad . \quad (40)$$

Now let us compute  $P(\text{Case 1}) = 1 - P(\text{Case 2})$ . As the  $x$  part of the string is chosen uniformly from  $\{0, 1\}^n$ , computing  $P(\text{Case 2})$  boils down to computing the number of  $x$ ’s that either satisfy the equations  $x \cdot s = 0$  and  $x \cdot s' = 1$  or  $x \cdot s = 1$  and  $x \cdot s' = 0$ . In both cases,  $x$  must satisfy two linearly independent relations. Thus, we conclude that the strings that satisfy  $x \cdot s = 0$  and  $x \cdot s' = 1$  form a vector space of dimension  $n - 2$ , which contains  $2^{n-2}$  strings. Thus, the number of strings that satisfy either  $x \cdot s = 0$  and  $x \cdot s' = 1$  or  $x \cdot s = 1$  and  $x \cdot s' = 0$  is  $2^{n-1}$  which is half of all the strings  $x$ . Thus,  $P(\text{Case 2}) = P(\text{Case 1}) = 1/2$ , and so the overall  $P_{\text{error}} = 1/2 * 1/2 = 1/4$ .  $\square$

In addition, we will need the notion of hypothesis selection and an algorithm that achieves a good approximation guarantee for it, due to Yatracos [39]:

**Theorem 6** (3-proper hypothesis selection). *Fix a class of distributions  $\mathcal{Q} = \{q_1, \dots, q_k\}$  and  $\epsilon, \delta > 0$ . Given  $O(\log |\mathcal{Q}|/\epsilon^2)$  samples from a target distribution  $p$  (which may not be in  $\mathcal{Q}$ ), there is an algorithm to output a distribution  $q^* \in \mathcal{Q}$  satisfying*

$$d_{TV}(p, q^*) \leq 3 \min_{i \in [k]} d_{TV}(p, q_i) + \epsilon \quad (41)$$

with probability at least  $1 - \delta$ .

Here, the word *proper* refers to the fact that the output distribution is required to be in  $\mathcal{Q}$ , a feature that we will require for the argument we are about to make. The connection between expectation values (the output of weak error mitigation) and statistical queries is encapsulated in the following observations (where for  $b \in \{0, 1\}^n$ ,  $Z^b := Z^{b_1} \otimes Z^{b_2} \otimes \dots \otimes Z^{b_n}$ ):

**Lemma 5** (Weak error mitigation on the circuits  $\mathcal{C}_s$  outputs statistical queries to the  $\text{PARITIES}_s$  distribution). *Consider error mitigation on the  $n+1$ -qubit  $\text{PARITIES}$  circuit  $\mathcal{C}_s$  (see Def. 6). That is, let  $\mathcal{M} = \{(\mathbb{I} + Z^b)/2\}_{b \in \{0,1\}^{n+1}}$ , and suppose we have a weak error mitigation algorithm  $\mathcal{A}(\mathcal{C}_s, \mathcal{N}, \mathcal{M})$  that, with probability  $1 - \delta$ , outputs  $\tau$ -accurate estimates*

$$|\hat{o}_i - \text{Tr}[\mathcal{C}_s(|0\rangle\langle 0|^{\otimes n+1})O_b]| < \tau \quad \text{for all } O_b \in \mathcal{M} \quad . \quad (42)$$

The outputs  $\hat{o}_i$  are valid responses to statistical queries with tolerance  $\tau$  for the distribution  $P_s$ .

The observation that weak error mitigation outputs statistical queries was also made in Ref. [61] which introduced the notion of a *quantum* statistical query (QSQs) in the quantum PAC learning setting. We will come back to this after we state our main theorem.

*Proof.* We will prove that each  $\tau$ -accurate expectation value output by  $\mathcal{A}$  that satisfies Eq. (42) takes the form of a response of a statistical query oracle to a statistical query of tolerance  $\tau$ , by specifying what  $c, q, D$  from Definition 5 correspond to. Let us denote the circuit's clean output state vector as  $\mathcal{C}_s |0\rangle^{\otimes n+1} = |\psi\rangle = 2^{-n/2} \sum_{x \in \{0,1\}^n} |x \otimes c_s(x)\rangle$ . Then it is easy to see, from inspecting the basis vectors in the support of  $|\psi\rangle$ , that the desired expectation values can be rewritten as

$$\text{Tr}(|\psi\rangle\langle\psi|(\mathbb{I} + Z^b)/2) = \mathbb{E}_{z \sim P_s}[q_b(z)] = \mathbb{E}_{x \sim \text{Unif}\{0,1\}^n}[q_b(x, c_s(x))] \quad , \quad (43)$$

where  $q_b : \{0,1\}^{n+1} \rightarrow \{0,1\}$  is defined as

$$q_b(z) = \frac{1 + \langle z | Z^b | z \rangle}{2} \quad . \quad (44)$$

That is,  $q = q_b, c = c_s$  and  $D = \text{Unif}\{0,1\}^n$ . □

The re-scaling of the observables  $Z^b$  is cosmetic; it is merely to obtain a set of observables whose eigenvalues take values in  $\{0,1\}$  due to the way we have defined statistical queries in Def 5. In a similar fashion, while this lemma is written for the observables  $Z^b$ , one could write an analogous lemma for observables that are all diagonal in some other common eigenbasis, by simply applying the corresponding basis change at the end of the circuit  $\mathcal{C}_s$  and re-scaling the value of the tolerance parameter  $\tau$  for the statistical query.

This observation will be crucial to the main theorem of this section (Theorem 7), which states that with sub-exponentially-many expectation values of observables which share the same eigenbasis, weak error mitigation cannot imply strong error mitigation in the worst-case. We now state this more formally:

**Theorem 7** (Exponentially-many expectation values of observables in the same eigenbasis are required to output samples from that basis). *There is a class of  $n$ -qubit circuits, such that for every circuit  $\mathcal{C}$  in the class (denote the clean circuit output state as  $|\psi\rangle = \mathcal{C}(|0\rangle)$ ) and the set of observables  $\mathcal{M} = \{Z^b\}_{b \in \{0,1\}^{n+1}}$  with  $m = o(\tau^2 \cdot 2^n)$ , the following holds:*

*No algorithm  $\mathcal{B}$  exists that takes as input the output of weak error mitigation  $\mathcal{A}(\mathcal{C}, \mathcal{N}, \mathcal{M})$ —the estimates  $\{\hat{z}_b\}_{b \in \{0,1\}^m}$  with  $|\hat{z}_b - \langle \psi | Z_b | \psi \rangle| < \tau$  for all  $b$ —and outputs  $O(n)$ -many samples from some distribution  $p$  where*

$$d_{TV}(p, D_{|\psi\rangle}) \leq 1/16 \quad , \quad (45)$$

where  $D_{|\psi\rangle}$  is the computational basis distribution on  $|\psi\rangle$ .

After we prove this, we will remark that not even the power to choose the observables *adaptively* will make it possible to transform a sub-exponential number of expectation values into linearly-many samples.

*Proof of Theorem 7.* Ref. [59] has shown that there is a class of  $n$ -qubit Clifford circuits whose output distributions are exactly the set of  $\text{PARITIES}$  distributions. Suppose to the contrary that for some circuit  $\mathcal{C}$  in this class whose output distribution is some  $P_s \in \text{PARITIES}$ , there exists some set of observables  $\mathcal{M}$  with  $|\mathcal{M}| = o(\tau^2 \cdot 2^n)$  and some algorithm  $\mathcal{B}$ , which takes as input the estimates  $\hat{z}_i$  of weak error mitigation up to error  $\tau$  and outputs  $O(n)$  samples from some distribution  $p$  such that  $d_{TV}(p, P_s) \leq 1/16$ .

We will show that we can use these samples to solve for the hidden string and thus solve the problem of  $\text{PARITIES}$ . However, this would imply a contradiction: Recall from Lemma 5 that each expectation value estimate is a statistical query to some distribution from the class of  $\text{PARITIES}$ , but as stated in Lemma 3, there is no *statistical query* algorithm to solve  $\text{PARITIES}$  with  $o(\tau^2 \cdot 2^n)$  statistical queries.

In the rest of this proof, we will explain how to use the samples output by the presumed  $\mathcal{B}$  to solve PARITIES. The key is simply to run the hypothesis selection algorithm of Ref. [39] (Theorem 6), with the set of candidate hypotheses being the set of all Clifford distributions encoding PARITIES. This algorithm will take as input the  $O(n)$  samples from  $p$  and by the guarantees of Theorem 6, will output a distribution  $P_{s'} \in \text{PARITIES}$  such that

$$d_{TV}(p, P_{s'}) \leq 3 \min_{\bar{s} \in \{0,1\}^n} d_{TV}(p, P_{\bar{s}}) + 1/16 \quad . \quad (46)$$

Now, consider that by the guarantees of  $\mathcal{B}$ ,  $3d_{TV}(p, P_s) + 1/16 = 1/4$  and so Eq. (46) yields that  $d_{TV}(p, P_{s'}) < 1/4$ . But as stated in Lemma 4, for all  $s, s'$ ,

$$d_{TV}(P_s, P_{s'}) \geq 1/2 \quad , \quad (47)$$

and so by the triangle inequality,

$$\arg \min_{\bar{s} \in \{0,1\}^n} d_{TV}(p, P_{\bar{s}}) = s \quad (48)$$

i.e., hypothesis selection recovers the hidden string of the PARITIES problem successfully. This implies the contradiction.  $\square$

**Extensions of our proof.** We remark that our bound holds against even an *adaptive* choice of observables in error mitigation because the statistical query hardness of PARITIES holds against an adaptive choice of statistical queries. However, we reiterate a limitation of our result: we can only rule out obtaining strong error mitigation from weak error mitigation with sub-exponentially-many *observables whose eigenbasis is the basis we desire to sample from*. It is natural to ask if we can lift this restriction. Here, we point out that the discussion in Ref. [62] shows that for ‘flat’ circuit output distributions, and a specially-crafted set of observables with multiple different eigenbases, only polynomially-many expectation values are needed to obtain a sampler (hence we cannot lift the restriction in general). We could also take a different tack: our proof is based on the hardness of PARITIES from *classical* statistical queries. As observed by Ref. [61], it is also possible to define *quantum* statistical queries for a given unknown distribution  $D$ , and unknown function  $c$ , and QSQs generalize classical statistical queries by allowing for observables not diagonal in the eigenbasis defined by  $c$ . Hence, if we knew of a problem that was hard for certain classes of *quantum* statistical queries (in the sense that exponentially-many QSQs are needed to solve them), our proof technique could then be applied to say something about the hardness of obtaining strong error mitigation from weak error mitigation without the restriction mentioned at the beginning of this paragraph. Unfortunately we do not know of any such problems.

### III. CONCEPTUAL BACKGROUND AND CONTRIBUTIONS

We first show a basic version of our theorem bounding the sample complexity of error mitigation in a circuit subject to depolarizing noise. This theorem already has all the features of our main results but is weaker, in the sense that we will only be able to show an exponential-in- $D$  dependence of the sample complexity, while our main result improves this to an exponential in both  $n, D$ .

#### A. Preliminaries

##### 1. Pauli operators and their properties

We denote the single-qubit Pauli operators by  $\mathbb{I}, X, Y, Z$ . Furthermore, for  $a = (a_1, a_2, \dots, a_n) \in \{0, 1\}^n$ , we let

$$Z^a := Z^{a_1} \otimes Z^{a_2} \otimes \dots \otimes Z^{a_n}, \quad (49)$$

$$X^a := X^{a_1} \otimes X^{a_2} \otimes \dots \otimes X^{a_n}, \quad (50)$$

where

$$Z^{a_i} = \begin{cases} \mathbb{I} & a_i = 0, \\ Z & a_i = 1, \end{cases} \quad (51)$$

$$X^{a_i} = \begin{cases} \mathbb{I} & a_i = 0, \\ X & a_i = 1. \end{cases} \quad (52)$$

We then define the Pauli group.

**Definition 7** (Pauli group and Pauli weight). *Let  $n \in \{1, 2, \dots\}$ . The Pauli group  $\mathcal{P}_n$ , by definition, consists of all operators of the form  $i^k X^a Z^b$ , where  $k \in \{0, 1, 2, 3\}$  and  $a, b \in \{0, 1\}^n$ . Let  $\mathcal{Q}_n := \mathcal{P}_n / \{\pm 1, \pm i\}$  be the quotient group that results from disregarding global phases in  $\mathcal{P}_n$ . For every Pauli operator  $P \in \mathcal{Q}_n$ , we denote by  $w(P)$  the weight of  $P$ , which is the number of qubits on which  $P$  acts non-trivially.*

In the following Lemma, we mention some basic facts about Pauli operators that we use repeatedly.

**Lemma 6** (Properties of Pauli operators). *We list some properties of the Pauli group.*

1. **Trace of products of Paulis.**

For  $n$ -qubit Paulis  $P, Q \in \mathcal{Q}_n$ ,

$$\text{Tr}(P \cdot Q) = \begin{cases} 2^n & P = Q, \\ 0 & \text{otherwise.} \end{cases} \quad (53)$$

2. **Pauli basis expansions.**

The  $n$ -fold Paulis form an orthogonal basis for all Hermitian operators in  $\mathcal{H}^{\otimes n}$ ; in particular, for  $a \in \{0, 1\}^n$ ,

$$|a\rangle\langle a| = \frac{1}{2^n} \sum_{b \in \{0, 1\}^n} (-1)^{a \cdot b} Z^b, \quad (54)$$

where  $a \cdot b = a_1 b_1 + a_2 b_2 + \dots + a_n b_n \pmod{2}$ .

2. **Depolarizing noise**

**Definition 8** (Depolarizing channels). *For  $M \in L(\mathbb{C}^d)$ , the  $d$ -dimensional depolarizing channel  $\mathcal{D}_p^{(d)}$ , for  $d \geq 2$ , is defined as*

$$\mathcal{D}_p^{(d)}(M) := pM + (1-p) \text{Tr}[M] \frac{\mathbb{I}}{d}, \quad p \in [-1/(d^2 - 1), 1]. \quad (55)$$

When the superscript for the dimension is omitted, we implicitly refer to the single-qubit depolarizing channel (with  $d = 2$ ):

$$\mathcal{D}_p(M) := \mathcal{D}_p^{(2)}(M) = pM + (1-p) \text{Tr}[M] \frac{\mathbb{I}}{2}, \quad p \in [-1/3, 1]. \quad (56)$$

We note that the single-qubit depolarizing channel has an alternate representation in terms of the Pauli operators

$$\mathcal{D}_p(\rho) = q_X X \rho X + q_Y Y \rho Y + q_Z Z \rho Z + q_I \rho, \quad (57)$$

where

$$q_X = q_Y = q_Z = \frac{1-p}{4} \quad (58)$$

and

$$q_I = \frac{1+3p}{4}. \quad (59)$$

In the context of depolarizing noise acting within an  $n$ -qubit circuit, a *global depolarizing channel* is simply a  $2^n$ -dimensional depolarizing channel. Using (53), we find that the *global* depolarizing channel on  $n$  qubits can be written as

$$\mathcal{D}_p^{(2^n)}(P) = pP + (1-p)\delta_{P,\mathbb{I}}\mathbb{I} \quad \forall P \in \mathcal{P}_n. \quad (60)$$

Alternatively, one could also model the noise within a circuit as an  $n$ -fold local (single-qubit) depolarizing channel  $\mathcal{D}_p^{\otimes n}$ . In particular, the single-qubit depolarizing channel has the property that

$$\mathcal{D}_p(X) = pX, \quad \mathcal{D}_p(Y) = pY, \quad \mathcal{D}_p(Z) = pZ, \quad \text{and} \quad \mathcal{D}_p(\mathbb{I}) = \mathbb{I}. \quad (61)$$

Based on these calculations, we make the following observation.

**Lemma 7** (Action of depolarizing noise on a Pauli string). *For all  $p \in [-1/3, 1]$  and  $P \in \mathcal{P}_n$ ,*

$$\mathcal{D}_p^{\otimes n}(P) = p^{w(P)}P. \quad (62)$$

*Proof.* This follows immediately from (57) and the definition of  $\mathcal{D}_p^{\otimes n}$ .  $\square$

### 3. Unitary 2-designs and Clifford unitaries

A *unitary  $t$ -design*, for  $t \in \{1, 2, \dots\}$ , is a finite ensemble  $\{(1/K, U_k)\}_{k=1}^K$  of unitaries such that [63, 64]

$$\frac{1}{K} \sum_{k=1}^K U^{\otimes t} \otimes (U^\dagger)^{\otimes t} = \int_U U^{\otimes t} \otimes (U^\dagger)^{\otimes t} dU, \quad (63)$$

where the integral on the right-hand side is with respect to the Haar measure on the unitary group. The  $n$ -qubit Clifford group  $\mathcal{C}_n$  forms a unitary 2-design [63, Theorem 1], where by definition the Clifford group is the normalizer of the Pauli group  $\mathcal{P}_n$  [65–67].

## B. Appetizer: mitigating depolarizing noise requires exponential-in-circuit depth samples

We are finally ready to present our ‘appetizer’ theorem. Let us first establish the notation:

**Definition 9** (Circuit with depolarizing noise). *For any noiseless circuit  $\mathcal{C}$  of depth  $D$  acting on  $n$  qubits, define the noisy circuit with noise parameter  $p$ ,  $\Phi_{\mathcal{C},p}$ , as follows:*

$$\Phi_{\mathcal{C},p} := \mathcal{C}^{(D)} \circ \mathcal{D}_p^{\otimes n} \dots \circ \mathcal{C}^{(2)} \circ \mathcal{D}_p^{\otimes n} \circ \mathcal{C}^{(1)} \quad . \quad (64)$$

where  $\mathcal{C}^{(i)}$  represents the  $i^{\text{th}}$  layer of the circuit  $\mathcal{C}$ , and  $\mathcal{D}_p$  is a local, single-qubit depolarizing channel.

Furthermore, denote the output state of the noisy circuit with input state  $\rho$  by

$$\omega_{p,\mathcal{C},\rho} := \Phi_{\mathcal{C},p}(\rho). \quad (65)$$

We will now present the key reduction we will use to bound the sample complexity of error mitigation. This reduction is visually represented in Figure 2.

To do so, let us step away from the error mitigation setting for a moment, and first consider the following problem of *state discrimination in the presence of noise*:

**Problem 3** (State discrimination in the presence of noise). *Fix a set  $\{\rho_0, \rho_1, \dots, \rho_N\}$  of  $N+1$  quantum states and a circuit  $\mathcal{C}$  acted upon by noise  $\mathcal{N}$ . Suppose that a distinguisher has knowledge of  $\mathcal{C}$  and  $\mathcal{N}$ , and is given access to copies of the state  $\Phi_{\mathcal{C},\mathcal{N}}(\rho_i)$ , defined analogously to (65), with the index  $i \in \{0, 1, \dots, N\}$  unknown. What is the fewest number of copies of  $\Phi_{\mathcal{C},\mathcal{N}}(\rho_i)$  needed in order to identify  $i$  with high probability, in terms of the noise strength?*

The following variation of Fano’s lemma can be used to address this question:

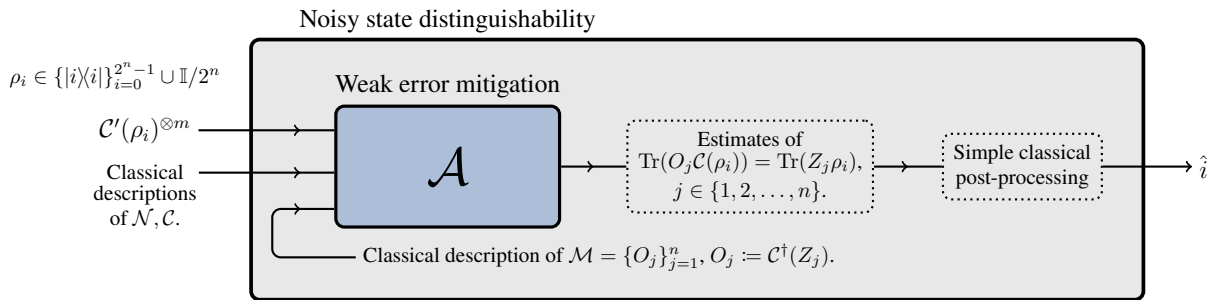


FIG. 2. To lower-bound the sample complexity of weak error mitigation, we show that it can be used as a subroutine to solve a constructed problem of distinguishing states under noise.

**Lemma 8** (Corollary 2.6 of Ref. [24]). *Let  $P_0, \dots, P_N$  be probability measures on some state space  $X$  such that*

$$\frac{1}{N+1} \sum_{k=0}^N D(P_k \| P_0) \leq \alpha \log(N) \quad (66)$$

for  $0 < \alpha < 1$ . Then the minimum average probability of error over tests  $\psi : X \rightarrow \{0, 1, \dots, N\}$  that distinguish the probability distributions  $P_0, \dots, P_N$  which we define as

$$\bar{p}_{e,N} := \inf_{\psi} \frac{1}{N+1} \sum_{j=0}^N P_j(\psi \neq j) \quad (67)$$

satisfies:

$$\bar{p}_{e,N} \geq \frac{\log(N+1) - \log(2)}{\log(N)} - \alpha. \quad (68)$$

Let us explain what this Lemma implies about the sample complexity of state discrimination in the presence of noise (Problem 3). Intuitively, dialing up the noise on the circuit  $\mathcal{C}$  makes the various output states  $\Phi_{\mathcal{C},\mathcal{N}}(\rho_i)$  less distinguishable from each other—and so, more copies of  $\Phi_{\mathcal{C},\mathcal{N}}(\rho_i)$  must be taken to determine the true identity of the input state. This Lemma formalizes this intuition. For let  $P_0, \dots, P_N$  be distributions over the output of the  $m$ -copy distinguisher in Problem 3 when the true state is  $\rho_0 \dots \rho_N$ , respectively. One can then show that, for  $\mathcal{N}$  being depolarizing noise of parameter  $p$  on generic depth- $D$  circuits, the average distance of  $P_1, \dots, P_N$  to  $P_0$  (the left-hand-side of (66)) scales as  $\alpha \log(N) = p^{2D} \log(N)$ . To get a constant probability of error in (68),  $\alpha$  must be constant. In order for  $\alpha$  to be constant, one needs to choose  $m$  scaling as  $p^{-2D}$ .

We can go one step further and use Lemma 8 to bound the performance of error mitigation. The proof proceeds according to the following two basic steps:

1. Embed an error mitigation protocol into a solver for a state discrimination problem. This means, we identify a state discrimination problem such that an error mitigation algorithm can be transformed, with no additional copies required, into a distinguisher that solves the state discrimination problem.
2. Compute Fano's lower bound for the number of copies needed for the state discrimination problem.

More formally, we now state and prove the following theorem:

**Theorem 8** (Appetizer: Error mitigation under depolarizing noise requires exponential-in-depth samples). *Let  $\mathcal{A}$  be an input state-agnostic weak error mitigation algorithm (see Definition 2) that takes as input  $m$  noisy copies  $\Phi_{\mathcal{C},p}(\rho)$ . Suppose that for any circuit  $\mathcal{C}$  acted upon by depolarizing noise of parameter  $p$ , and any set of observables  $\mathcal{M} = \{O_i\}_{i \in [n]}$ ,  $\mathcal{A}$  is able to produce estimates  $\{\hat{o}_i\}_{i \in [n]}$  such that with probability at least  $1 - \delta$ ,*

$$|\hat{o}_i - \text{Tr}[O_i \mathcal{C}(\rho)]| \leq \epsilon \quad \text{for all } i \in [n]. \quad (69)$$

Then there exists an observable set  $\mathcal{M}$  and a set of input states  $\rho$  such that, as long as  $\epsilon < 1/2$ , the number of noisy copies,  $m$ , needed by  $\mathcal{A}$ , is lower-bounded as

$$m = \Omega(p^{-2D}) \quad (70)$$

for any circuit  $\mathcal{C}$  on  $n$  qubits of depth  $D$ .

After we prove this theorem, we remark on extensions of our proof to EM protocols with other types of input state access, including those protocols that are input state-aware.

*Proof of Theorem 8.* Let  $\mathcal{A}$  be an input state-agnostic weak error mitigation algorithm and consider the observable set  $\mathcal{M} = \{\mathcal{C}^\dagger(Z_j)\}_{j=1}^n$  where  $Z_j$  is the Pauli  $Z$  acting on qubit  $j$ . For ease of notation, we will denote by  $\hat{z}_j$  the estimate of  $\text{Tr}[\mathcal{C}^\dagger(Z_j)\mathcal{C}(\rho_i)]$ . We are going to identify a set  $S$  of states (which is also the set of states referenced in the Theorem), such that to solve Problem 3 when the unknown state is chosen from  $S$ , it suffices to run  $\mathcal{A}$  on copies of the unknown state and then do simple post-processing.

*a. Step 1: A noisy state identification problem that can be solved by weak error mitigation.* Consider choosing the unknown state  $\rho_i$ ,  $i \in [2^n + 1]$  in Problem 3 from the following  $N + 1 = 2^n + 1$  options:

- $\rho_N = \mathbb{I}/2^n$ , the maximally-mixed state. If this option is chosen, say we are in **Case 1**.
- $\{\rho_x := |x\rangle\langle x|\}_{x \in \{0,1\}^n}$ . If one of these options is chosen, say we are in **Case 2**.

(We emphasize that the state discrimination algorithm must output  $i$ , the label of the state, rather than simply decide whether it is in Case 1 or 2.) We will now show that Problem 3 can be solved with access to an error mitigation algorithm  $\mathcal{A}$ . First suppose hypothetically that the error mitigation algorithm were *perfect*; i.e., when given  $m$  copies of  $\Phi_{\mathcal{C},\mathcal{N}}(\rho_i)$ ,  $\mathcal{A}$  outputs estimates  $\hat{z}_j$  that are *exact*:

$$\hat{z}_j = \text{Tr}[\mathcal{C}^\dagger(Z_j)\mathcal{C}(\rho_i)] = \text{Tr}[Z_j\rho_i] \quad \forall j \in [n] \quad . \quad (71)$$

(Later we will show that the analysis below goes through if we assume only approximate estimates.) But consider that, in Case 1, where  $\rho_i$  is  $\rho_N = \mathbb{I}/2^n$ , the true expected value of the observables is

$$\text{Tr}[Z_j\mathbb{I}/2^n] = 0 \quad \forall j \in [n] \quad (72)$$

i.e., a perfect error mitigation algorithm should estimate  $\hat{z}_j = 0$  for every observable.

On the other hand, in Case 2, when  $\rho_i = \rho_x$  for some  $x = (x_1, x_2, \dots, x_n) \in \{0,1\}^n$ , the true expected value of the observables is

$$\text{Tr}[|x\rangle\langle x|Z_j] = 2x_j - 1 \quad \forall j \in [n] \quad , \quad (73)$$

i.e., a perfect error mitigation algorithm would output the estimate  $\hat{z}_j = 1$  if  $x_j = 1$  and  $\hat{z}_j = -1$  if  $x_j = 0$ . Hence, if the error mitigation algorithm were perfect, one could perfectly distinguish Case 1 and Case 2; furthermore in the latter case, concatenating the set  $\{(1 + \hat{z}_j)/2\}_{j=1}^n$  results in an  $n$ -bit string corresponding to  $x = (x_1, x_2, \dots, x_n)$ , the label of the actual input state. A perfect error mitigation algorithm thus solves the state identification problem perfectly.

Now we argue that state identification can still be solved by an error mitigation algorithm that returns only *approximate* estimates. By assumption, with probability  $1 - \delta$ , the estimates are accurate up to additive error  $\epsilon$ . As long as  $\epsilon < 1/2$ , we find

$$|\hat{z}_j - \text{Tr}[\rho_i Z_j]| \leq 1/2 \quad \forall j \in [n] \quad , \quad (74)$$

(c.f. Eq. (71)). The *true* values, of course, remain the same: Eq. (72) and Eq. (73) still hold in Case 1 and Case 2 respectively. In particular, let  $\hat{z} = (\hat{z}_1, \hat{z}_2, \dots, \hat{z}_n)$  be the estimates produced by  $\mathcal{A}$  when the input state is  $\rho_N = \mathbb{I}/2^n$  (Case 1) and let  $\hat{z}_x = (\hat{z}_{1,x}, \hat{z}_{2,x}, \dots, \hat{z}_{n,x})$  be the estimates produced by  $\mathcal{A}$  when the input state is  $\rho_x = |x\rangle\langle x|$  (Case 2). Then Eq. (74) says that with probability  $1 - \delta$ ,

$$-\frac{1}{2} < \hat{z}_i < \frac{1}{2} \quad \forall i \quad (\text{Case 1}) \quad \text{and} \quad \begin{cases} \frac{1}{2} < \hat{z}_{j,x} < \frac{3}{2} & \text{if } x_j = 0, \\ -\frac{3}{2} < \hat{z}_{j,x} < -\frac{1}{2} & \text{if } x_j = 1. \end{cases} \quad (\text{Case 2}) \quad (75)$$

Now, let  $\text{EstInput}$  be the algorithm that takes the output of  $\mathcal{A}$ , which is the set of estimates  $\{\hat{z}_j\}_{j \in [n]}$ , and does the following:

1. If there exists  $j \in [n]$  such that  $\hat{z}_j \in (-1/2, 1/2)$ , it outputs that  $\hat{x} = N$ , i.e., the unknown state is  $\rho_N$ .
2. Else, it computes the  $n$ -bit string  $\hat{x} = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n) \in \{0,1\}^n$  as

$$\hat{x}_j = \begin{cases} 0 & \text{if } \hat{z}_j \geq 0 \\ 1 & \text{if } \hat{z}_j < 0 \end{cases} \quad , \quad (76)$$



and outputs that the unknown state is  $\rho_{\hat{x}}$ .

It is easy to see from Eq. (75) that this procedure correctly identifies the unknown state, and hence solves Problem 3 with success probability  $1 - \delta$ .

*b. Step 2: Computing Fano's lower bound for state identification* The string  $\hat{x}$  output by the above procedure is probabilistic. In particular, let  $P_i(\hat{x})$  for  $i \in [2^n + 1]$  be the probability that running  $\text{EstInput} \circ \mathcal{A}$  on the unknown state  $\rho_i$  will output  $\hat{x}$ . Since the last step of our procedure is to sample from  $P_i$  and then report the output, this last step is a single-sample test to distinguish probability distributions. By Fano's Lemma (Lemma 8), any such test must fail with probability no less than

$$\frac{\log(N+1) - \log(2)}{\log(N)} - \alpha \xrightarrow{N \text{ large}} 1 - \alpha \quad (77)$$

where  $\alpha \log(N)$  is defined to be an *upper bound* to the quantity

$$\frac{1}{N+1} \sum_{k=0}^N D(P_k \| P_N) \quad . \quad (78)$$

We will show later (in Proposition 1) that we can bound the average relative entropy between  $P_k$  and  $P_N$  as follows:

$$\frac{1}{N+1} \sum_{k=0}^N D(P_k \| P_N) \leq p^{2D} m \log(N), \quad (79)$$

so it suffices to set  $\alpha = p^{2D} m$ . Therefore, in order for the test to have a constant failure probability  $\delta$ , it must take at least  $m = p^{-2D} (1 - \delta)$ -many samples.  $\square$

Eq. (78) is a critical quantity in Fano's Lemma and our argument, so let us interpret it: take the maximally-mixed state  $\mathbb{I}/2^n$  to be the reference state and the probability distribution output by our noisy state discrimination algorithm on  $\mathbb{I}/2^n$  to be the reference probability distribution  $P_N$ . Eq. (78) examines the divergence between the output of the algorithm initialized with one of the states  $\rho_x$ , and that when initialized on the reference state. Intuitively, if this quantity is high, it indicates that the states are "far apart", and hence the distinguishing algorithm could potentially have a lower probability of error.

Indeed, bounding the distinguishability of the states in our set (which becomes the left-hand-side of Eq. (78) after one application of data-processing) is the technical crux of our argument. We present a basic bound in Proposition 1 and we will gradually refine this bound in the next few sections to get progressively more severe dependencies of the sample complexity of error mitigation on various parameters of the problem. In this Proposition, we will use notation that is more explicit about the inputs to error mitigation:  $\mathcal{A}(\mathcal{C}, p, \mathcal{M}, \rho_x)$  will mean the error mitigation algorithm provided with  $m$  copies of the noisy input state  $\Phi_{\mathcal{C}, p}(\rho_x)$  and where the task is to predict the observables in the set  $\mathcal{M}$ .

**Proposition 1** (Simple bound on average divergence). *Let  $P_0, P_1, \dots, P_{N-1}$  for  $N = 2^n$  be the probability distributions over the outputs of the algorithm for noisy state discrimination described in the proof of Theorem 8, run on the different input states  $\rho_0, \dots, \rho_{N-1}$ . That is,  $P_x$  for  $x \in \{0, 1\}^n$  is the probability distribution output by*

$$\text{EstInput} \circ \mathcal{A}(\mathcal{C}, p, \mathcal{M}, \rho_x) \quad , \quad (80)$$

and similarly let  $P_N$  be the probability distribution over the output of

$$\text{EstInput} \circ \mathcal{A}(\mathcal{C}, p, \mathcal{M}, \mathbb{I}/2^n) \quad . \quad (81)$$

Then it holds that

$$\frac{1}{N+1} \sum_{k=0}^N D(P_k \| P_N) \leq p^{2D} m \log(N), \quad (82)$$

where  $m$  is the number of times  $\mathcal{A}(\mathcal{C}, p, \mathcal{M}, \rho_x)$  has to invoke the noisy circuit  $\Phi_{\mathcal{C}, p}$ .

*Proof.* Firstly, note that since  $\rho_N := (\mathbb{I}/2^n)^{\otimes m}$ , we can explicitly write down the output of the noisy circuit on this state as  $\Phi_{\mathcal{C}, p}(\rho_N) = (\mathbb{I}/2^n)^{\otimes m}$ , as the maximally-mixed state is the fixed point of depolarizing noise.

Since  $P_k = \text{EstInput} \circ \mathcal{A}(\mathcal{C}, p, \mathcal{M}, \rho_k)$  and  $P_N = \text{EstInput} \circ \mathcal{A}(\mathcal{C}, p, \mathcal{M}, \mathbb{I}/2^n)$ , by the data-processing inequality we have

$$D(P_k \| P_N) \leq D(\Phi_{\mathcal{C}, p}(\rho_k)^{\otimes m} \| \mathbb{I}/2^n),$$

and so it suffices to bound the quantity  $D(\Phi_{\mathcal{C}, p}(\rho_k)^{\otimes m} \| \mathbb{I}/2^n)$ . The rest of this proof shows that for  $k \in \{0, 1\}^n$ :

$$D(\Phi_{\mathcal{C}, p}(\rho_k)^{\otimes m} \| (\mathbb{I}/2^n)^{\otimes m}) \leq p^{2D} nm \quad (83)$$

by the unitary invariance of the relative entropy, the strong data-processing inequality for depolarizing noise and the additivity of the relative entropy. To see this, observe that

$$D(\Phi_{\mathcal{C}, p}(\rho_k)^{\otimes m} \| (\mathbb{I}/2^n)^{\otimes m}) = m D(\Phi_{\mathcal{C}, p}(\rho_k) \| \mathbb{I}/2^n) \quad (84)$$

$$\leq m p^{2D} D(\rho_k \| \mathbb{I}/2^n) \quad (85)$$

$$= m p^{2D} D(\mathcal{U}^\dagger(|k\rangle\langle k|) \| \mathbb{I}/2^n) \quad (86)$$

$$= m p^{2D} D(|k\rangle\langle k| \| \mathbb{I}/2^n) \quad (87)$$

$$= p^{2D} mn, \quad (88)$$

where the first equality follows from additivity of relative entropy, the first inequality is strong data-processing for  $D$  layers of depolarizing noise with parameter  $p$ , [68]

$$D(\Phi_{\mathcal{C}, p}(\rho_k) \| \mathbb{I}/2^n) \leq p^{2D} D(\rho_k \| \mathbb{I}/2^n), \quad (89)$$

the third equality follows from unitary invariance of relative entropy, and the fourth equality follows from calculation. Lastly, we clearly have  $D(P_N \| P_N) = 0 < p^{2D} nm$ . Therefore

$$\frac{1}{N+1} \sum_{k=0}^N D(P_k \| P_N) < p^{2D} nm \quad (90)$$

as desired.  $\square$

Theorem 8 states that a general input state-agnostic error mitigation protocol should have a worst-case sample complexity scaling exponentially in  $D$ . So far, our proof method does *not* encompass input state-aware error mitigation protocols, which are those that, in addition to knowing the noiseless circuit, also know its input state. Such knowledge would be an unauthorized ‘back door’ for the error mitigation algorithm used inside the solver for noisy state discrimination.

Can we remedy this lack? We offer two affirmative answers. The first follows immediately from close inspection of our proof method: our results extend to even error mitigation protocols working with a guarantee that the input state is chosen from a known set of size  $N+1$ —a situation somewhere between input-state agnosticism and full input-state awareness. Secondly, we will next show that a modification of our argument allows us to make a statement about the complexity of fully input-state aware error mitigation algorithms.

### C. The input-state aware case

We cannot expect lower bounds on the sample complexity of the general error mitigation task in the input-state aware setting, because there always exists a trivial error mitigation algorithm that does not access the quantum device—hence uses zero samples—and instead simulates the quantum circuit to obtain the desired properties of the noiseless state. This would, of course, not be possible in the setting of the previous subsection, without the additional information about the input state to the circuit. While we previously found sample complexity bounds on *any* input state-agnostic error mitigation algorithm outputting expectation value estimates, we now change our perspective: we ask how many samples would be needed by an error mitigation algorithm whose outputs *are meaningfully different* from one that obtains comparable results through purely classical means. In other words, if an error mitigation algorithm is successful but it invokes the quantum device an insufficient number of times, then there exists an equivalent classical algorithm. A tool we will use is the following bound on two-hypothesis testing due to Le Cam:

**Lemma 9** (Two-hypothesis testing [24], Theorem 2.2iii). *For two probability distributions  $P_0, P_1$  satisfying*

$$\|P_0 - P_1\|_{TV} \leq \alpha \quad , \quad (91)$$

*the minimum probability of error of any test that draws a single sample and distinguishes  $P_0$  and  $P_1$  is*

$$P_{\text{error}} \geq \frac{1 - \alpha}{2} \quad . \quad (92)$$

Our results on the convergence of random quantum circuits to the maximally mixed state can easily be used to obtain bounds in the total variation distance by an application of Pinsker's inequality and the data processing inequality.

In the following theorem we prove that our construction can be used to make a statement about the sample complexity of input state aware error mitigation under depolarizing noise, as long as one demands that such error mitigation makes meaningful use of the quantum inputs. We then remark on how our technique can be extended even to the case of non-unital noise.

**Theorem 9** (Input state-awareness does not break the curse of exponentiality). *Let  $\mathcal{A}$  be an input state-aware weak error mitigation algorithm that takes:*

- *Quantum input: At most  $m$  copies of the noisy output state  $\Phi_{\mathcal{C}, \mathcal{N}}(\rho)$ .*
- *Classical input,  $CL(\mathcal{C}, \mathcal{N}, \rho)$ : Classical descriptions of  $\mathcal{C}$ , the noise channel  $\mathcal{N}$  acting on  $\mathcal{C}$  and the input state to  $\mathcal{C}$ , which we call  $\rho$ ,*

*and outputs a distribution over possible outcomes. Then either:*

- **Case I:** *For some  $c > 1$ ,  $mc^{-\tilde{\mathcal{O}}(nD)} = \Omega(1)$ , i.e., even for input-state aware error mitigation algorithms, exponential in  $n, D$  samples are needed; or*
- **Case II:** *no algorithm exists can distinguish between the outputs of  $\mathcal{A}(\Phi_{\mathcal{C}, \mathcal{N}}(\rho)^{\otimes m}, CL(\mathcal{C}, \mathcal{N}, \rho))$  and  $\mathcal{A}(\mathbb{I}/2^{\otimes n}, CL(\mathcal{C}, \mathcal{N}, \rho))$  with an error probability significantly bounded away from  $2/3$ .*

*Proof.* Suppose we are not in Case I. This implies that  $mc^{-\tilde{\mathcal{O}}(nD)} = \Theta(1)$ . In Sections IV and V we proved there exists a circuit  $\mathcal{C}^*$  such that for any algorithm  $\mathcal{A}$  run on either  $m$  copies of the state  $\Phi_{\mathcal{C}^*, \mathcal{N}}(\rho)$  or  $m$  copies of the maximally-mixed state,

$$D(\mathcal{A}(\Phi_{\mathcal{C}^*, \mathcal{N}}(\rho)^{\otimes m}, CL(\mathcal{C}^*, \mathcal{N}, \rho)) \| \mathcal{A}(\mathbb{I}/2^{n \otimes m}, CL(\mathcal{C}^*, \mathcal{N}, \rho))) \leq mD(\Phi_{\mathcal{C}^*, \mathcal{N}}(\rho) \| \mathbb{I}/2^n) \leq mc^{-\tilde{\mathcal{O}}(nD)} = \Theta(1) \quad , \quad (93)$$

therefore Lemma 9 implies that the minimum probability of error distinguishing the output of  $\mathcal{A}(\Phi_{\mathcal{C}^*, \mathcal{N}}(\rho)^{\otimes m}, CL(\mathcal{C}, \mathcal{N}, \rho))$  from the output of  $\mathcal{A}(\mathbb{I}/2^{n \otimes m}, CL(\mathcal{C}, \mathcal{N}, \rho))$  is  $1/2 - o(1)$ , proving the claim.  $\square$

We can prove a similar statement in the case of non-unital noise. However, unitary circuits affected by product non-unital noise are qualitatively different: whereas in the case of unital circuits the outputs will converge to the maximally mixed state asymptotically, non-unital random circuits do not converge to a single state. For instance, as we will show later in the setting of our toy example of noise being applied after 2-design, what happens instead is that after we apply  $D$  layers of 2-designs followed by noise, yielding the evolution  $T_D$ , we have

$$\mathbb{E} [\|T_D(\rho_i - \rho_j)\|_1] \leq 2^{\frac{n+1}{2}} q_n^{\frac{D}{2}}, \quad (94)$$

where  $q_n = \mathcal{O}(c_1^n)$  for  $c_1 < 1$  is an exponentially small in  $n$  parameter that depends on the noise channel and  $\rho_i, \rho_j$  orthogonal inputs. That is, any two inputs will be mapped to essentially the same state  $\sigma_D$  after  $D$  applications of the circuit. However, it is in general not possible to compute this state efficiently to simulate it classically and unless the noise is unital the sequence  $\sigma_1, \sigma_2, \dots$  does not converge. Thus, in the argument we used above, the maximally mixed state would have to be replaced by this state  $\sigma_D$ . Thus, the same conclusions would hold, but the output of the error mitigation algorithm would be indistinguishable from the output given as input copies of a fixed state  $\sigma_D$  that is independent of the input  $\rho$ . We cannot exclude the possibility that sampling from  $\sigma_D$  gives some extra computational power, but as in general there is no way to compute  $\sigma_D$ , it is unclear how to profit from it. In summary, the output of the error mitigation algorithm given the samples from the noisy device would be indistinguishable from that of the same error mitigation algorithm given as input copies of a state  $\sigma_D$  that is independent of the input.

*Interlude* — Let us review the conceptual contribution of this section and look ahead to the next. Theorem 8 leveraged a relation between a noisy state distinguishability problem and a related error mitigation problem to show

that the worst-case sample complexity of error mitigation needs to depend exponentially on the depth of the circuit whose errors are being mitigated. The proof of this theorem depended crucially on upper-bounding the quantity

$$\frac{1}{N+1} \sum_{k=0}^N D(\Phi_{\mathcal{C},p}(\rho_k) \parallel \Phi_{\mathcal{C},p}(\mathbb{I}/2^n)) \quad , \quad (95)$$

which is a sort of average distance between the output of the noisy circuit (whose errors we would like to mitigate) and its fixed point, which is the maximally-mixed state. This quantity controls the sample complexity of noisy state discrimination via Fano's Lemma, and in the chain of inequalities (84) to (88) of the proof of Proposition 1, we upper-bounded it by  $\alpha \log N$  where we computed  $\alpha$  in terms of  $m$  and  $D$ . Plugging into Fano's Lemma then completed the argument and yielded the lower bound on sample complexity  $m$ .

In the next section, we will refine the upper-bound on (95), so that  $\alpha$  has an exponential dependence not only on  $D$ , but also on  $n$ . Plugging this into Fano's Lemma yields, in turn, a worst-case sample complexity  $m$  that depends exponentially on *both* of these quantities. We will achieve this improvement by carefully constructing the input states  $\mathcal{C}^{-1}(|x\rangle\langle x|)$ . In fact, the key will be a perspicacious choice of  $\mathcal{C}$ , which leads to a worst-case-bound *better* than the bound on the right-hand-side of Eq. (89), which was proven for general  $\mathcal{C}$ .

#### IV. TECHNICAL BACKGROUND AND CONTRIBUTIONS

In this section, we show that there exists a 'rapidly mixing' quantum circuit  $\mathcal{C}^*$ . In so doing we prove a stronger version of our theorem. This circuit is rapidly mixing in the sense that, its output states

$$\Phi_{\mathcal{C}^*,p}(\mathbb{I}/2^n \otimes |0\rangle\langle 0|^{\otimes s}), \Phi_{\mathcal{C}^*,p}(\rho_k \otimes |0\rangle\langle 0|^{\otimes s}), \quad k \in \{0, 1\}^n, \quad (96)$$

for the input states  $\{\rho_k\}$  and the maximally mixed state (concatenated with ancilla qubits), satisfies

$$\frac{1}{N+1} \sum_{k=0}^N D(\Phi_{\mathcal{C}^*,p}(\rho_k \otimes |0\rangle\langle 0|^{\otimes s}) \parallel \Phi_{\mathcal{C}^*,p}(\mathbb{I}/2^n \otimes |0\rangle\langle 0|^{\otimes s})) < c^{\Omega(nD)}, \quad (97)$$

for some constant  $c < 1$  (c.f. Eq. (95)). In the previous section, in Eq. (90), for a generic  $\mathcal{C}$ , we could only prove that the average distinguishability could be bounded by  $p^{\Omega(D)}nm$ , which does not have the  $n$  dependence. Thus when  $D = O(\log n)$ , the old bound is exponentially weaker in  $n$ . After we prove the improved bound for our particular circuit  $\mathcal{C}^*$ , we will then wrap up by again plugging this bound into the information-theoretic outer loop of our argument, which uses the sample complexity of noisy state distinguishability to lower bound that of error mitigation.

##### A. Technical background

In fact, we do not construct  $\mathcal{C}^*$  explicitly. Instead, we let  $\mathcal{C}^*$  be a concatenation of blocks, i.e.,

$$\mathcal{C}^* = \tilde{\mathcal{C}}_\ell \circ \dots \circ \tilde{\mathcal{C}}_2 \circ \tilde{\mathcal{C}}_1, \quad (98)$$

where every *block* is a circuit  $\tilde{\mathcal{C}}_j$  sampled independently from an *ensemble* of circuits (described in Lemma 10). We will then show that the *expected*  $\mathcal{C}^*$ , when affected by depolarizing noise and run on the input states  $\mathbb{I}/2^n$  and  $\rho_k$ ,  $k \in \{0, 1\}^n$ , displays the scaling we claim in Eq. (97). We can then use the probabilistic method to conclude that there *exists* one particular circuit that achieves the same bound. Every block in  $\mathcal{C}^*$  is constructed via the following prescription by Cleve, Leung, Liu and Wang [33]:

**Lemma 10** (Exact unitary 2-designs from Clifford gates [33]). *With circuits  $\tilde{\mathcal{C}}$  consisting of  $O(n \log^2 n \log \log n)$  single- and two-qubit Clifford gates and  $\tilde{O}(n)$  auxiliary qubits, in  $O(\log^2(n))$  depth, it is possible to implement an exact unitary 2-design on  $n$  qubits. The auxiliary qubits all start in the state  $|0\rangle\langle 0|$  and are returned to this state at the end of the circuit.*

In other words, the above is a prescription for constructing Clifford circuits  $\tilde{\mathcal{C}}$  that form unitary 2 designs on a subset of the qubits. Each such circuit has the property that

$$\tilde{\mathcal{C}}(\rho \otimes |0\rangle\langle 0|^{\otimes s}) = \mathcal{C}(\rho) \otimes |0\rangle\langle 0|^{\otimes s}, \quad (99)$$

where  $s$  is the number of auxiliary qubits and  $\mathcal{C}$  is a unitary sampled from an exact unitary 2-design.

We want to consider a noisy version of the circuit  $\mathcal{C}^*$  defined in (98). This means that in between every layer of single- and two-qubit Clifford gates comprising each circuit  $\tilde{\mathcal{C}}_j$ , there is a layer of depolarizing noise on every qubit. In particular, the noisy version of the block  $\tilde{\mathcal{C}}_j$  is of the form

$$\Phi_{\tilde{\mathcal{C}}_j, \vec{p}_j} = \mathcal{D}_{p_{D'}}^{\otimes(n+s)} \circ \tilde{\mathcal{C}}_j^{(D')} \circ \dots \circ \mathcal{D}_{p_2}^{\otimes(n+s)} \circ \tilde{\mathcal{C}}_j^{(2)} \circ \mathcal{D}_{p_1}^{\otimes(n+s)} \circ \tilde{\mathcal{C}}_j^{(1)}, \quad (100)$$

where  $\vec{p}_j = (p_1, p_2, \dots, p_{D'})$  is a vector of single-qubit depolarizing noise parameters and  $D'$  is the depth of the construction of Ref. [33], as specified in Lemma 10 above. The superscript indexes the layer of the circuit and the subscript indexes the block. The noisy version of the overall circuit  $\mathcal{C}^*$ , which is the subject of our analysis, is then

$$\Phi_{\mathcal{C}^*, \vec{p}} := \Phi_{\tilde{\mathcal{C}}_\ell, \vec{p}_\ell} \circ \dots \circ \Phi_{\tilde{\mathcal{C}}_2, \vec{p}_2} \circ \Phi_{\tilde{\mathcal{C}}_1, \vec{p}_1}. \quad (101)$$

We now show that we can simplify the form of this noisy circuit by making use of an important property of Clifford gates, which can even be taken to be their defining property, which is that they map Pauli operators to other Pauli, up to a sign. This leads to the following result.

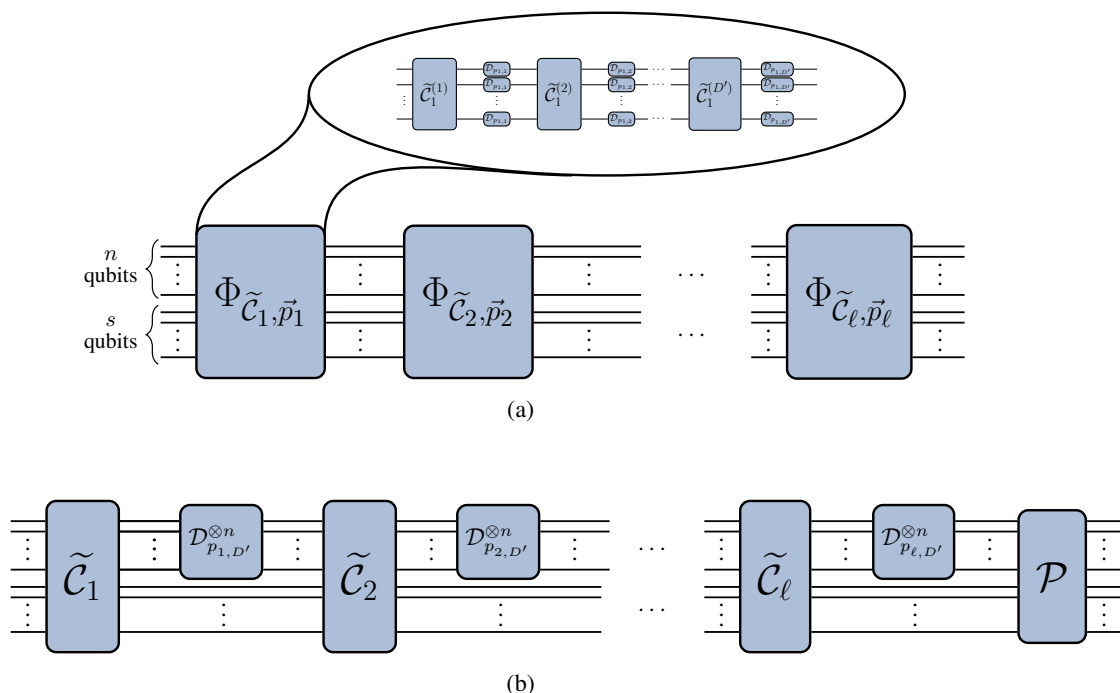


FIG. 3. A sketch of the constructions used in this subsection. Figure a) depicts the noisy circuit that results from our prescription of iterating the construction in Ref. [33]  $\ell$  times (refer to Eq. (101)) and using our model of local depolarizing noise. Figure b) depicts how we can re-write the circuit in Figure a) by pushing the depolarizing noise on the ancilla qubits to the end of the circuit, by invoking Lemma 12.

**Lemma 11.** *Let  $\mathcal{P}$  be a Pauli channel and let  $\mathcal{C}$  be a Clifford unitary. Then there exists a Pauli channel  $\mathcal{P}'$  such that*

$$\mathcal{C} \circ \mathcal{P} = \mathcal{P}' \circ \mathcal{C}. \quad (102)$$

*Proof.* Let  $\mathcal{P}' = \mathcal{C} \circ \mathcal{P} \circ \mathcal{C}^\dagger$ . It is clear that Eq. (102) holds. Furthermore, we see that letting  $\mathcal{P} = \sum_{P \in \{I, X, Y, Z\}^{\otimes n}} q_P P(\cdot) P$ , then

$$\mathcal{P}'(\rho) = \sum_{P \in \{I, X, Y, Z\}^{\otimes n}} q_P C P C^\dagger(\cdot) C^\dagger P C.$$

As  $C$  is a Clifford unitary,  $C P C^\dagger$  will also be a Pauli string, up to a sign. Thus, we conclude that the Pauli operators

of  $\mathcal{P}'$  are also Paulis and it is a Pauli channel. □

**Lemma 12** (Pushing Paulis to the end of the circuit). *Consider the noisy circuit defined in Eq. (101). There exists a Pauli channel  $\mathcal{P}_{n+s}$ , acting on all  $n + s$  qubits, such that*

$$\mathcal{N} = \mathcal{P}_{n+s} \circ (\mathcal{D}_{p_\ell}^{\otimes n} \otimes \text{id}_s) \circ \tilde{\mathcal{C}}_\ell \circ \dots \circ (\mathcal{D}_{p_2}^{\otimes n} \otimes \text{id}_s) \circ \tilde{\mathcal{C}}_2 \circ (\mathcal{D}_{p_1}^{\otimes n} \otimes \text{id}_s) \circ \tilde{\mathcal{C}}_1. \quad (103)$$

*In other words, all of the noise acting on the  $s$  auxiliary qubits and within the circuit for each design can be pushed to the end.*

*Proof.* First, note that all gates of the circuit are Clifford and all the noisy channels are assumed to be product of 1-qubit Pauli noise channels. As observed in Lemma 11, given a Pauli channel followed by a Clifford gate, we can always obtain the same transformation by first applying the gate followed by a (typically different) Pauli channel. Furthermore, Pauli channels all commute with each other. To see this, note that Pauli strings are the eigenvalues of Pauli channels. As they share a common eigenbasis, they all commute.

Thus, we can take all the individual depolarizing channels either acting on the auxiliary systems or within the blocks that generate each design and commute them to end of the circuit. This gives the desired representation. □

We briefly remark on how the above calculations will be used in our proof. We will again use the observation of Eq. (2) of the Central Lemma 1, but now with the noisy circuits we have specially constructed:  $\mathcal{C}' = \Phi_{\mathcal{C}^*, \bar{p}}$ . That is, consider a new instance of the noisy state distinguishability problem (Problem 3), in which the task is to identify the label  $k$  given a noisy state from the set

$$\Phi_{\mathcal{C}^*, \bar{p}}(|k\rangle\langle k| \otimes |0\rangle\langle 0|^{\otimes s}), \quad k \in \{0, 1\}^n, \quad \text{and } \Phi_{\mathcal{C}^*, \bar{p}}(\mathbb{I}_n/2^n \otimes |0\rangle\langle 0|^{\otimes s}) \quad (104)$$

Let us denote  $\rho_k = |k\rangle\langle k| \otimes |0\rangle\langle 0|^{\otimes s}$ . We then want to consider the quantity

$$D(\Phi_{\mathcal{C}^*, \bar{p}}(\rho_k) \| \Phi_{\mathcal{C}^*, \bar{p}}(\mathbb{I}_n/2^n \otimes |0\rangle\langle 0|^{\otimes s})). \quad (105)$$

Lemma 12 will simplify the task of bounding this quantity. Recall the result in (103). In particular, let

$$\tilde{\mathcal{N}} := (\mathcal{D}_{p_\ell}^{\otimes n} \otimes \text{id}_s) \circ \tilde{\mathcal{C}}_\ell \circ \dots \circ (\mathcal{D}_{p_2}^{\otimes n} \otimes \text{id}_s) \circ \tilde{\mathcal{C}}_2 \circ (\mathcal{D}_{p_1}^{\otimes n} \otimes \text{id}_s) \circ \tilde{\mathcal{C}}_1, \quad (106)$$

so that  $\Phi_{\mathcal{C}^*, \bar{p}} = \mathcal{P}_{n+s} \circ \tilde{\mathcal{N}}$ . This re-writing is useful because recursive application of (99) implies that the state  $\mathbb{I}_n/2^n \otimes |0\rangle\langle 0|^{\otimes s}$  is a fixed-point of  $\tilde{\mathcal{N}}$ , a fact that we use in Eq. (107) below. From the data-processing inequality, we obtain

$$\begin{aligned} D(\Phi_{\mathcal{C}^*, \bar{p}}(\rho_k) \| \Phi_{\mathcal{C}^*, \bar{p}}(\mathbb{I}_n/2^n \otimes |0\rangle\langle 0|^{\otimes s})) &= D((\mathcal{P}_{n+s} \circ \tilde{\mathcal{N}})(\rho_k) \| (\mathcal{P}_{n+s} \circ \tilde{\mathcal{N}})(\mathbb{I}_n/2^n \otimes |0\rangle\langle 0|^{\otimes s})) \\ &\leq D(\tilde{\mathcal{N}}(\rho_k) \| \mathbb{I}_n/2^n \otimes |0\rangle\langle 0|^{\otimes s}) \\ &\leq D\left(\left(\bigcirc_{j=1}^\ell \mathcal{D}_{p_j}^{\otimes n} \circ \mathcal{C}_j\right)(|k\rangle\langle k|) \otimes |0\rangle\langle 0|^{\otimes s} \| \tilde{\mathcal{N}}(\mathbb{I}_n/2^n \otimes |0\rangle\langle 0|^{\otimes s})\right) \\ &= D\left(\left(\bigcirc_{j=1}^\ell \mathcal{D}_{p_j}^{\otimes n} \circ \mathcal{C}_j\right)(|k\rangle\langle k|) \| \mathbb{I}_n/2^n\right), \end{aligned} \quad (107)$$

where we note that the channel  $\bigcirc_{j=1}^\ell \mathcal{D}_{p_j}^{\otimes n} \circ \mathcal{C}_j$  acts only on the  $n$  system qubits and not on the auxiliary ones. We can thus focus on the quantity

$$D\left(\left(\bigcirc_{j=1}^\ell \mathcal{D}_{p_j}^{\otimes n} \circ \mathcal{C}_j\right)(|k\rangle\langle k|) \| \mathbb{I}_n/2^n\right) \quad (108)$$

for the remainder of this section, where the circuits  $\mathcal{C}_j$  are independently drawn from a unitary 2-design. Since the uniform ensemble of Clifford unitaries forms a unitary 2-design, we assume from now on that the  $\mathcal{C}_j$  are Clifford unitary channels.

### B. Bound for multiple layers of depolarizing noise

Let us recall that our goal is to bound the *average distinguishability* quantity

$$\frac{1}{N+1} \sum_{k=0}^N D(P_k \| P_N). \quad (109)$$

It turns out that we can obtain a  $k$ -independent bound on  $D(P_k \| P_N)$ ; to do so, we invoke the following consequence of Lemma 2, which says that we can bound the expected relative entropy for each  $k$  by calculating the expected purity

$$\begin{aligned} \mathbb{E}_{C \sim \mathcal{E}} \left[ D \left( \left( \bigcirc_{j=1}^{\ell} \mathcal{D}_{p_j}^{\otimes n} \circ C_j \right) (|k\rangle\langle k|) \left\| \frac{\mathbb{I}}{2^n} \right\| \right) \right] &\leq n + \mathbb{E}_{C \sim \mathcal{E}} \left[ \log \text{Tr} \left[ \left( \bigcirc_{j=1}^{\ell} \mathcal{D}_{p_j}^{\otimes n} \circ C_j \right) (|k\rangle\langle k|)^2 \right] \right] \leq \\ n + \log \mathbb{E}_{C \sim \mathcal{E}} \text{Tr} \left[ \left( \bigcirc_{j=1}^{\ell} \mathcal{D}_{p_j}^{\otimes n} \circ C_j \right) (|k\rangle\langle k|)^2 \right], \end{aligned} \quad (110)$$

where  $k \in \{0, 1\}^n$  and the last step follows from Jensen's inequality. Let us now compute that purity. In Section V we will compute the expected purity of the outcome of  $l$  applications of a channel with a 2-design for arbitrary channels. In particular, in Prop. 4 we show that:

$$\mathbb{E}_{C \sim \mathcal{E}} \text{Tr} \left[ \left( \bigcirc_{j=1}^{\ell} \mathcal{D}_{p_j}^{\otimes n} \circ C_j \right) (|k\rangle\langle k|)^2 \right] = (q_n^{\ell} + 2^{-n}), \quad (111)$$

where for a qubit depolarizing channel with depolarizing probability  $p$  we have:

$$q_n = \frac{(1 - \frac{3}{4}p(2-p))^n - 2^{-2n}}{1 - 2^{-2n}}. \quad (112)$$

Inserting this bound in Eq. (110) we conclude that:

$$\mathbb{E}_{C \sim \mathcal{E}} \left[ D \left( \left( \bigcirc_{j=1}^{\ell} \mathcal{D}_{p_j}^{\otimes n} \circ C_j \right) (|k\rangle\langle k|) \left\| \frac{\mathbb{I}}{2^n} \right\| \right) \right] \leq n + \log(q_n^{\ell} + 2^{-n}) = \log(2^n q_n^{\ell} + 1) \leq 2^n q_n^{\ell} \quad (113)$$

In particular, if

$$\ell > \frac{1}{2 - \log_2(4 - 3p(2-p))}, \quad (114)$$

then

$$2^n q_n^{\ell} \leq c^n \quad (115)$$

for a constant  $c < 1$ .

### C. Putting it all together

We finally feed the results of the previous subsections into the information-theoretic outer loop already used in Section III, and obtain a tighter bound on sample complexity. For conciseness, we state Theorem 10 for depolarizing noise that acts with the same strength after every circuit layer; however, as demonstrated in Section IV B, the theorem holds with minor modifications for the case when the noise parameter differs for every layer.

**Theorem 10** (Error mitigation with depolarizing noise requires exponential-in- $n$ ,  $D$  samples). *Let  $\mathcal{A}$  be an input state-agnostic weak error mitigation algorithm (see Definition 2) that takes as input  $m$  noisy copies of the output state of a circuit  $\mathcal{C}$  acted upon by local depolarizing noise of parameter  $p$ . Suppose that for any circuit  $\mathcal{C}$  acting on  $n$  qubits of depth  $D$ , where*

$$D = \Omega \left( \frac{\log^2(n)}{\log \left( \frac{4}{1+3p^2} \right)} \right), \quad (116)$$

and the set of observables  $\mathcal{M} = \{Z_i\}_{i=1}^n$ ,  $\mathcal{A}(\mathcal{C}, \mathcal{M}, p)$  is able to produce estimates  $\{\hat{z}_i\}_{i \in [n]}$  such that with probability at least  $1 - \delta$ ,

$$|\hat{z}_i - \text{Tr}[Z_i \mathcal{C}(\rho)]| \leq \epsilon \quad \text{for all } i \in [n]. \quad (117)$$

Then as long as  $\epsilon < 1/2$ , there exists a constant  $c \geq 1$ , such that the number of noisy copies,  $m$ , needed by  $\mathcal{A}(\mathcal{C}, \mathcal{M}, p)$ , is lower-bounded as

$$m = c^{-\tilde{\mathcal{O}}(nD)}, \quad (118)$$

where the notation  $\tilde{\mathcal{O}}()$  hides polylogarithmic factors in  $n$ .

*Proof.* Let us first note that we must rescale  $n$  by a constant factor because it now refers to the size of both system and ancillas (the constant factor is because only  $s = \tilde{\mathcal{O}}(n)$  ancillas are required, according to [33]), while our bounds from the previous subsections were in terms of only the number of system qubits.

From Equation (113) and linearity of expectation we conclude that

$$\mathbb{E}_{\mathcal{C}} \left[ \frac{1}{N+1} \sum_{k=0}^N D \left( \mathcal{C}(|k\rangle\langle k| \otimes |0\rangle\langle 0|) \left\| \sum_E p(E) \frac{\mathbb{I}}{2^n} \otimes \sigma_E \right. \right) \right] \leq c^{\Omega(n\ell)}. \quad (119)$$

Here, the expectation is over the distribution induced by constructing a circuit  $\mathcal{C}$  by sequentially performing the randomized circuit construction of Ref. [33],  $\ell \geq \frac{1}{\log\left(\frac{4}{1+3p^2}\right)}$  times. Here we note that, as stated in the previous subsection, the ancillary states  $\sigma_E$  are a function of  $\mathcal{C}$ . By the probabilistic method, there thus exists at least one circuit  $\tilde{\mathcal{C}}$  of this form, with corresponding states  $\{\sigma_E\}_E$ , such that the average relative entropy is upper-bounded by the same quantity: the circuit's output states

$$\tilde{\rho}_k := \tilde{\mathcal{C}}(|k\rangle\langle k| \otimes |0\rangle\langle 0|) \quad \text{for } k = 0, \dots, N-1, \quad (120)$$

$$\tilde{\rho}_N := \tilde{\mathcal{C}} \left( \sum_E p(E) \frac{\mathbb{I}}{2^n} \otimes \sigma_E \right) \quad (121)$$

for  $N = 2^n$ , satisfy

$$\frac{1}{N+1} \sum_{k=0}^N D(\rho_k || \rho_N) \leq c^{\mathcal{O}(n\ell)} = c^{\mathcal{O}(nD/\log^2(n))}. \quad (122)$$

Wrapping up, we plug this into the Central Lemma (Lemma 1): we lower bound the sample complexity of mitigating the depolarizing noise in the circuit  $\mathcal{C}$  by the sample complexity of noisy state discrimination of the states  $\{\tilde{\rho}_0, \dots, \tilde{\rho}_N\}$ . This proceeds in much the same way as we proved Theorem 8, except that we do not even require all the estimates output by weak error mitigation for the state discrimination problem: we may throw away the information in the estimates  $\hat{z}_j$  for  $j > n$  (i.e., the Paulis on the ancillary qubits).  $\square$

## D. Strengthening our results

In this last subsection, we strengthen our results so that they hold for circuits that are restricted to use local gates only. We also improve the depth dependence and introduce a dependence on the light-cone size.

### 1. Geometrically local circuits

Our circuit construction works by iterating the construction of Ref. [33]  $D$  times. The reader may observe, however, that the above construction requires all-to-all connectivity. Bearing in mind that error mitigation is usually applied to near-term quantum circuits for which only some geometrically local architecture is available, we now discuss how our results transfer to the case when we limit our attention only to error mitigation on such circuits. We observe that, for one-dimensional circuit architectures, every gate between non-nearest neighbor qubits can be replaced by at most  $\mathcal{O}(n)$  nearest-neighbor gates, by first doing a cascade of  $n-1$  sequential SWAPs to bring the two input qubits next to each other, and then applying the appropriate gate on these now-nearest neighbors. In  $d$  dimensions, the number



of swaps that must be done is at most  $\mathcal{O}(n^{1/d})$ —as the maximum distance between two qubits scales as the radius of the circuit architecture graph. Thus, unsurprisingly, with limited connectivity, the exponential in  $n$  cost of error mitigation implied by our results only kicks in at depths in which the light-cone of all qubits is the entire system. And, for  $d$  dimensional architectures, this is  $\mathcal{O}(n^{1/d})$ .

## 2. Smaller depths and light-cone size dependence

Furthermore, we can easily use the construction above to also obtain a large relative entropy contraction at smaller depths. Indeed, suppose we partition the qubits into  $s$  disjoint subsets  $S_1, \dots, S_s$  such that each subset has maximal size  $\lceil n/s \rceil$ . After that, we independently apply the construction above *on each* subset. Then at depth

$$D \geq O\left(\frac{\log^2(n/s)}{\log\left(\frac{4}{1+3p^2}\right)}\right) \quad (123)$$

on each block the relative entropy will satisfy

$$\frac{1}{N+1} \sum_{k=0}^N D(\rho_k || \rho_N) \leq O(c^{-nD/s}). \quad (124)$$

As the state will be product across different subsets, the total relative entropy will be at most

$$\mathcal{O}\left(sc^{-nd/s}\right). \quad (125)$$

In particular, by picking, say  $s \sim n/\log^2(n)$ , we see that there are circuits such that at depths  $\mathcal{O}(\text{poly}(\log \log(n)))$ , the relative entropy will decay towards 0 faster than any polynomial in the number of qubits. Furthermore, another property this construction enjoys is that it only involves quantum circuits with light-cones of size at most  $\lceil n/s \rceil$ . Indeed, as each of the unitaries only acts on subset of qubits of this size, the light-cone of each qubit is of that size. From this we obtain the following.

**Theorem 11** (Error mitigation with depolarizing noise with small depth and light-cones). *Let  $\mathcal{A}$  be an input state-agnostic weak error mitigation algorithm (see Definition 2) that takes as input  $m$  noisy copies of the output state of a circuit  $\mathcal{C}$  acted upon by local depolarizing noise of parameter  $p$ . Suppose that for some  $L > 0$  for any circuit  $\mathcal{C}$  with light-cone of size  $L$  acting on  $n$  qubits of depth  $D$ , where*

$$D \geq O\left(\frac{\log^2(L)}{\log\left(\frac{4}{1+3p^2}\right)}\right), \quad (126)$$

and the set of observables  $\mathcal{M} = \{Z_i\}_{i=1}^n$ ,  $\mathcal{A}(\mathcal{C}, \mathcal{M}, p)$  is able to produce estimates  $\{\hat{z}_i\}_{i \in [n]}$  such that with probability at least  $1 - \delta$ ,

$$|\hat{z}_i - \text{Tr}[Z_i \mathcal{C}(\rho)]| \leq \epsilon \quad \text{for all } i \in [n]. \quad (127)$$

Then as long as  $\epsilon < 1/2$ , there exists a constant  $c \geq 1$ , such that the number of noisy copies,  $m$ , needed by  $\mathcal{A}(\mathcal{C}, \mathcal{M}, p)$ , is lower-bounded as

$$m = \frac{L}{n} c^{-\tilde{\mathcal{O}}(LD)}, \quad (128)$$

where the notation  $\tilde{\mathcal{O}}()$  hides polylogarithmic factors in  $n$ . In particular, by picking  $L = \Omega(\log^2(n))$ , we see that there are circuits of depth  $\text{poly}(\log \log(n))$  that require a superpolynomial number of samples to error-mitigate.

The theorem above shows that as long as light-cones are slightly above  $\log(n)$  (namely  $\log^2(n)$ ) and depths are slightly above constant ( $\log^2(\log(n))$ ), error mitigation is no longer efficient in general. However, note that it is possible to compute the expectation value of observables with light-cones of size  $\mathcal{O}(\log(n))$  efficiently classically. This proves that in general error mitigation can only provide a better exponent in the scaling of an algorithm to estimate expectation values when compared with classical algorithms.

## V. BEYOND UNITAL NOISE

So far all the results we have discussed only apply to local random circuits with unital noise. It is thus natural to ask how the picture changes when the noise is non-unital. This is because many physically relevant noise models, such as amplitude damping, are non-unital. However, going beyond the non-unital model brings with it several technical complications. First, as we show, in general the noisy circuit will not converge anymore to a fixed point. Second, the entropy of a state does not necessarily increase under a non-unital map. As our previous results were based on entropic considerations, our arguments do not apply anymore. Furthermore, if the outputs of the noisy circuit are not full-rank, as could be the case for noise models like amplitude damping, the relative entropy between the outputs given different inputs might even be infinitely large. This motivates us to look at a different distance measure as well. Indeed, we resort in this section to the standard technique of controlling the trace distance through the Hilbert-Schmidt distance.

In light of these complications, we now discuss a toy model for the convergence of noisy circuits where it is still possible to obtain explicit bounds on their convergence. More precisely, we consider the model in which we have a circuit composed of (global) 2-designs followed by a noisy channel, and then we iterate this construction.

We show below that if the noisy channel is a qubit product channel, in expectation the trace distance between outputs are exponentially small in both the depth and the number of qubits. Although global 2-designs require at least polylogarithmic depth to implement and we only consider the case in which the noise acts after we implement it, we believe that our results illustrate that in principle our conclusions in the previous sections should carry over to the non-unital setting for deep enough unitary circuits. We start by computing moments of noisy random circuits that will be crucial for our conclusions later on.

### A. Purity and overlap change after one noisy gate

Consider the state

$$\rho_{\text{out},i} := (\mathcal{N} \circ \mathcal{U})(\rho_i), \quad (129)$$

where  $\rho_i$  is an arbitrary  $d$ -dimensional state,  $\mathcal{N}$  is a quantum channel, and  $\mathcal{U}(\cdot) = U(\cdot)U^\dagger$ , with  $U$  a unitary. The idea is that  $\mathcal{N}$  represents the noise after we apply the gate. Let us investigate how the overlap and purity of states changes after we apply the 2-designs and the noise on top. Note that by controlling overlaps and purities it is then straightforward to control the Hilbert-Schmidt distance. Let us start by expressing the overlap in a more convenient way. The overlap between the two output states  $\rho_{\text{out},i}, \rho_{\text{out},j}$  is given by

$$\begin{aligned} \text{Tr}[\rho_{\text{out},i}\rho_{\text{out},j}] &= \text{Tr}[(\mathcal{N} \circ \mathcal{U})(\rho_i)(\mathcal{N} \circ \mathcal{U})(\rho_j)] \\ &= \text{Tr}[F((\mathcal{N} \circ \mathcal{U})(\rho_i) \otimes (\mathcal{N} \circ \mathcal{U})(\rho_j))], \end{aligned} \quad (130)$$

where  $F$  is the swap operator. Now, suppose that  $\mathcal{N}$  has the following Kraus decomposition

$$\mathcal{N}(\cdot) = \sum_{\ell=1}^r K_\ell(\cdot)K_\ell^\dagger. \quad (131)$$

Using this, we get that the expected overlap of the output states for  $U$  a 2-design is

$$\begin{aligned} \mathbb{E}_U[\text{Tr}[\rho_{\text{out},i}\rho_{\text{out},j}]] &= \sum_{\ell,\ell'=1}^r \mathbb{E}_U\left[\text{Tr}\left[F\left((K_\ell \otimes K_{\ell'})(U \otimes U)(\rho_i \otimes \rho_j)(U^\dagger \otimes U^\dagger)(K_\ell^\dagger \otimes K_{\ell'}^\dagger)\right)\right]\right] \\ &= \alpha \sum_{\ell,\ell'=1}^r \text{Tr}\left[F((K_\ell \otimes K_{\ell'})(K_\ell^\dagger \otimes K_{\ell'}^\dagger))\right] + \beta \sum_{\ell,\ell'=1}^r \text{Tr}\left[F(K_\ell \otimes K_{\ell'})F(K_\ell^\dagger \otimes K_{\ell'}^\dagger)\right], \end{aligned} \quad (132)$$

where for the last line we used the general fact (see, e.g., Ref. [69, Example 7.25]) that for 2-designs

$$\mathbb{E}_U[(U \otimes U)X(U^\dagger \otimes U^\dagger)] = \underbrace{\left(\frac{\text{Tr}[X]}{d^2-1} - \frac{\text{Tr}[FX]}{d(d^2-1)}\right)}_{\alpha} \mathbb{I} \otimes \mathbb{I} + \underbrace{\left(\frac{\text{Tr}[FX]}{d^2-1} - \frac{\text{Tr}[X]}{d(d^2-1)}\right)}_{\beta} F. \quad (133)$$

The involved quantities take the values

$$\mathrm{Tr}\left[F((K_\ell \otimes K_{\ell'}) (K_\ell^\dagger \otimes K_{\ell'}^\dagger))\right] = \mathrm{Tr}[K_\ell K_\ell^\dagger K_{\ell'} K_{\ell'}^\dagger], \quad (134)$$

$$\mathrm{Tr}\left[F(K_\ell \otimes K_{\ell'}) F(K_\ell^\dagger \otimes K_{\ell'}^\dagger)\right] = \mathrm{Tr}[K_{\ell'} K_\ell^\dagger \otimes K_\ell K_{\ell'}^\dagger] = |\mathrm{Tr}[K_\ell K_\ell^\dagger]|^2, \quad (135)$$

for all  $\ell, \ell' \in \{1, 2, \dots, r\}$  and

$$\alpha = \frac{1}{d^2 - 1} - \frac{\mathrm{Tr}[\rho_i \rho_j]}{d(d^2 - 1)}, \quad (136)$$

$$\beta = \frac{\mathrm{Tr}[\rho_i \rho_j]}{d^2 - 1} - \frac{1}{d(d^2 - 1)}. \quad (137)$$

We can rephrase these quantities in terms of global properties of the channel as

$$\sum_{\ell=1}^r K_\ell K_\ell^\dagger = \mathcal{N}(\mathbb{I}), \quad (138)$$

$$\sum_{\ell, \ell'=1}^r |\mathrm{Tr}[K_\ell K_{\ell'}^\dagger]|^2 = \mathrm{Tr}[(\Gamma^\mathcal{N})^2], \quad (139)$$

where

$$\Gamma^\mathcal{N} := (\mathrm{id} \otimes \mathcal{N})(|\Gamma\rangle\langle\Gamma|), \quad |\Gamma\rangle := \sum_{j=0}^{d-1} |j, j\rangle, \quad (140)$$

is the Choi representation of  $\mathcal{N}$ . Therefore, the expected overlap of the output states is

$$\mathbb{E}_U[\mathrm{Tr}[\rho_{\mathrm{out},i} \rho_{\mathrm{out},j}]] = \left(\frac{1}{d^2 - 1} - \frac{\mathrm{Tr}[\rho_i \rho_j]}{d(d^2 - 1)}\right) \mathrm{Tr}[\mathcal{N}(\mathbb{I})^2] + \left(\frac{\mathrm{Tr}[\rho_i \rho_j]}{d^2 - 1} - \frac{1}{d(d^2 - 1)}\right) \mathrm{Tr}[(\Gamma^\mathcal{N})^2]. \quad (141)$$

Note that both  $\mathrm{Tr}[(\Gamma^\mathcal{N})^2]$  and  $\mathrm{Tr}[\mathcal{N}(\mathbb{I})^2]$  scale with the underlying dimension of the space. In order to more easily grasp the scaling of the formula above, it is fruitful to look at the rescaled quantities

$$\eta = \mathrm{Tr}[\mathcal{N}(\mathbb{I}/d)^2], \quad \nu = \mathrm{Tr}[(\Gamma^\mathcal{N})^2/d^2]. \quad (142)$$

Note that  $\eta$  is just the purity of the output of the maximally mixed state and, thus, takes values in  $[d^{-1}, 1]$ , whereas  $\nu$  is just the purity of the output of the maximally entangled state if we act with the channel on one half of the system (i.e., the Choi state) and takes values in  $[d^{-2}, 1]$ . In terms of these newly defined quantities, we obtain

$$\mathbb{E}_U[\mathrm{Tr}[\rho_{\mathrm{out},i} \rho_{\mathrm{out},j}]] = \left(\frac{d^2}{d^2 - 1} - \frac{d \mathrm{Tr}[\rho_i \rho_j]}{(d^2 - 1)}\right) \eta + \left(\frac{d^2 \mathrm{Tr}[\rho_i \rho_j]}{d^2 - 1} - \frac{d}{(d^2 - 1)}\right) \nu. \quad (143)$$

**Remark 12.** *It is well-known that twirling any quantum channel with a unitary 2-design will lead to a global depolarizing noise in expectation, i.e.,  $\mathbb{E}_{U \sim \mathrm{Haar}}[\mathcal{U}^\dagger \circ \mathcal{N} \circ \mathcal{U}] = \mathcal{D}_p^{(d)}$  [70]. Note, however, that the problem at hand is slightly different, as we are interested in an expectation value that contains two copies of the channel. In that case, the resulting action cannot be effectively of a global depolarizing channel. To illustrate this point more concretely, assume that the noise is actually a unitary gate. In that case, the purity or overlap between two different initial states should remain constant for all unitaries, which differs from the depolarizing noise picture discussed before. For example, let us assume that  $\mathcal{N}$  is a unitary channel and  $\rho_i = \rho_j = \rho$ , which means that  $\mathcal{N}(\mathbb{I}) = \mathbb{I}$  and  $\mathrm{Tr}[(\Gamma^\mathcal{N})^2] = d^2$ . Then, we obtain*

$$\begin{aligned} \mathbb{E}_U[\mathrm{Tr}[\rho_{\mathrm{out},i} \rho_{\mathrm{out},j}]] &= \alpha d + \beta d^2 \\ &= d \left(\frac{1}{d^2 - 1} - \frac{\mathrm{Tr}[\rho^2]}{d(d^2 - 1)}\right) + d^2 \left(\frac{\mathrm{Tr}[\rho^2]}{d^2 - 1} - \frac{1}{d(d^2 - 1)}\right) \\ &= \mathrm{Tr}[\rho^2], \end{aligned} \quad (144)$$

as expected.

### B. Convergence of global random circuits

From Eq. (143), we obtain a recursive formula for the overlap and/or purity of outputs of independent 2–designs interspersed by noise. Letting

$$T_t = \bigcirc_{l=1}^t \mathcal{N} \circ \mathcal{U}_l \quad (145)$$

and  $\rho_{i,t} = T_t(\rho_i)$  for some initial state  $\rho_i$  we have:

$$\mathbb{E}_U[\text{Tr}[\rho_{i,t+1}\rho_{j,t+1}]] = \left( \frac{d^2}{d^2-1} - \frac{d \text{Tr}[\rho_{i,t}\rho_{j,t}]}{(d^2-1)} \right) \eta + \left( \frac{d^2 \text{Tr}[\rho_{i,t}\rho_{j,t}]}{d^2-1} - \frac{d}{(d^2-1)} \right) \nu \quad (146)$$

$$= \frac{d(d\nu - \eta)}{d^2-1} \text{Tr}[\rho_{i,t}\rho_{j,t}] + \frac{d(d\eta - \nu)}{d^2-1}. \quad (147)$$

The structure of the recursion becomes more apparent when discussed in abstract terms. We have that

$$\begin{aligned} f(t) &= Af(t-1) + B \\ &= A(Af(t-2) + B) + B \\ &= A^2f(t-2) + AB + B \\ &= A^2(Af(t-3) + B) + (A+1)B \\ &= A^3f(t-3) + (A^2 + A + 1)B \\ &= A^t f(0) + (A^{t-1} + \dots + A + 1)B. \end{aligned} \quad (148)$$

The right term is nothing but a geometric series, which under the assumption  $0 \leq A < 1$  yields

$$f(t) = A^t f(0) + \frac{1 - A^t}{1 - A} B. \quad (149)$$

Applied to our setting, we obtain

$$\mathbb{E}_U[\text{Tr}[\rho_{t,i}\rho_{t,j}]] = \left( \frac{d(d\nu - \eta)}{d^2-1} \right)^t \text{Tr}[\rho_{0,i}\rho_{0,j}] + \left[ 1 - \left( \frac{d(d\nu - \eta)}{d^2-1} \right)^t \right] \frac{d(d\eta - \nu)}{d^2-1 - d(d\nu - \eta)}. \quad (150)$$

The exponential behavior in  $t$  of this quantity and the applicability of the geometric series is determined by the term  $d(d\nu - \eta)/(d^2 - 1)$ . The fact that this term is always positive and strictly smaller than 1 if the channel is not unitary, indicating an exponential decay in  $t$ , is a consequence of the following proposition:

**Proposition 2.** *Let  $\mathcal{N} : \mathbb{M}_d \rightarrow \mathbb{M}_d$  be a quantum channel. Define  $\eta$  and  $\nu$  as in Eq. (142). Then*

$$d\nu - \eta \geq 0. \quad (151)$$

Furthermore,

$$\frac{d^2\nu - d\eta}{d^2-1} = 1 \quad (152)$$

if and only if the underlying channel is unitary.

*Proof.* First, note that for any linear superoperator  $\mathcal{L}$  we have  $\text{Tr}[[\Gamma]\langle\Gamma|(\text{id} \otimes \mathcal{L})(|\Gamma\rangle\langle\Gamma|)] = \text{Tr}[\mathcal{L}]$ , where by  $\text{Tr}[\mathcal{L}]$  we mean the trace of  $\mathcal{L}$  as a linear map. Thus, we see that  $\nu = d^{-2} \text{Tr}[\mathcal{N}^\dagger \mathcal{N}]$  is just the average of the square of the singular values of quantum channel. Now, note that  $d^{-\frac{1}{2}}\mathbb{1}$  is a normalized vector with respect to the Hilbert-Schmidt norm. Furthermore, let  $X_i$ ,  $1 \leq i \leq d^2 - 1$  be an orthonormal, self-adjoint basis for the space of traceless matrices.

Then, using the fact that

$$|\Gamma\rangle\langle\Gamma| = \frac{\mathbb{I}}{\sqrt{d}} \otimes \frac{\mathbb{I}}{\sqrt{d}} + \sum_{i=1}^{d^2-1} \bar{X}_i \otimes X_i, \quad (153)$$

we obtain

$$\nu = d^{-2} \text{Tr}[\mathcal{N}^\dagger \mathcal{N}] = d^{-2} \text{Tr}[\mathcal{N}(d^{-\frac{1}{2}}I)^2] + d^{-2} \sum_{i=1}^{d^2-1} \text{Tr}[\mathcal{N}(X_i)^2] = d^{-1}\eta + d^{-2} \sum_{i=1}^{d^2-1} \text{Tr}[\mathcal{N}(X_i)^2]. \quad (154)$$

As the second term in the sum above is clearly positive, we obtain the claim in Eq. (151). To obtain the claim in Eq. (152), note that the expression is clearly monotone increasing in  $\nu$  and decreasing in  $\eta$ . The maximum value for  $\nu$  is 1, which is attained for a unitary channel (pure Choi states are unitaries). Furthermore, for unitary channels we have  $\eta = d^{-1}$ . In this case, the constant contribution of Eq. (146) vanishes as expected.  $\square$

In particular, it follows that unless the channel is unitary, the expected trace norm between outputs will be exponentially small. To conclude that, we will first study the convergence in the Hilbert-Schmidt norm.

**Proposition 3.** *Let  $T_\ell = \bigcirc_{l=1}^\ell \mathcal{N} \circ \mathcal{U}_l$  be a random quantum channel where  $\mathcal{N} : \mathbb{M}_d \rightarrow \mathbb{M}_d$  is a fixed quantum channel and  $\mathcal{U}_l$  are drawn independently from a unitary 2-design on a  $d$ -dimensional space. Furthermore, let  $\eta, \nu$  be defined as in Eq. (142) and*

$$q = \frac{d(d\nu - \eta)}{d^2 - 1}. \quad (155)$$

Then, for any orthogonal pure states  $\rho_i, \rho_j$  we have

$$\mathbb{E}_U [\|T_D(\rho_i - \rho_j)\|_2^2] = 2q^D \quad (156)$$

and

$$\mathbb{E}_U [\|T_\ell(\rho_i - \rho_j)\|_1] \leq \sqrt{2dq^D} \quad (157)$$

*Proof.* First, observe that

$$\mathbb{E}_U [\|T_\ell(\rho_i - \rho_j)\|_2^2] = \mathbb{E}_U [\text{Tr}[\rho_{\ell,i}^2]] + \mathbb{E}_U [\text{Tr}[\rho_{\ell,j}^2]] - 2\mathbb{E}_U [\text{Tr}[\rho_{\ell,i}\rho_{\ell,j}]], \quad (158)$$

where  $\rho_{\ell,i} = T_\ell(\rho_i)$ . Inserting the expression in Eq. (150) for each term we get

$$\mathbb{E}_U [\|T_\ell(\rho_i - \rho_j)\|_2^2] = q^\ell (\text{Tr}[\rho_i^2] + \text{Tr}[\rho_j^2] - 2\text{Tr}[\rho_i\rho_j]), \quad (159)$$

as the rest of the terms cancel. Using that  $\text{Tr}[\rho_i\rho_j] = 0$  and  $\text{Tr}[\rho_i^2] = \text{Tr}[\rho_j^2] = 1$  we arrive at Eq. (156). Next, from Jensen's inequality, it follows that

$$\mathbb{E} [\|T_\ell(\rho_i - \rho_j)\|_1]^2 \leq \mathbb{E} [\|T_\ell(\rho_i - \rho_j)\|_2^2]. \quad (160)$$

Furthermore, we have the standard inequality  $\|X\|_1 \leq \sqrt{d}\|X\|_2$  for all Hermitian  $X$ . Combining this observation with the claim in Eq. (160) yields Eq. (157).  $\square$

Let us investigate the case of  $d = 2^n$  ( $n$  qubits) and  $\mathcal{N} = \bigotimes_{i=1}^n \mathcal{N}_i$  (i.e.,  $\mathcal{N}$  is a product channel) in a bit more detail. Furthermore, assume for simplicity that all  $\mathcal{N}_i$  are the same. As the purity is multiplicative, we obtain  $\eta_{\mathcal{N}} = \eta_{\mathcal{N}_i}^n$ ,  $\nu_{\mathcal{N}} = \nu_{\mathcal{N}_i}^n$  and

$$q_n = \frac{2^{2n}\nu_{\mathcal{N}_i}^n - 2^n\eta_{\mathcal{N}_i}^n}{2^{2n} - 1}. \quad (161)$$

Thus, we see that the 2-norm decays exponentially in the number of qubits, even after one layer. In particular, we obtain the following corollary.

**Corollary 1.** *In the same setting of Proposition 3, assume further that  $\mathcal{N}$  is an  $n$ -qubit product channel and define  $q_n$  as in Eq. (161). Then we have*

$$\mathbb{E} [\|T_\ell(\rho_i - \rho_j)\|_1] \leq 2^{\frac{n+1}{2}} q_n^{\frac{\ell}{2}}. \quad (162)$$

*Proof.* The statement then follows from Proposition 3 and the discussion above, as

$$\mathbb{E} [\|T_\ell(\rho_i - \rho_j)\|_1] \leq 2^{\frac{n+1}{2}} q_n^{\frac{\ell}{2}}. \quad (163)$$

□

We can combine the statement above with sub-additivity of the trace distance to conclude that for  $m$  copies of the outputs of the noisy circuit we have

$$\mathbb{E} [\|T_\ell^{\otimes m}(\rho_i^{\otimes m} - \rho_j^{\otimes m})\|_1] \leq m 2^{\frac{n+1}{2}} \left( \frac{2^{2n} \nu_{\mathcal{N}_i}^n - 2^n \eta_{\mathcal{N}_i}^n}{2^{2n} - 1} \right)^{\frac{\ell}{2}}. \quad (164)$$

As explained before, successful error mitigation can be used to distinguish orthogonal input states. However, it is well-known that the trace distance bounds the probability of successful distinction between two states. Thus, we conclude the following.

**Theorem 13** (Error mitigation with non-unital noise). *Any input agnostic error mitigation procedure that works successfully for outputs of  $T_\ell$  would require a number of copies  $m$  that grows as*

$$m \geq 2^{-\frac{n+1}{2}} \left( \frac{2^{2n} \nu_{\mathcal{N}_i}^n - 2^n \eta_{\mathcal{N}_i}^n}{2^{2n} - 1} \right)^{-\frac{\ell}{2}}. \quad (165)$$

As discussed before,  $\nu_{\mathcal{N}_i} < 1$  unless we have a unitary channel. Thus  $m$  will be  $\Omega(c^{-n\ell})$  for some  $c < 1$  at a number of iterations  $\ell = \mathcal{O}(1)$ , as

$$\left( \frac{2^{2n} \nu_{\mathcal{N}_i}^n - 2^n \eta_{\mathcal{N}_i}^n}{2^{2n} - 1} \right)^{\frac{\ell}{2}} \leq 2^{-\frac{n+1}{2}} \quad (166)$$

for  $D = \mathcal{O}(1)$ .

The input aware case is a bit more subtle. What we conclude from the results above is that all input states will be mapped by  $T_D$  to the same state up to an exponential small correction. Thus, the noisy circuit is essentially a replacer channel and any error mitigation algorithm with sub-exponential samples would have the same performance given copies of this one state. However, unlike it was the case with the maximally mixed state, it is unclear that it is possible to classically simulate that state or that it is not a valuable computational resource. However, it is of course also unclear how a practical error mitigation might profit from access to samples from an (unknown) fixed state.

### C. Bounds on the probability of successful virtual distillation

It is straightforward to use the result introduced above to bound the probability of virtual distillation protocols [20, 21] succeeding under non-unital noise. As discussed before, the probability of such protocols succeeding is bounded by  $\text{Tr}[\rho^2]$ . If we introduce the quantity

$$r_n = \frac{2^{2n} \eta_{\mathcal{N}_i}^n - 2^n \nu_{\mathcal{N}_i}^n}{2^{2n} - 1}, \quad (167)$$

we can show the following:

**Proposition 4.** *In the same setting of Proposition 3, assume further that  $\mathcal{N}$  is an  $n$ -qubit product channel and define  $q_n$  as in Eq. (161) and  $r_n$  as in Eq. (167). Then we have for all pure states  $\rho$*

$$\mathbb{E} [\text{Tr} [T_\ell(\rho)^2]] = \left( q_n^\ell + \frac{(1 - q_n^\ell) r_n}{1 - q_n} \right). \quad (168)$$

Furthermore, both  $r_n$  and  $q_n$  are exponentially small in  $n$  unless the quantum channel  $\mathcal{N}_i$  is unitary or a replacer channel for a pure state, i.e.,  $\mathcal{N}_i(\sigma) = |\psi\rangle\langle\psi|$  for all  $\sigma$  and some  $|\psi\rangle$ .

*Proof.* The claim in Eq. (168) follows from Eq. (150), inserting the definition of  $q_n$  and  $r_n$ . As  $q_n \leq \frac{\nu_{\mathcal{N}_i}^n}{1-2^{-2n}}$  and  $\nu_{\mathcal{N}_i} = 1$  if, and only if,  $\mathcal{N}_i$  is a unitary channel, we see that  $q_n$  will decay exponentially unless the channel is unitary. Let us now analyse  $r_n$ , assuming the channel is not unitary. Note that it will also decay exponentially if we manage to show that  $\eta_{\mathcal{N}_i} < 1$  if the channel is not a replacer with a pure state. As  $\eta_{\mathcal{N}_i}$  is the purity of the output of the maximally mixed state, if  $\eta_{\mathcal{N}_i} = 1$  then the maximally mixed state is mapped to a pure state  $|\psi\rangle\langle\psi|$ . We will now show that this implies that all states are mapped to the same pure state. By linearity, it suffices to show the claim for a pure input state  $|\phi\rangle\langle\phi|$ . For this, note that

$$\frac{1}{2}|\phi\rangle\langle\phi| \leq \frac{\mathbb{I}}{2} \implies \frac{1}{2}\mathcal{N}_i(|\phi\rangle\langle\phi|) \leq \mathcal{N}_i\left(\frac{\mathbb{I}}{2}\right) = |\psi\rangle\langle\psi|.$$

The fact that  $\frac{1}{2}\mathcal{N}_i(|\phi\rangle\langle\phi|) \leq |\psi\rangle\langle\psi|$  implies that  $\frac{1}{2}\mathcal{N}_i(|\phi\rangle\langle\phi|) = \frac{1}{2}|\psi\rangle\langle\psi|$ . Indeed, if it was not the case,  $\mathcal{N}_i(|\phi\rangle\langle\phi|)$  would have nonzero support on the subspace orthogonal to  $|\psi\rangle$ , which would contradict the inequality. Thus, we conclude that  $\eta_{\mathcal{N}_i} < 1$  for non-replacer channels and hence that  $r_n$  is exponentially small in  $n$ .  $\square$

Thus, it follows that the probability of success of virtual distillation, which is upper-bounded by Eq. (168) will also be exponentially small even after a *single* layer of noise unless we are in the unitary or replacer case. But for unitary errors virtual distillation does not have any effect, as the input state is anyways pure. And for replacer channels it is clear that, even though the procedure will succeed with probability 1, we will just output a fixed pure product state. So in the cases where the success probability is not exponentially small, the procedure is of no use. However, note that it will not converge to the minimal possible value of the purity ( $2^{-n}$ ) unless the channel is unital, but rather to  $r_n$ . This is to be expected, as the noisy quantum circuit will not converge to the maximally mixed state in this case.

## VI. OUTLOOK

In this work, we have established a general rigorous framework that encapsulates large classes of schemes for quantum error mitigation that are being used in practice, but also others that are primarily of conceptual value. For these schemes, all of which come along with little additional experimental effort, we have identified severe information-theoretic limitations, limitations that are exponentially tighter than what was previously known. As our results feature a *worst-case* construction of circuits, further empirical investigation is needed to determine whether typical circuit instances for near-term quantum algorithms will fall into the regime we have described. These results do not imply that one should not perform quantum error mitigation to tackle suitably small noise levels in existing quantum architectures. At the same time, this work identifies substantial limitations that arise when one aims at scaling up such ideas to near-term quantum devices. In fact, this work can be seen as a strong indication that in the medium term, some form of quantum error correction involving quantum redundancy will presumably be necessary. It may also mean new schemes might be required that are intermediate between mere quantum error mitigation and resource expensive fault tolerant quantum computing. It is the hope that the present work inspires such further steps aimed at combating quantum noise in near-term quantum devices, to move universal quantum computers closer to reality.

## VII. ACKNOWLEDGMENTS

This work has been supported by the BMBF (RealistiQ), the BMWK (PlanQK), the DFG (CRC 183), the Einstein Foundation (Einstein Research Unit on Quantum Devices), the Quant-ERA (HQCC), and the Alexander von Humboldt Foundation. The research is also part of the Munich Quantum Valley (K-8), which is supported by the Bavarian state government with funds from the Hightech Agenda Bayern Plus. DSF acknowledges financial support from the VILLUM FONDEN via the QMATH Centre of Excellence (Grant no. 10059), the QuantERA ERA-NET Cofund in Quantum Technologies implemented within the European Union's Horizon 2020 Program (QuantAlgo) via the Innovation Fund Denmark and from the European Research Council (grant agreement no. 81876). DSF acknowledges that this work benefited from a government grant managed by the Agence Nationale de la Recherche under the Plan

France 2030 with the reference ANR-22-PETQ-0007.

- 
- [1] R. P. Feynman, Quantum mechanical computers, *Found. Phys.* **16**, 507 (1986).
- [2] D. Deutsch, Quantum theory, the Church-Turing principle and the universal quantum computer, *Proc. Roy. Soc. A* **400**, 97 (1985).
- [3] P. W. Shor, Algorithms for quantum computation: discrete logarithms and factoring, *Proc. 50th Ann. Symp. Found. Comp. Sc.*, 124 (1994).
- [4] P. W. Shor, Scheme for reducing decoherence in quantum computer memory, *Phys. Rev. A* **52**, R2493 (1995).
- [5] A. M. Steane, Error correcting codes in quantum theory, *Phys. Rev. Lett.* **77**, 793 (1996).
- [6] D. Gottesman, An introduction to quantum error correction and fault-tolerant quantum computation, (2009), [arXiv:0904.2557](https://arxiv.org/abs/0904.2557).
- [7] E. T. Campbell, B. M. Terhal, and C. Vuillot, Roads towards fault-tolerant universal quantum computation, *Nature* **549**, 172 (2017).
- [8] D. Litinski, A game of surface codes: Large-scale quantum computing with lattice surgery, *Quantum* **3**, 12 (2019).
- [9] Y. Li and S. C. Benjamin, Efficient variational quantum simulator incorporating active error minimization, *Phys. Rev. X* **7**, 021050 (2017).
- [10] K. Temme, S. Bravyi, and J. M. Gambetta, Error mitigation for short-depth quantum circuits, *Phys. Rev. Lett.* **119**, 180509 (2017).
- [11] S. Endo, S. C. Benjamin, and Y. Li, Practical quantum error mitigation for near-future applications, *Phys. Rev. X* **8**, 031027 (2018).
- [12] P. Suchsland, F. Tacchino, M. H. Fischer, T. Neupert, P. K. Barkoutsos, and I. Tavernelli, Algorithmic error mitigation scheme for current quantum processors, *Quantum* **5**, 492 (2021).
- [13] Y. Kim, C. J. Wood, T. J. Yoder, S. T. Merkel, J. M. Gambetta, K. Temme, and A. Kandala, Scalable error mitigation for noisy quantum circuits produces competitive expectation values, (2021), [arXiv:2108.09197](https://arxiv.org/abs/2108.09197).
- [14] C. Piveteau, D. Sutter, S. Bravyi, J. M. Gambetta, and K. Temme, Error mitigation for universal gates on encoded qubits, *Phys. Rev. Lett.* **127**, 200505 (2021).
- [15] V. Russo, A. Mari, N. Shammah, R. LaRose, and W. J. Zeng, Testing platform-independent quantum error mitigation on noisy quantum computers, (2022), [arXiv:2210.07194](https://arxiv.org/abs/2210.07194).
- [16] Z. Cai, R. Babbush, S. C. Benjamin, S. Endo, W. J. Huggins, Y. Li, J. R. McClean, and T. E. O'Brien, Quantum error mitigation, [arXiv:2210.00921](https://arxiv.org/abs/2210.00921).
- [17] P. Czarnik, A. Arrasmith, P. J. Coles, and L. Cincio, Error mitigation with Clifford quantum-circuit data, *Quantum* **5**, 592 (2022).
- [18] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, The theory of variational hybrid quantum-classical algorithms, *New J. Phys.* **18**, 023023 (2016).
- [19] F. Verstraete, M. M. Wolf, and J. I. Cirac, Quantum computation and quantum-state engineering driven by dissipation, *Nature Phys.* **5**, 633 (2009).
- [20] W. J. Huggins, S. McArdle, T. E. O'Brien, J. Lee, N. C. Rubin, S. Boixo, K. B. Whaley, R. Babbush, and J. R. McClean, Virtual distillation for quantum error mitigation, *Phys. Rev. X* **11**, 041036 (2021).
- [21] B. Koczor, Exponential error suppression for near-term quantum devices, *Phys. Rev. X* **11**, 031057 (2022).
- [22] C. Piveteau, D. Sutter, and S. Woerner, Quasiprobability decompositions with reduced sampling overhead, *npj Quant. Inf.* **8**, 12 (2022).
- [23] S. Endo, S. C. Benjamin, and Y. Li, Practical quantum error mitigation for near-future applications, *Phys. Rev. X* **8**, 031027 (2018).
- [24] A. B. Tsybakov, *Introduction to non-parametric estimation* (Springer New York, 2009).
- [25] R. Takagi, S. Endo, S. Minagawa, and M. Gu, Fundamental limits of quantum error mitigation, *npj Quant. Inf.* **8**, 114 (2022).
- [26] R. Takagi, H. Tajima, and M. Gu, Universal sample lower bounds for quantum error mitigation (2022), [arXiv:2208.09178](https://arxiv.org/abs/2208.09178).
- [27] G. De Palma, M. Marvian, C. Rouzé, and D. Stilck Franca, Limitations of variational quantum algorithms: a quantum optimal transport approach (2022), [arXiv:2204.03455](https://arxiv.org/abs/2204.03455).
- [28] S. Wang, P. Czarnik, A. Arrasmith, M. Cerezo, L. Cincio, and P. J. Coles, Can error mitigation improve trainability of noisy variational quantum algorithms? (2021), [arXiv:2109.01051](https://arxiv.org/abs/2109.01051).
- [29] S. Thanasilp, S. Wang, M. Cerezo, and Z. Holmes, Exponential concentration and untrainability in quantum kernel methods, (2022), [arXiv:2208.11060](https://arxiv.org/abs/2208.11060).
- [30] A. Deshpande, P. Niroula, O. Shtanko, A. V. Gorshkov, B. Fefferman, and M. J. Gullans, Tight bounds on the convergence of noisy random circuits to the uniform distribution (2021), [arXiv:2112.00716](https://arxiv.org/abs/2112.00716).
- [31] D. Stilck Franca and R. García-Patrón, Limitations of optimization algorithms on noisy quantum devices, *Nature Phys.* **17**, 1221 (2021).
- [32] S. Wang, E. Fontana, M. Cerezo, K. Sharma, A. Sone, L. Cincio, and P. J. Coles, Noise-induced barren plateaus in variational quantum algorithms, *Nature Comm.* **12**, 6961 (2021).



- [33] R. Cleve, D. Leung, L. Liu, and C. Wang, Near-linear constructions of exact unitary 2-designs, *Quant. Inf. Comp.* **16**, 721–756 (2016).
- [34] E. Farhi, J. Goldstone, and S. Gutmann, A quantum approximate optimization algorithm, (2014), [arXiv:1411.4028](https://arxiv.org/abs/1411.4028).
- [35] L. Zhou, S.-T. Wang, S. Choi, H. Pichler, and M. Lukin, Quantum approximate optimization algorithm: Performance, mechanism, and implementation on near-term devices, *Phys. Rev. X* **10**, 021067 (2020).
- [36] V. Feldman, A general characterization of the statistical query complexity, in *Proceedings of the 2017 Conference on Learning Theory* (PMLR, 2017) pp. 785–830.
- [37] L. Reyzin, Statistical queries and statistical algorithms: Foundations and applications (2020), [arXiv:2004.00557](https://arxiv.org/abs/2004.00557).
- [38] A. Blum, M. Furst, J. Jackson, M. Kearns, Y. Mansour, and S. Rudich, Weakly learning DNF and characterizing statistical query learning using Fourier analysis, in *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '94 (Association for Computing Machinery, New York, NY, USA, 1994) p. 253–262.
- [39] Y. G. Yatacos, Rates of convergence of minimum distance estimators and Kolmogorov's entropy, *Ann. Stat.* **13**, 768 (1985).
- [40] S. Bravyi, A. Kliesch, R. Koenig, and E. Tang, Obstacles to variational quantum optimization from symmetry protection, *Phys. Rev. Lett.* **125**, 260505 (2020).
- [41] E. Farhi, D. Gamarnik, and S. Gutmann, The quantum approximate optimization algorithm needs to see the whole graph: Worst case examples, (2020), [arXiv:2005.08747](https://arxiv.org/abs/2005.08747).
- [42] L. Eldar and A. W. Harrow, Local Hamiltonians whose ground states are hard to approximate, in *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)* (2017) pp. 427–438.
- [43] A. Anshu, N. P. Breuckmann, and C. Nirkhe, NLTS Hamiltonians from good quantum codes, (2022), [arXiv:2206.13228](https://arxiv.org/abs/2206.13228).
- [44] D. S. França, S. Strelchuk, and M. Studziński, Efficient classical simulation and benchmarking of quantum processes in the Weyl basis, *Phys. Rev. Lett.* **126**, 210502 (2021).
- [45] P. Rall, D. Liang, J. Cook, and W. Kretschmer, Simulation of qubit quantum circuits via Pauli propagation, *Phys. Rev. A* **99**, 062337 (2019).
- [46] S. Bravyi, D. Gosset, and R. Movassagh, Classical algorithms for quantum mean values, *Nature Phys.* **17**, 337 (2021).
- [47] Note that this metric is marginally weaker than Eq. (27), which requires the algorithm to succeed with  $1 - \delta$  probability at predicting expectation values for *all* observables.
- [48] K. Tsubouchi, T. Sagawa, and N. Yoshioka, Universal cost bound of quantum error mitigation based on quantum estimation theory (2022), [arXiv:2208.09385](https://arxiv.org/abs/2208.09385).
- [49] M. Mosonyi and F. Hiai, On the quantum Rényi relative entropies and related capacity formulas, *IEEE Trans. Inf. Th.* **57**, 2474 (2011).
- [50] T. Giurgica-Tiron, Y. Hindy, R. LaRose, A. Mari, and W. J. Zeng, Digital zero noise extrapolation for quantum error mitigation, in *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)* (2020) pp. 306–316, [2005.10921](https://arxiv.org/abs/2005.10921).
- [51] Z. Cai, Multi-exponential error extrapolation and combining error mitigation techniques for NISQ applications, *npj Quant. Inf.* **7**, 80, 2007.01265.
- [52] A. Lowe, M. H. Gordon, P. Czarnik, A. Arrasmith, P. J. Coles, and L. Cincio, Unified approach to data-driven quantum error mitigation, *Phys. Rev. Res.* **3**, 033098.
- [53] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, Variational quantum algorithms, *Nature Rev. Phys.* **3**, 625.
- [54] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke, W.-K. Mok, S. Sim, L.-C. Kwek, and A. Aspuru-Guzik, Noisy intermediate-scale quantum algorithms, *Rev. Mod. Phys.* **94**, 015004 (2022).
- [55] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, A variational eigenvalue solver on a photonic quantum processor, *Nature Comm.* **5**, 4213.
- [56] J. Bretagnolle and C. Huber, Estimation des densités : Risque minimax, in *Séminaire de Probabilités XII*, Vol. 649, edited by C. Dellacherie, P. A. Meyer, and M. Weil (Springer Berlin Heidelberg) pp. 342–363.
- [57] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, Quantum state tomography via compressed sensing, *Phys. Rev. Lett.* **105**, 150401 (2010).
- [58] M. Kearns, Efficient noise-tolerant learning from statistical queries, *J. ACM* **45**, 983–1006 (1998).
- [59] M. Hinsche, M. Ioannou, A. Nietner, J. Haferkamp, Y. Quek, D. Hangleiter, J.-P. Seifert, J. Eisert, and R. Sweke, A single  $t$ -gate makes distribution learning hard (2022), [arXiv:22207.03140](https://arxiv.org/abs/22207.03140).
- [60] V. Kanade, Advanced machine learning - Hilary term 2017 7: Statistical query learning, <https://www.cs.ox.ac.uk/people/varun.kanade/teaching/AML-HT2017/lectures/lecture07.pdf>, [Online; accessed 2022-08-05].
- [61] S. Arunachalam, A. B. Grilo, and H. Yuen, Quantum statistical query learning, (2020), [arXiv:2002.08240](https://arxiv.org/abs/2002.08240).
- [62] D. S. França and R. Garcia-Patron, A game of quantum advantage: Linking verification and simulation, (2020), [arXiv:2011.12173](https://arxiv.org/abs/2011.12173).
- [63] C. Dankert, R. Cleve, J. Emerson, and E. Livine, Exact and approximate unitary 2-designs and their application to fidelity estimation, *Phys. Rev. A* **80**, 012304 (2009).
- [64] D. Gross, K. Audenaert, and J. Eisert, Evenly distributed unitaries: On the structure of unitary designs, *J. Math. Phys.* **48**, 052104 (2007).
- [65] D. Gottesman, Stabilizer codes and quantum error correction, (1997), [quant-ph/9705052](https://arxiv.org/abs/quant-ph/9705052).

- [66] A. Calderbank, E. Rains, P. Shor, and N. Sloane, Quantum error correction via codes over  $GF(4)$ , *IEEE Trans. Inf. Theory* **44**, 1369 (1998).
- [67] D. DiVincenzo, D. Leung, and B. Terhal, Quantum data hiding, *IEEE Trans. Inf. Theory* **48**, 580 (2002).
- [68] C. Hirche, C. Rouzé, and D. S. França, *On contraction coefficients, partial orders and approximation of capacities for quantum channels* (2020).
- [69] J. Watrous, *The theory of quantum information* (Cambridge University Press, 2018).
- [70] J. Emerson, R. Alicki, and K. Życzkowski, Scalable noise estimation with random unitary operators, *J. Opt. B* **7**, S347 (2005).