Quantum advantages for Pauli channel estimation

Senrui Chen, Sisi Zhou, Alireza Seif, and Liang Jiang

Phys. Rev. A **105**, 032435 — Published 22 March 2022

DOI: 10.1103/PhysRevA.105.032435

Quantum advantages for Pauli channel estimation

Senrui Chen,[1] Sisi Zhou,[1,2] Alireza Seif,[1] and Liang Jiang[1]

[1] *Pritzker School of Molecular Engineering, The University of Chicago, Illinois 60637, USA*
[2] *Institute for Quantum Information and Matter, California Institute of Technology, Pasadena, CA 91125, USA*

(Dated: March 1, 2022)

We show that entangled measurements provide an exponential advantage in sample complexity for Pauli channel estimation, which is both a fundamental problem and a practically important subroutine for benchmarking near-term quantum devices. The specific task we consider is to simultaneously learn all the eigenvalues of an n-qubit Pauli channel to $\pm\varepsilon$ precision. We give an estimation protocol with an n-qubit ancilla that succeeds with high probability using only $\mathcal{O}(n/\varepsilon^2)$ copies of the Pauli channel, while prove that any ancilla-free protocol (possibly with adaptive control and channel concatenation) would need at least $\Omega(2^{n/3})$ rounds of measurement. We further study the advantages provided by a small number of ancillas. For the case that a k-qubit ancilla ($k \leq n$) is available, we obtain a sample complexity lower bound of $\Omega(2^{(n-k)/3})$ for any non-concatenating protocol, and a stronger lower bound of $\Omega(n2^{n-k})$ for any non-adaptive, non-concatenating protocol, which is shown to be tight. We also show how to apply the ancilla-assisted estimation protocol to a practical quantum benchmarking task in a noise-resilient and sample-efficient manner, given reasonable noise assumptions. Our results provide a practically-interesting example for quantum advantages in learning and also bring new insight for quantum benchmarking.

I. INTRODUCTION

One important challenge for the Noisy Intermediate-Scale Quantum (NISQ) era [1] is to demonstrate quantum advantages on near-term devices. Recent works have made groundbreaking progress towards demonstrating *quantum computational advantages* [2–5], which means quantum computers can efficiently solve certain computational problems outside the reach of the most advanced classical computers. However, computation is not the only aspect quantum computers can achieve meaningful advantages. Learning is yet another possibility, where the basic question is whether quantum computers (with resources such as *quantum memory* and *entangled measurements*) can help us learn certain properties of a physical system more efficiently. This kind of quantum advantages have been explored by several recent works [6–10] with positive examples including mixedness testing [6], unitarity testing [7], Pauli expectation values estimation [8], *etc.* However, all these quantum advantages proposals so far either focus on artificial problems or have no noise-resilient implementation. It is thus highly desirable to identify a practically-interesting learning task that can be used to demonstrate a robust quantum advantage in the NISQ era.

A particularly interesting type of learning tasks, known as *quantum benchmarking* [11], aims at characterizing the noise on a quantum device. This is yet another challenge for the NISQ era which is crucial to building better quantum hardware. Pauli noise is one of the most important quantum noise models: On the one hand, it can describe a wide range of incoherent noise, including dephasing, deplorizing, bit-flip, etc. On the other hand, the recently developed "randomized compiling" technique [12, 13] can tailor any noise model in a universal gate set into Pauli noise. Many benchmarking protocols also rely on twirling the noise into a Pauli channel before extracting any information [14–17]. Because of the important role played by Pauli channels, it is of great interest to study the estimation of these objects in an efficient and practical way. Despite a long line of research [18–26], the ultimate sample complexity for Pauli channel estimation has not yet been fully characterized.

In this work, we demonstrate an *exponential* quantum advantage for Pauli channel estimation. We show that, for the task of estimating the eigenvalues of an n-qubit Pauli channel (*i.e.*, *Pauli eigenvalues*) to additive error ε in l_∞ distance, there exists a measurement protocol assisted with an n-qubit ancilla that succeeds with high probability using $\mathcal{O}(n/\varepsilon^2)$ samples, while any ancilla-free protocol (possibly with adaptive control and channel concatenation) would require at least $\Omega(2^{n/3})$ rounds of measurements. As a byproduct, this provides a lower bound for randomized benchmarking (RB) [15, 27] based Pauli noise estimation protocol, resolving an open problem raised in [23].

We then study the sample efficiency advantages provided by a restricted amount of ancilla. While an ancilla larger than n qubits will not help further improve the efficiency, given a k-qubit ($0 \leq k \leq n$) ancilla, we obtain a lower bound of $\Omega(2^{(n-k)/3})$ for any non-concatenating protocol, and a stronger lower bound of $\Omega(n2^{n-k})$ for non-adaptive and non-concatenating protocols. The latter is shown to be tight by an explicitly constructed protocol (see Algorithm 1).

Finally, we show how to apply the ancilla-assisted estimation protocol in the practical task of benchmarking Pauli gates. Inspired by RB-type methods [15, 23, 27], we design a protocol that is both robust against state-preparation-and-measurement (SPAM) errors and exponentially more sample-efficient than any ancilla-free scheme, under reasonable noise assumptions. This proto-

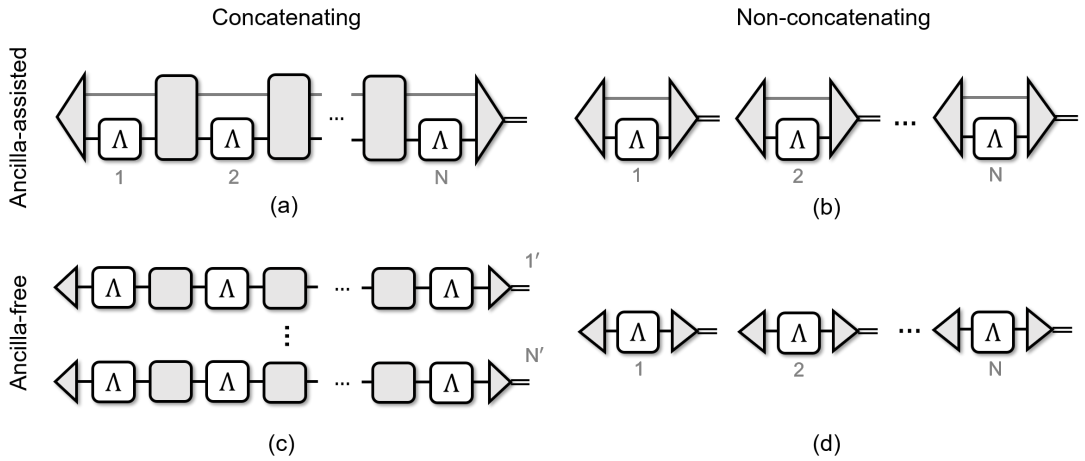

FIG. 1. Different measurement models on N copies of an unknown channel Λ. The gray triangles denote state preparation and measurement, and the gray boxes denote some known processing channels. (a) *Fully entangled measurement*: the most general way to measure a quantum channel where arbitrarily large entanglement and quantum memory is allowed; (b) *Ancilla-assisted non-concatenating measurement*: For each sample of Λ, one input some entangled state and conduct an entangled measurement. The measurement must be completely destructive, which means no quantum memory is allowed; (c) *Ancilla-free concatenating measurement*: One is allowed to sequally apply multiple copies of Λ and some processing channels before applying a single round of measurement, with no ancilla allowed. Here N' denotes the number of measurement rounds which is no larger than the number of samples; (d) *Un-entangled measurement*: Neither ancilla nor concatenation is allowed. Additionally, measurement models (b), (c), (d) can be either *adaptive* or *non-adaptive*. If the measurement setting at a certain round depends on previous measurement outcomes, it is an adaptive protocol. Otherwise, it is non-adaptive.

col can be used to experimentally demonstrate a robust and practical quantum advantage on NISQ devices.

II. PRELIMINARIES

A. Pauli group and Pauli channels

For an n-qubit Hilbert space, define P^n to be the Pauli group modulo the non-physical phase. P^n is an Abelian group isomorphic to $\mathbb{Z}_2^{2n}$, so we can use elements of $\mathbb{Z}_2^{2n}$ to uniquely label elements of P^n. Specifically, we view every $a \in \mathbb{Z}_2^{2n}$ as a $2n$-bit string $a = a_{x,1}a_{z,1}a_{x,2}a_{z,2}\cdots a_{x,n}a_{z,n}$ corresponding to the Pauli operator

$$P_a = \otimes_{k=1}^n i^{a_{x,k}a_{z,k}} X^{a_{x,k}} Z^{a_{z,k}}$$

where the phase is chosen to ensure Hermiticity. We also define a sympletic inner product $\langle\cdot,\cdot\rangle$ within $\mathbb{Z}_2^{2n}$ as

$$\langle a, b\rangle = \sum_{k=1}^{n} (a_{x,k}b_{z,k} + a_{z,k}b_{x,k}) \mod 2.$$

One can verify that $P_aP_b = (-1)^{\langle a,b\rangle}P_bP_a$ [28].

An n-qubit Pauli channel Λ is a quantum channel of the following form

$$\Lambda(\cdot) = \sum_{a\in\mathbb{Z}_2^{2n}} p_a P_a(\cdot)P_a, \tag{1}$$

where $\boldsymbol{p} := \{p_a\}_a$ is called the *Pauli error rates*. An important property of Pauli channels is that their eigenoperators are exactly the 4^n Pauli operators. Thus, an alternative expression for Λ is

$$\Lambda(\cdot) = \frac{1}{2^n}\sum_{b\in\mathbb{Z}_2^{2n}} \lambda_b \operatorname{Tr}(P_b(\cdot))P_b, \tag{2}$$

where $\boldsymbol{\lambda} := \{\lambda_b\}_b$ is called the *Pauli eigenvalues* [23, 25]. These two sets of parameters, $\boldsymbol{p}$ and $\boldsymbol{\lambda}$, are related by the Walsh-Hadamard transform

$$\lambda_b = \sum_{a\in\mathbb{Z}_2^{2n}} p_a(-1)^{\langle a,b\rangle}, \quad p_a = \frac{1}{4^n}\sum_{b\in\mathbb{Z}_2^{2n}} \lambda_b(-1)^{\langle a,b\rangle}. \tag{3}$$

Both $\boldsymbol{p}$ and $\boldsymbol{\lambda}$ are physically interesting parameters: The Pauli error rates are directly related to the error thresholds in fault-tolerant quantum computation [29, 30] and have been the quantities of interest for many quantum benchmarking protocols [16, 17, 23–25]; The Pauli eigenvalues quantify how well a Pauli observable is preserved through the noise channel (hence also known as *Pauli fidelities*) and have applications in quantum error mitigation (see *e.g.* [31]). In this work, we will focus on the estimation for $\boldsymbol{\lambda}$.

B. Comparison of different measurement models

To study the advantages provided by different kinds of quantum resources, we categorize measurement strategies

into ancilla-assisted *vs.* ancilla-free, concatenating *vs.* non-concatenating, and adaptive *vs.* non-adaptive measurements. See Fig. 1 and explanations in the captions. We remark that, recent works on quantum advantages in property learning [7, 8] have been focusing on the difference between the full-fledged entangled measurement Fig. 1(a) and the un-entangled measurement Fig. 1(d). Here, we introduce two intermediate measurement models Fig. 1(b) and Fig. 1(c) to separately study the role of ancilla and concatenation. There are also practical reasons to care about those intermediate models. As an example, the ancilla-free concatenating measurement model Fig. 1(c) exactly describes most existing RB protocols (see *e.g.* [32]), including the RB-type Pauli channel estimation protocol in Ref. [23]. Indeed, we show that it is the ancilla that provides an exponential advantage in sample complexity for Pauli channel estimation, while concatenation does not help improve the sample efficiency. These results further advances our understanding of quantum advantages in property learning [6–9].

Regarding these measurement models, one further question to ask is whether a restricted amount of ancilla can provide any sample efficiency advantages. In this work, we will characterize the advantages provided by a k-qubit ancilla ($0 \leq k \leq n$) for the non-concatenating measurement models. This kind of quantitative trade-off relations between quantum resources and sample efficiency has not been explored in previous works [6–9] and may potentially lead to a new resource-theoretical interpretation of quantum entanglement [33].

III. UPPER BOUNDS

Our goal is to estimate the Pauli eigenvalues $\boldsymbol{\lambda}$ of an n-qubit Pauli channel Λ to ε precision in l_∞ distance, *i.e.*, estimating each λ_a to ε additive error.

When an n-qubit ancillary system is available, a simple protocol is as follows: prepare n Bell pairs, input one qubit from each pair to the Pauli channel, and apply a Bell measurement on the output. Since the Pauli channel can be viewed as randomly applying one of the 4^n Pauli operators P_a with probability p_a, and each P_a is mapped to a unique measurement outcome [34], we are effectively sampling from the probability distribution $\boldsymbol{p}$. The sample of $\boldsymbol{p}$ can then be used to construct an estimator for $\boldsymbol{\lambda}$ according to the Walsh-Hadamard transform. As shown in Theorem 1, this protocol has sample complexity $\mathcal{O}(n)$; on the other hand, when no ancilla is allowed, there is no way to sample from $\boldsymbol{p}$ and simultaneously estimate all elements of $\boldsymbol{\lambda}$ from a single measurement setting. Intuitively, this is what makes the task difficult.

In the following, we give a unified estimation protocol using k ancilla qubits for $0 \leq k \leq n$. Roughly speaking, we divide the n qubits of the Pauli channel into two disjoint subsystems containing k and $n-k$ qubits, respectively, and deal with them separately. For the first subsystem, we introduce a k-qubit ancilla, input k Bell pairs to both systems, and apply a Bell measurement; For the second subsystem, we will use *stabilizer states* input and *syndrome measurements* to be defined later. The measurement scheme is depicted in Fig. 2.

A rigorous description of the protocol is given in Algorithm 1. Here, $|\Psi_v\rangle$ represents the Bell states on the $2k$-qubit subsystem defined as

$$|\Psi_v\rangle = P_v \otimes I\,|\Psi^+\rangle\,, \quad |\Psi^+\rangle = \frac{1}{2^k}\sum_{i=0}^{2^k-1}|i\rangle\,|i\rangle\,. \tag{4}$$

A *stabilizer group* S on the $(n-k)$-qubit subsystem is a group of 2^{n-k} commuting Pauli operators. Mathematically, S can be viewed as an $(n-k)$-dimensional subspace of $\mathbb{Z}_2^{2(n-k)}$. The *stabilizer states* $|\phi_e^{\mathsf{S}}\rangle$ are the simultaneous eigenstates of all Pauli operators in S, which can be expressed as

$$|\phi_e^{\mathsf{S}}\rangle\langle\phi_e^{\mathsf{S}}| = \frac{1}{2^{n-k}}\sum_{s\in\mathsf{S}}(-1)^{\langle s,e\rangle}P_s, \tag{5}$$

for $e \in \mathsf{S}^\perp := \mathbb{Z}_2^{2(n-k)}/\mathsf{S}$ known as the *error syndrome*. Note that $\{|\phi_e^{\mathsf{S}}\rangle\}_e$ forms an orthonormal basis for the $(n-k)$-qubit subsystem. The *stabilizer covering* O is a set of stabilizer groups $\{\mathsf{S}_i\}_i$ such that every Pauli operator belongs to at least one S_i [23].

Algorithm 1 k-qubit-ancilla-assisted Pauli channel estimation

Input: (1) N copies of an n-qubit Pauli channel Λ. (2) A stabilizer covering of P^{n-k} denoted as O.
Output: Estimates $\widehat{\boldsymbol{\lambda}}$ for the Pauli eigenvalues of Λ.

$\widehat{\lambda}_a := 0$, $N_a := 0$ **for all** $a \in \mathbb{Z}_2^{2n}$.
for $\mathsf{S} \in \mathsf{O}$ **do**
  for $i = 1$ **to** $\lfloor N/|\mathsf{O}|\rfloor$ **do**
    Input $|\Psi_0\rangle \otimes |\phi_0^{\mathsf{S}}\rangle$ to $\mathbb{1}\otimes\Lambda$.
    Measure in basis $\{|\Psi_v\rangle \otimes |\phi_e^{\mathsf{S}}\rangle\}$ with outcomes v, e.
    for $u \in \mathbb{Z}_2^{2k}, s \in \mathsf{S}$ **do**
      $\widehat{\lambda}_{u\oplus s}$ += $(-1)^{\langle u,v\rangle+\langle s,e\rangle}$, $N_{u\oplus s}$ += 1.
$\widehat{\lambda}_a := \widehat{\lambda}_a/N_a$ **for all** $a \in \mathbb{Z}_2^{2n}$.
return $\widehat{\boldsymbol{\lambda}} := \{\widehat{\lambda}_a\}_a$.
(Note: $\oplus$ stands for string concatenation.)

Theorem 1. *Algorithm 1 gives an estimate $\widehat{\boldsymbol{\lambda}}$ for the Pauli eigenvalues $\boldsymbol{\lambda}$ of any n-qubit Pauli channels that satisfies*

$$|\widehat{\lambda}_a - \lambda_a| \leq \varepsilon, \quad \forall a \in \mathbb{Z}_2^{2n} \tag{6}$$

with success probability at least $1-\delta$, given the following number of samples

$$N = \mathcal{O}(|\mathsf{O}| \times n\varepsilon^{-2}\log\delta^{-1}). \tag{7}$$

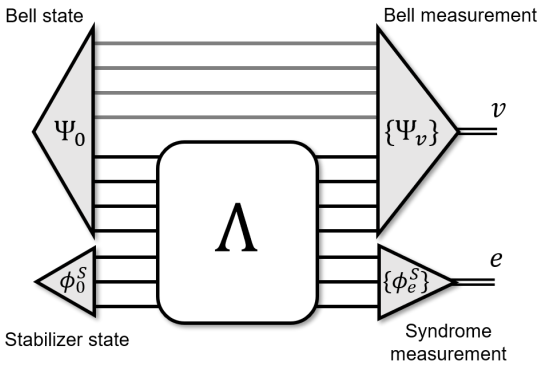

FIG. 2. A single round of measurement for the k-qubit-assisted Pauli channel estimation protocol in Algorithm 1. Here, a 4-qubit ancilla is used to estimate a 7-qubit Pauli channel.

Proof of Theorem 1. The probability distribution of measurement outcomes at Line 5 in Algorithm 1 can be calculated as

$$p(v,e) = \frac{1}{2^{n+k}} \sum_{u\in\mathbb{Z}_2^{2k}} \sum_{s\in\mathsf{S}} \lambda_{u\oplus s}(-1)^{\langle u,v\rangle}(-1)^{\langle s,e\rangle}. \quad (8)$$

See Appendix A for more details. Therefore, $p(v,e)$ and $\lambda_{u\oplus s}$ are related by the Walsh-Hadamard transform. Taking the inverse transform, we get

$$\lambda_{u\oplus s} = \sum_{v\in\mathbb{Z}_2^{2k}} \sum_{e\in\mathsf{S}^\perp} p(v,e)(-1)^{\langle u,v\rangle+\langle s,e\rangle}. \quad (9)$$

Thus, $(-1)^{\langle u,v\rangle+\langle s,e\rangle}$ is an un-biased estimator of $\lambda_{u\oplus s}$. According to Hoeffding's bound, $N_0 = \mathcal{O}(\varepsilon^{-2}\log\delta_0^{-1})$ samples are enough to estimate a single $\lambda_{u\oplus s}$ to additive precision ε with success probability at least $1-\delta_0$. Since every λ_a is covered by some stabilizer group $\mathsf{S}\in\mathsf{O}$, a total sample complexity of $N = \mathcal{O}(|\mathsf{O}|\times n\varepsilon^{-2}\log\delta^{-1})$ is enough to estimate all λ_a to additive error ε simultaneously with success probability at least $1-\delta$, by setting $\delta_0 := 4^{-n}\delta$ and applying the union bound. □

Corollary 2. *There exists a k-qubit-ancilla-assisted non-adaptive non-concatenating Pauli channel estimation protocol achieving $|\widehat{\lambda}_a - \lambda_a| \le \varepsilon$ for all $a\in\mathbb{Z}_2^{2n}$ with probability $1-\delta$ using $N = \mathcal{O}(n2^{n-k}\varepsilon^{-2}\log\delta^{-1})$ samples.*

Proof. This follows from the existence of a stabilizer covering for P^{n-k} of size $2^{n-k}+1$ [35], which in turn follows from the existence of $2^{n-k}+1$ mutually-unbiased bases for $(n-k)$-qubit systems [36]. □

We remark that, the choice of stabilizer covering in Corollary 2 gives the optimal sample complexity for all *non-adaptive* and *non-concatenating* Pauli eigenvalues estimation protocols, as proved in the next section. From a practical point of view, it involves $(n-k)$-qubit stabilizer states that might be difficult to prepare. A more experimental friendly version is to choose the stabilizer covering generated by all 3^{n-k} possible Pauli measurements, in which case only Pauli eigenstates preparation and Pauli measurements are required (on the $(n-k)$-qubit subsystem), at the expense of a sub-optimal sample complexity $N = \mathcal{O}(n3^{n-k}\varepsilon^{-2}\log\delta^{-1})$.

IV. LOWER BOUNDS

We have established a sample complexity upper bound of $\mathcal{O}(n2^{n-k})$ for non-adaptive non-concatenating k-qubit-ancilla-assisted Pauli eigenvalues estimation protocols, which implies a $\mathcal{O}(n)$ upper bound for the n-qubit ancilla case and a $\mathcal{O}(n2^n)$ upper bound for the ancilla-free case. In the following Theorem 3, we provide corresponding lower bounds to justify the exponential advantage provided by ancilla in this task.

Theorem 3. *For any estimation protocol that give an estimate $\widehat{\boldsymbol{\lambda}}$ for the Pauli eigenvalues $\boldsymbol{\lambda}$ of an arbitrary unknwon n-qubit Pauli channel Λ such that*

$$|\widehat{\lambda}_a - \lambda_a| \le 1/2, \quad \forall a\in\mathbb{Z}_2^{2n} \quad (10)$$

holds with high probability, the number of samples of Λ must satisfies (recall Fig. 1)

(A) $N = \Omega(n2^{n-k})$, for non-adaptive non-concatenating k-qubit-ancilla measurements.

(B) $N = \Omega(2^{(n-k)/3})$, for adaptive non-concatenating k-qubit-ancilla measurements.

(C) $N \ge N' = \Omega(2^{n/3})$, for adaptive concatenating ancilla-free measurements, where N' stands for the number of measurement rounds.

(D) $N = \Omega(n)$, for fully entangled measurements.

Indeed, Theorem 3 and Corollary 2 establishes an exponential advantage of ancilla-assisted measurements over ancilla-free measurements even with channel concatenation (as in the RB-type Pauli channel estimation protocols in Ref. [23]). Furthermore, for the non-concatenating cases, we see a roughly matching bounds for all ancilla size $0\le k\le n$, which can be interpreted as that a small number of ancilla ($k = o(n)$) would not help much in improving the sample efficiency. We also see from (D) that the sample complexity of Algorithm 1 with n ancilla qubits is optimal among all entangled strategies, thus we need not study protocols with more than n ancilla qubits.

Sketch of the proof. Our proof generalizes the techniques of Huang *et al.* [8] in proving lower bounds for learning Pauli expectation values of quantum states. The key is to construct the following set of Pauli channels

$$\left\{\Lambda_{(a,s)}(\cdot) = \frac{1}{2^n}\left(I\,\mathrm{Tr}(\cdot) + sP_a\,\mathrm{Tr}(P_a(\cdot))\right)\right\}_{a,s}, \quad (11)$$

for $a\in\{1,\cdots,4^n-1\}$ and $s=\pm1$. An estimation protocol satisfying the assumption of Theorem 3 is able to

identify an arbitrary element of this set using N copies of the channel. We can then use information-theoretical arguments to lower bound N for (A); The bounds in (B) and (C) are proved by reducing the learning problem to a channel discrimination problem between the completely-deplorizing channel and the channels in Eq. (11); To prove the bound in (D), we first use *teleportation stretching* [37, 38] to reduce any estimation protocols on N copies of the Pauli channel into a POVM measurement on N copies of their Choi states [39, 40], and then apply the Holevo's theorem [41]. See Appendix B for a full proof of Theorem 3. □

V. APPLICATIONS IN QUANTUM BENCHMARKING

In above, we have described an ancilla-assisted Pauli channel estimation protocol which provides exponential advantages over any ancilla-free protocols. Several issues need to be addressed before applying these protocols to a practical quantum noise characterization setting. Firstly, in most cases, we do not have a "black-box" access to the Pauli noise channels of interest. Instead, they are often attached with some applied quantum gates. We must consider the effect of such gates in our protocol. Secondly, the state preparation and measurement (SPAM) process would inevitably suffer from error. We would like to minimize the effect of such errors.

A recent progress by Flammia and Wallman [23] provides a way to address these issues using ideas from randomized benchmarking (see also [14, 42]). Their task is to benchmark the Pauli error rates of the Pauli gate set. By concatenating $m+1$ layers of random Pauli gates, they effectively obtain the mth power of the Pauli twirl for the noise channel (plus a single Pauli correction gate at the end), under a gate-independent, time-stationary, and Markovian (GTM) noise assumption. They then describe a protocol to estimate the quantity $A_a\lambda_a^m$ for all $a \in \mathbb{Z}_2^{2n}$, where $\boldsymbol{\lambda} \coloneqq \{\lambda_a\}_a$ is the Pauli eigenvalues of the noise channel of interest, and A_a is a SPAM related constant that is independent of m. By repeating this estimation procedure for different concatenating length m and applying a single-exponential fitting of $A_a\lambda_a^m$ for each $a \in \mathbb{Z}_2^{2n}$, one obtained an SPAM-robust estimation for $\boldsymbol{\lambda}$.

Importantly, the protocol in [23] is ancilla-free, which means an exponential number of measurements is necessary to approximate $\boldsymbol{\lambda}$ to small error in l_∞ distance, according to our Theorem 3 (C). In this section, we explain how the ancilla-assisted Pauli channel estimation protocol described above can be extended to this Pauli gates benchmarking setting, which is able to estimate $\boldsymbol{\lambda}$ exponentially more sample-efficiently as well as being SPAM-robust. The new protocol uses gate concatenation and single-exponential fitting, and can be viewed as a generalization of the methods in [23].

Let us start by defining the task and specifying our assumptions. The task is to characterize the noise of the n-qubit Pauli gate set. We assume the noise satisfies the GTM condition, which means every noisy implementation of Pauli gates can be written as

$$\widetilde{\mathcal{P}}_a = \mathcal{P}_a\Lambda_G, \quad a \in \mathbb{Z}_2^{2n}, \tag{12}$$

for an a-independent quantum channel Λ_G. The calligraphic $\mathcal{P}_a$ is the channel representation of the Pauli gate P_a, *i.e.*, $\mathcal{P}_a(\cdot) \coloneqq P_a(\cdot)P_a$. Our specific goal is to estimate the Pauli eigenvalues of the Pauli twirl of Λ_G, which is defined as

$$\Lambda \coloneqq \frac{1}{4^n}\sum_{a\in\mathbb{Z}_2^{2n}} \mathcal{P}_a\Lambda_G\mathcal{P}_a. \tag{13}$$

In addition, we assume there to be an n-qubit ancillary systems that can be entangled with the main system. The ancilla will basically be used as a quantum memory, see Fig. 3. A crucial assumption we need is that, the noise on the ancilla is negligible except for the entangled state preparation and measurement procedure. In other word, the noise channel on the ancilla is independent of the concatenating length on the main system. In practice, this requires (1) the crosstalk between the ancilla and the main system is negligible when applying gates only on the main system, and (2) the coherence time of the ancilla is much longer compared to the time of applying gates on the main system. We expect that these assumptions can be satisfied by *e.g.*, a near-term ion trap platform (see [43, 44]). Ion trap system typically has very long coherence time; Besides, after preparing the entangled state, one can shuttle the ions to separate the ancilla and the main system during gate applications and shuttle them back for the entangled measurement [44]. This step can minimize the crosstalk, and the errors introduced there can be viewed as SPAM error (independent of the concatenating length) so our protocol will be naturally robust against them. Techniques like dynamical decoupling [45, 46] can also make realistic devices better satisfy these assumptions.

The benchmarking protocol is described in Algorithm 2. In the following, we will show the correctness and give a rough analysis on the sample efficiency. A more rigorous analysis on the sample complexity, optimization of the concatenating length M and the number of repetitions R, and other aspects of the protocol are left for future research. We also remark that, for simplicity, we focus on the case where $k = n$ ancillary qubits are available. For a restricted number of ancillary qubits $0 < k < n$, one can also design a similar benchmarking protocol by hybridizing the $k = n$ protocol here and the $k = 0$ protocol in [23]. We omit the details about this hybrid protocol.

Theorem 4. *Given the aforementioned two assumptions about the noise model, the estimator given at Line 6 in Algorithm 2 satisfies*

$$\mathbb{E}\left[\widehat{F}_a^{(k)}(m)\right] = A_a\lambda_a^m, \quad \forall a \in \mathbb{Z}_2^{2n}, \tag{14}$$

where A_a, defined in Eq. (C1), is a noise-dependent constant that is independent of m.

Algorithm 2 SPAM-robust ancilla-assisted Pauli gate benchmarking

Input: (1) List of concatenating length M. (2) Number of Repetitions R. (3) Noisy implementation of Pauli gates $\widetilde{\mathcal{P}}_a = \mathcal{P}_a \Lambda_G$.
Output: SPAM-robust estimates $\widehat{\boldsymbol{\lambda}}$ for the Pauli eigenvalues of Λ as defined in Eq. (13).

for $m \in \mathsf{M}$ **do**
  for $k = 1$ **to** R **do**
    Prepare the (noisy) Bell state $|\widetilde{\Psi}^+\rangle$ between the ancillary system and the main system.
    Sequentially apply $m+1$ random (noisy) n-qubit Pauli gates $\{\widetilde{P}_{a_t}\}_{t=0}^m$ to the main system.
    Apply the (noisy) Bell measurement $\{|\widetilde{\Psi}_v\rangle\}_v$ with outcome v.
    $\widehat{F}_a^{(k)}(m) \coloneqq (-1)^{\langle a,v\rangle + \sum_{t=0}^m \langle a,a_t\rangle}$ **for all** $a \in \mathbb{Z}_2^{2n}$.
for $a \in \mathbb{Z}_2^{2n}$ **do**
  $\widehat{F}_a(m) \coloneqq \frac{1}{R}\sum_{k=1}^R \widehat{F}_a^{(k)}(m)$.
  Fit $\widehat{F}_a(m)$ to the single-exponential decay model $\widehat{A}_a \widehat{\lambda}_a^m$.
return $\widehat{\boldsymbol{\lambda}} \coloneqq \{\widehat{\lambda}_a\}_a$.

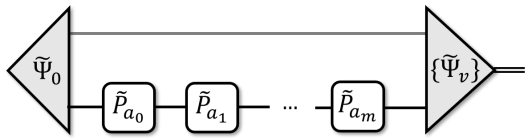

FIG. 3. A single round of measurement for the n-qubit ancilla-assisted Pauli gate benchmarking protocol. Here $\widetilde{\Psi}_0$ and $\{\widetilde{\Psi}_v\}$ stands for the (noisy) Bell states/measurements. $\{\widetilde{P}_{a_t}\}_{t=0}^m$ stands for a sequence of random (noisy) Pauli gates to be characterized.

The proof is given in Appendix C. Theorem 4 guarantees that, given sufficiently many concatenating length m and circuit samples R, Algorithm 2 can indeed converge to the true Pauli eigenvalues $\boldsymbol{\lambda}$, in a SPAM-error resilient manner. Since $\widehat{F}_a^{(k)}(m)$ takes value from $\{1,-1\}$, Hoeffding's bound says that a constant number of samples is enough to estimate its expectation to constant additive error with $1-o(1)$ success probability, for any specific a and m. By the union bound, $\mathcal{O}(n)$ samples are enough for this to hold for all $a \in \mathbb{Z}_2^{2n}$ simultaneously with high probability. If we further assume that the noise is weak, so that both A_a and λ_a are lower bounded by some constant, then a constant number of m and the above-achieved constant additive precision is enough for a small final estimation error for $\boldsymbol{\lambda}$ in l_∞ distance, which implies a total sample complexity of $\mathcal{O}(n)$. Therefore, Algorithm 2 is indeed exponentially more sample-efficient than any ancilla-free protocol [23] for this task. A more rigorous analysis about the sample complexity advantages under realistic SPAM noise models is left for future study.

VI. SUMMARY AND OUTLOOK

In this work, we show a provable quantum advantage provided by entangled measurements for a learning task [7, 8] of Pauli channel estimation, which is a practically useful tool urgently needed to characterize large quantum systems. For quantum benchmarking, our results provide fundamental efficiency limits for Pauli noise estimation, which partly solve an open problem raised in [23]. We also describe how the ancilla-assisted Pauli channel estimation protocol can be applied to a practical quantum benchmarking tasks in a noise-resilient and sample-efficient manner. Our results provide a promising tool for both characterizing near-term quantum devices and demonstrating quantum advantages in those systems

Several interesting questions remain to be explored in the future, including exploring the quantum advantages in learning other properties of Pauli channels (e.g., the Pauli error rates [47]); analyzing the sample complexity for learning Pauli channels with more specific structures; and analyzing the experimental performance of our algorithms in comparison to other ancilla-free protocols [13, 16].

ACKNOWLEDGMENTS

We acknowledge support from the ARO (W911NF-18-1-0020, W911NF-18-1-0212), ARO MURI (W911NF-16-1-0349), AFOSR MURI (FA9550-19-1-0399, FA9550-21-1-0209), DoE Q-NEXT, NSF (EFMA-1640959, OMA-1936118, EEC-1941583), NTT Research, and the Packard Foundation (2013-39273). S.Z. acknowledges funding provided by the Institute for Quantum Information and Matter, an NSF Physics Frontiers Center (NSF Grant PHY-1733907). A.S. is supported by a Chicago Prize Postdoctoral Fellowship in Theoretical Quantum Science.

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

[2] A. W. Harrow and A. Montanaro, Quantum computational supremacy, Nature **549**, 203 (2017).

[3] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell, *et al.*, Quantum supremacy using a programmable superconducting processor, Nature **574**, 505 (2019).

[4] H.-S. Zhong, H. Wang, Y.-H. Deng, M.-C. Chen, L.-C. Peng, Y.-H. Luo, J. Qin, D. Wu, X. Ding, Y. Hu, *et al.*, Quantum computational advantage using photons, Science **370**, 1460 (2020).

[5] Y. Wu, W.-S. Bao, S. Cao, F. Chen, M.-C. Chen, X. Chen, T.-H. Chung, H. Deng, Y. Du, D. Fan, *et al.*, Strong quantum computational advantage using a superconducting quantum processor, arXiv preprint arXiv:2106.14734 (2021).

[6] S. Bubeck, S. Chen, and J. Li, Entanglement is necessary for optimal quantum property testing, in *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE, 2020) pp. 692–703.

[7] D. Aharonov, J. Cotler, and X.-L. Qi, Quantum algorithmic measurement, arXiv preprint arXiv:2101.04634 (2021).

[8] H.-Y. Huang, R. Kueng, and J. Preskill, Information-theoretic bounds on quantum advantage in machine learning, Phys. Rev. Lett. **126**, 190505 (2021).

[9] S. Chen, J. Li, and R. O'Donnell, Toward instance-optimal state certification with incoherent measurements, arXiv preprint arXiv:2102.13098 (2021).

[10] Z. M. Rossi, J. Yu, I. L. Chuang, and S. Sugiura, Quantum advantage for noisy channel discrimination, arXiv preprint arXiv:2105.08707 (2021).

[11] J. Eisert, D. Hangleiter, N. Walk, I. Roth, D. Markham, R. Parekh, U. Chabaud, and E. Kashefi, Quantum certification and benchmarking, Nature Reviews Physics **2**, 382 (2020).

[12] J. J. Wallman and J. Emerson, Noise tailoring for scalable quantum computation via randomized compiling, Physical Review A **94**, 052325 (2016).

[13] A. Hashim, R. K. Naik, A. Morvan, J.-L. Ville, B. Mitchell, J. M. Kreikebaum, M. Davis, E. Smith, C. Iancu, K. P. O'Brien, *et al.*, Randomized compiling for scalable quantum computing on a noisy superconducting quantum processor, arXiv preprint arXiv:2010.00215 (2020).

[14] A. Erhard, J. J. Wallman, L. Postler, M. Meth, R. Stricker, E. A. Martinez, P. Schindler, T. Monz, J. Emerson, and R. Blatt, Characterizing large-scale quantum computers via cycle benchmarking, Nature communications **10**, 1 (2019).

[15] E. Magesan, J. M. Gambetta, and J. Emerson, Scalable and robust randomized benchmarking of quantum processes, Physical review letters **106**, 180504 (2011).

[16] R. Harper, S. T. Flammia, and J. J. Wallman, Efficient learning of quantum noise, Nature Physics **16**, 1184 (2020).

[17] Y. Liu, M. Otten, R. Bassirianjahromi, L. Jiang, and B. Fefferman, Benchmarking near-term quantum computers via random circuit sampling, arXiv preprint arXiv:2105.05232 (2021).

[18] A. Fujiwara and H. Imai, Quantum parameter estimation of a generalized pauli channel, Journal of Physics A: Mathematical and General **36**, 8093 (2003).

[19] M. Hayashi, Quantum channel estimation and asymptotic bound, in *Journal of Physics: Conference Series*, Vol. 233 (IOP Publishing, 2010) p. 012016.

[20] A. Chiuri, V. Rosati, G. Vallone, S. Pádua, H. Imai, S. Giacomini, C. Macchiavello, and P. Mataloni, Experimental realization of optimal noise estimation for a general pauli channel, Physical review letters **107**, 253602 (2011).

[21] L. Ruppert, D. Virosztek, and K. Hangos, Optimal parameter estimation of pauli channels, Journal of Physics A: Mathematical and Theoretical **45**, 265305 (2012).

[22] D. Collins, Mixed-state pauli-channel parameter estimation, Physical Review A **87**, 032301 (2013).

[23] S. T. Flammia and J. J. Wallman, Efficient estimation of pauli channels, ACM Transactions on Quantum Computing **1**, 1 (2020).

[24] R. Harper, W. Yu, and S. T. Flammia, Fast estimation of sparse quantum noise, PRX Quantum **2**, 010322 (2021).

[25] S. T. Flammia and R. O'Donnell, Pauli error estimation via population recovery, Quantum **5**, 549 (2021).

[26] M. Mohseni, A. T. Rezakhani, and D. A. Lidar, Quantum-process tomography: Resource analysis of different strategies, Physical Review A **77**, 032322 (2008).

[27] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland, Randomized benchmarking of quantum gates, Physical Review A **77**, 012307 (2008).

[28] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition*, 10th ed. (Cambridge University Press, USA, 2011).

[29] P. W. Shor, Fault-tolerant quantum computation, in *Proceedings of 37th Conference on Foundations of Computer Science* (IEEE, 1996) pp. 56–65.

[30] D. Aharonov and M. Ben-Or, Fault-tolerant quantum computation with constant error rate, SIAM Journal on Computing 10.1137/S0097539799359385 (2008).

[31] S. Chen, W. Yu, P. Zeng, and S. T. Flammia, Robust shadow estimation, PRX Quantum **2**, 030348 (2021).

[32] J. Helsen, I. Roth, E. Onorati, A. H. Werner, and J. Eisert, A general framework for randomized benchmarking, arXiv preprint arXiv:2010.07974 (2020).

[33] E. Chitambar and G. Gour, Quantum resource theories, Reviews of Modern Physics **91**, 025001 (2019).

[34] This can be viewed as a consequence of the well-known *superdense coding* protocol [48].

[35] J. Lawrence, Č. Brukner, and A. Zeilinger, Mutually unbiased binary observable sets on n qubits, Physical Review A **65**, 032320 (2002).

[36] W. K. Wootters and B. D. Fields, Optimal state-determination by mutually unbiased measurements, Annals of Physics **191**, 363 (1989).

[37] S. Pirandola, R. Laurenza, C. Ottaviani, and L. Banchi, Fundamental limits of repeaterless quantum communications, Nature communications **8**, 1 (2017).

[38] S. Pirandola, R. Laurenza, C. Lupo, and J. L. Pereira, Fundamental limits to quantum channel discrimination, npj Quantum Information **5**, 1 (2019).

[39] M.-D. Choi, Completely positive linear maps on complex matrices, Linear algebra and its applications **10**, 285 (1975).

[40] A. Jamiołkowski, Linear transformations which preserve trace and positive semidefiniteness of operators, Reports on Mathematical Physics **3**, 275 (1972).

[41] A. S. Holevo, Bounds for the quantity of information transmitted by a quantum communication channel, Problemy Peredachi Informatsii **9**, 3 (1973).

[42] J. Helsen, X. Xue, L. M. Vandersypen, and S. Wehner, A new class of efficient randomized benchmarking protocols, npj Quantum Information **5**, 1 (2019).

[43] K. Wright, K. Beck, S. Debnath, J. Amini, Y. Nam, N. Grzesiak, J.-S. Chen, N. Pisenti, M. Chmielewski, C. Collins, *et al.*, Benchmarking an 11-qubit quantum computer, Nature communications **10**, 1 (2019).

[44] J. M. Pino, J. M. Dreiling, C. Figgatt, J. P. Gaebler, S. A. Moses, M. Allman, C. Baldwin, M. Foss-Feig, D. Hayes, K. Mayer, *et al.*, Demonstration of the trapped-ion quantum ccd computer architecture, Nature **592**, 209 (2021).

[45] L. Viola, E. Knill, and S. Lloyd, Dynamical decoupling of open quantum systems, Physical Review Letters **82**, 2417 (1999).

[46] J. Bylander, S. Gustavsson, F. Yan, F. Yoshihara, K. Harrabi, G. Fitch, D. G. Cory, Y. Nakamura, J.-S. Tsai, and W. D. Oliver, Noise spectroscopy through dynamical decoupling with a superconducting flux qubit, Nature Physics **7**, 565 (2011).

[47] It is shown in [25] that the Pauli error rates for an n-qubit Pauli channel can be estimated to small error in l_∞ distance using un-entangled measurements with $\mathcal{O}(\log n)$ samples, so there is no large advantages for entangled strategies. Still, one could expect a sample-efficiency advantage under other practically-interesting metric of estimation precision.

[48] C. H. Bennett and S. J. Wiesner, Communication via one-and two-particle operators on einstein-podolsky-rosen states, Physical review letters **69**, 2881 (1992).

[49] S. T. Flammia, D. Gross, Y.-K. Liu, and J. Eisert, Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators, New Journal of Physics **14**, 095022 (2012).

[50] J. Haah, A. W. Harrow, Z. Ji, X. Wu, and N. Yu, Sample-optimal tomography of quantum states, IEEE Transactions on Information Theory **63**, 5628 (2017).

[51] I. Roth, R. Kueng, S. Kimmel, Y.-K. Liu, D. Gross, J. Eisert, and M. Kliesch, Recovering quantum gates from few average gate fidelities, Physical review letters **121**, 170502 (2018).

[52] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, 2012).

[53] C. H. Bennett, D. P. DiVincenzo, J. A. Smolin, and W. K. Wootters, Mixed-state entanglement and quantum error correction, Physical Review A **54**, 3824 (1996).

[54] M. M. Wilde, *Quantum information theory* (Cambridge University Press, 2013).

Appendix A: Details about the proof of Theorem 1

In this section, we provide details in the derivation of Eq. (8), *i.e.*, the distribution of measurement outcomes at Line 5 in Algorithm 1. For clarity, denote the k-qubit Hilbert space of the ancilla as A, and divide the n-qubit Hilbert space of the main system into a k-qubit subspace B and an $(n-k)$-qubit subspace C. The input state to the channel is $|\Psi_0\rangle_{AB}\otimes|\phi_0^{\mathsf{S}}\rangle_C$, and the measurement basis is $\{|\Psi_v\rangle_{AB}\otimes|\phi_e^{\mathsf{S}}\rangle_C\}_{v,e}$. Here, $|\Psi_v\rangle_{AB}$ are Bell states and can be expressed as

$$\begin{aligned}|\Psi_v\rangle\langle\Psi_v| &\coloneqq (P_v\otimes I)|\Psi^+\rangle\langle\Psi^+|(P_v\otimes I)\\ &= \frac{1}{4^k}\sum_{u\in\mathbb{Z}_2^{2k}} P_vP_uP_v\otimes P_u^T\\ &= \frac{1}{4^k}\sum_{u\in\mathbb{Z}_2^{2k}} (-1)^{\langle u,v\rangle}P_u\otimes P_u^T,\end{aligned} \tag{A1}$$

And the stabilizer state $|\phi_e^{\mathsf{S}}\rangle_C$ is defined as

$$|\phi_e^{\mathsf{S}}\rangle\langle\phi_e^{\mathsf{S}}| \coloneqq \frac{1}{2^{n-k}}\sum_{s\in\mathsf{S}}(-1)^{\langle s,e\rangle}P_s, \tag{A2}$$

for $e\in\mathsf{S}^\perp\coloneqq\mathbb{Z}_2^{2(n-k)}/\mathsf{S}$. We remark that, the stabilizer state is well-defined for all $e\in\mathbb{Z}_2^{2(n-k)}$, but $|\phi_a^{\mathsf{S}}\rangle$ and $|\phi_b^{\mathsf{S}}\rangle$ represent the same state if $a+b\in\mathsf{S}$ (bitwise modulo 2 sum), so we only need to consider the quotient space of $\mathbb{Z}_2^{2(n-k)}$ over S. When calculating the sympletic inner product $\langle s,e\rangle$, one should understand e as an arbitrary representative of the coset it stands for.

Therefore, the measurement outcome distribution can be calculated as

$$\begin{aligned}p(v,e) &= \mathrm{Tr}\left((|\Psi_v\rangle\langle\Psi_v|_{AB}\otimes|\phi_e^{\mathsf{S}}\rangle\langle\phi_e^{\mathsf{S}}|_C)\mathbb{1}^A\otimes\Lambda^{BC}(|\Psi_0\rangle\langle\Psi_0|_{AB}\otimes|\phi_0^{\mathsf{S}}\rangle\langle\phi_0^{\mathsf{S}}|_C)\right)\\ &= \frac{1}{4^{n+k}}\mathrm{Tr}\left(\sum_{u,u'\in\mathbb{Z}_2^{2k}}\sum_{s,s'\in\mathsf{S}}\lambda_{u'\oplus s'}\left((-1)^{\langle u,v\rangle}P_u\otimes P_u^T\otimes(-1)^{\langle s,e\rangle}P_s\right)_{ABC}\left(P_{u'}\otimes P_{u'}^T\otimes P_{s'}\right)_{ABC}\right)\\ &= \frac{1}{2^{n+k}}\sum_{u\in\mathbb{Z}_2^{2k}}\sum_{s\in\mathsf{S}}\lambda_{u\oplus s}(-1)^{\langle u,v\rangle}(-1)^{\langle s,e\rangle},\end{aligned} \tag{A3}$$

which is exactly Eq. (8) in the main text. It is then obvious that $p(v,e)$ and $\lambda_{u\oplus s}$ are related by the Walsh-Hadamard transform (see [23, Lemma 4]).

Appendix B: Proof of the lower bounds

1. A tight lower bound for non-adaptive and non-concatenating strategies

In this section, we prove a matching lower bound of $\Omega(n2^{n-k})$ for all non-adaptive and non-concatenating k-qubit ancilla-assisted Pauli channel estimation protocols, which include the ancilla-free strategies ($k=0$) and n-qubit ancilla assisted strategies ($k=n$) as two special cases. Recall that, by a non-adaptive and non-concatenating protocol we mean that, for each sample of the Pauli channel Λ, we prepare an $n+k$ qubits state, input it to $\Lambda\otimes\mathbb{1}$, and apply a POVM measurement on the joint output state. The input state and measurement setting for the ith sample is not allowed to depend on previous measurement outcomes, nor do we allow concatenating multiple samples of Λ in a single round of measurement (which is used in RB-type Pauli error estimation protocols [23]).

Theorem 5. *For any non-adaptive, non-concatenating k-qubit ancilla-assisted protocols that give an estimate $\widehat{\boldsymbol{\lambda}}$ of the Pauli eigenvalues $\boldsymbol{\lambda}$ of an arbitrary unknown n-qubit Pauli channel such that*

$$|\widehat{\lambda}_a-\lambda_a|<\frac{1}{2},\quad \forall a\in\mathbb{Z}_2^{2n} \tag{B1}$$

holds with high probability, the number of samples of Λ required is at least $\Omega(n2^{n-k})$.

Our proof techniques generalize the information-theoretical arguments by Huang *et al.* [8], which in turn stem from previous work on sample complexity lower bounds for quantum tomography [49–51]. Consider a communication protocol between Alice and Bob where their share the following "codebook":

$$(a,s)\in\{1,\cdots,4^n-1\}\times\{\pm1\}\ \longrightarrow\ \Lambda_{(a,s)}(\cdot)=\frac{1}{2^n}\left(I\,\mathrm{Tr}(\cdot)+sP_a\,\mathrm{Tr}(P_a(\cdot))\right). \tag{B2}$$

Now, Alice picks at uniform random one out of the $2(4^n-1)$ possible pairs of (a,s) and then send N copies of the channel $\Lambda_{(a,s)}$ to Bob. If there exists a Pauli channel estimation protocol using N samples and satisfying the assumption of Theorem 5, Bob can use that protocol to uniquely determine Alice's choice of (a,s) with high probability, since the Pauli eigenvalues of any $\Lambda_{(a,s)}$ only take values from $\{-1,0,+1\}$. Suppose Bob's input state and POVM outcome for the ith sample is $\{\rho_i,E_i\}$. According to Fano's inequality, the mutual information between the random variable pair (a,s) and Bob's measurement results has the following lower bound

$$I\left((a,s):\{\rho_1,E_1\},...,\{\rho_N,E_N\}\right)\geq\Omega(\log(2(4^n-1)))=\Omega(n). \tag{B3}$$

We also know by assumption that the measurement outcomes $\{\rho_i,E_i\}$ are independent from each other, conditioned on (a,s). The chain rule of mutual information then gives that

$$\sum_{i=1}^{N}I\left((a,s),\{\rho_i,E_i\}\right)=I((a,s):\{\rho_1,E_1\},...,\{\rho_N,E_N\})\geq\Omega(n). \tag{B4}$$

We will show that $I\left((a,s):\{\rho_i,E_i\}\right)\leq\mathcal{O}(2^{k-n})$ in the following lemma. This would then give us the desired sample complexity lower bound $N\geq\Omega(n2^{n-k})$, which completes the proof of Theorem 5.

Lemma 6. $I\left((a,s):\{\rho_i,E_i\}\right)\leq\dfrac{2^k}{2^n-1}$.

Proof. First notice that it suffices to consider pure state input and rank-1 POVM measurements. The latter comes from the fact that every POVM measurement can be viewed as a coarse-graining of some rank-1 POVM measurement. To see the former, consider the underlying distribution $p((a,s),\{\rho_i,E_i\})=p(a,s)p(\{E_i,\rho_i\}|a,s)$. The mutual information $I((a,s):\{\rho_i,E_i\})$ is convex about $p(\{E_i,\rho_i\}|a,s)$ when fixing $p(a,s)$ (see *e.g.*, [52, Theorem 2.7.4]), thus using mixed state input can never provide a larger mutual information.

Thanks to this observation, we can without loss of generality let the input state be $|A\rangle$ and let the POVM measurement be $\{w_j2^{n+k}|B_j\rangle\langle B_j|\}_j$, where $|A\rangle,|B_j\rangle\in\mathbb{C}^{2^n\times2^k}$ are unit vectors, and $\sum_j w_j=1$ by normalization. We also abuse notations a little bit to let A and B_j denote the $2^n\times2^k$ matrices that satisfy

$$|A\rangle=\sum_{p=0}^{2^n-1}\sum_{q=0}^{2^k-1}\langle p|A|q\rangle\,|p\rangle\,|q\rangle\,,\quad |B_j\rangle=\sum_{p=0}^{2^n-1}\sum_{q=0}^{2^k-1}\langle p|B_j|q\rangle\,|p\rangle\,|q\rangle\,, \tag{B5}$$

where $\{|p\rangle\}$ and $\{|q\rangle\}$ are computational basis states. The normalization condition of $|A\rangle$ and $|B_j\rangle$ is equivalent to

$$\mathrm{Tr}(A^\dagger A)=\mathrm{Tr}(B_j^\dagger B_j)=1. \tag{B6}$$

We also define $C_j\coloneqq B_jA^\dagger$ which is a $2^n\times2^n$ matrix of rank less or equal to 2^k.

The mutual information between (a,s) and a single round of measurement outcome j can be upper bounded as

$$\begin{aligned}I((a,s):j)&=H(j)-H(j|a,s)\\&=-\sum_j\left(\mathop{\mathbb{E}}_{(a,s)}p(j|a,s)\right)\log\left(\mathop{\mathbb{E}}_{(a,s)}p(j|a,s)\right)+\mathop{\mathbb{E}}_{(a,s)}\sum_j p(j|a,s)\log p(j|a,s)\\&\leq\sum_j\frac{\mathbb{E}_{(a,s)}[p(j|a,s)^2]-\mathbb{E}_{(a,s)}[p(j|a,s)]^2}{\mathbb{E}_{(a,s)}[p(j|a,s)]},\end{aligned} \tag{B7}$$

where the inequality follows from the fact that $\log(x)\leq\log(y)+\frac{x-y}{y}$ in which we take $x\coloneqq p(j|a,s)$ and $y\coloneqq\mathbb{E}_{(a,s)}[p(j|a,s)]$.

The conditional probability $p(j|a,s)$ can be calculated as

$$\begin{aligned} p(j|a,s) &= w_j 2^{n+k} \langle B_j| \Lambda_{(a,s)} \otimes \mathbb{1}(|A\rangle\langle A|) |B_j\rangle \\ &= w_j \sum_{b=0}^{4^k} \left(\langle B_j|I \otimes P_b|B_j\rangle \langle A|I \otimes P_b|A\rangle + s\langle B_j|P_a \otimes P_b|B_j\rangle \langle A|P_a \otimes P_b|A\rangle \right) \\ &= w_j \sum_{b=0}^{4^k} \left(\mathrm{Tr}(B_j^\dagger B_j P_b^T) \,\mathrm{Tr}(A^\dagger A P_b^T) + s \,\mathrm{Tr}(B_j^\dagger P_a B_j P_b^T) \,\mathrm{Tr}(A^\dagger P_a A P_b^T) \right) \\ &= w_j 2^k \left(\mathrm{Tr}(B_j^\dagger B_j A^\dagger A) + s \,\mathrm{Tr}(B_j^\dagger P_a B_j A^\dagger P_a A) \right) \\ &= w_j 2^k \left(\mathrm{Tr}(C_j^\dagger C_j) + s \,\mathrm{Tr}(P_a C_j P_a C_j^\dagger) \right), \end{aligned} \tag{B8}$$

where we expand the identity channel as $\mathbb{1}(\cdot) = 2^{-k} \sum_{b=0}^{4^k} P_b \,\mathrm{Tr}(P_b(\cdot))$ in the second line, and use the fact $2^{-k} \sum_{b=0}^{4^k} P_b \otimes P_b$ equals to the swap operator in the fourth line.

The average value and second moment of $p(j|a,s)$ according to the distribution of (a,s) are

$$\begin{aligned} &\mathop{\mathbb{E}}_{(a,s)} [p(j|a,s)] = w_j 2^k \,\mathrm{Tr}(C_j^\dagger C_j), \\ &\mathop{\mathbb{E}}_{(a,s)} [p(j|a,s)^2] = w_j^2 4^k \left(\mathrm{Tr}^2(C_j^\dagger C_j) + \frac{1}{4^n - 1} \sum_{a=1}^{4^n-1} \mathrm{Tr}^2(P_a C_j P_a C_j^\dagger) \right). \end{aligned} \tag{B9}$$

Hence we have the following bound for the mutual information

$$I\left((a,s) : j\right) \leq \sum_j w_j 2^k \,\mathrm{Tr}(C_j^\dagger C_j) \left(\frac{1}{4^n - 1} \sum_{a=1}^{4^n-1} \frac{\mathrm{Tr}(P_a C_j^\dagger P_a C_j)^2}{\mathrm{Tr}(C_j^\dagger C_j)^2} \right). \tag{B10}$$

Now we further calculate the R.H.S. of the above inequality. Let $M = P_a C_j^\dagger P_a C_j$. Notice that

$$\mathrm{rank}(M) \leq \mathrm{rank}(C_j) \leq 2^k, \tag{B11}$$

which means there exists a rank-2^k projector Π such that $\mathrm{Tr}(M) = \mathrm{Tr}(M\Pi)$. According to Cauchy-Schwarz inequality,

$$\begin{aligned} \mathrm{Tr}(M)^2 &= \mathrm{Tr}(M\Pi)^2 \\ &\leq \mathrm{Tr}(MM^\dagger) \,\mathrm{Tr}(\Pi^\dagger \Pi) \\ &= \mathrm{Tr}(C_j C_j^\dagger P_a C_j C_j^\dagger P_a) \times 2^k \end{aligned} \tag{B12}$$

Substitute this into Eq. (B10),

$$\begin{aligned} I((a,s) : j) &\leq \sum_j w_j 2^k \,\mathrm{Tr}(C_j^\dagger C_j) \left(\frac{2^k}{4^n - 1} \sum_{a=1}^{4^n-1} \frac{\mathrm{Tr}(C_j C_j^\dagger P_a C_j C_j^\dagger P_a)}{\mathrm{Tr}(C_j^\dagger C_j)^2} \right) \\ &= \sum_j w_j 2^k \,\mathrm{Tr}(C_j^\dagger C_j) \left(\frac{2^k}{4^n - 1} \times \frac{\sum_{a=0}^{4^n-1} \mathrm{Tr}(C_j C_j^\dagger P_a C_j C_j^\dagger P_a) - \mathrm{Tr}(C_j C_j^\dagger C_j C_j^\dagger)}{\mathrm{Tr}(C_j^\dagger C_j)^2} \right) \\ &= \sum_j w_j 2^k \,\mathrm{Tr}(C_j^\dagger C_j) \left(\frac{2^k}{4^n - 1} \times \frac{2^n \,\mathrm{Tr}(C_j^\dagger C_j)^2 - \mathrm{Tr}(C_j C_j^\dagger C_j C_j^\dagger)}{\mathrm{Tr}(C_j^\dagger C_j)^2} \right) \\ &\leq \sum_j w_j 2^k \,\mathrm{Tr}(C_j^\dagger C_j) \frac{2^k}{2^n - 1} \\ &= \frac{2^k}{2^n - 1}, \end{aligned} \tag{B13}$$

where the third line uses the following formula of Pauli twirling,

$$\frac{1}{4^n}\sum_{a=0}^{4^n} P_a X P_a = \frac{1}{2^n}\operatorname{Tr}(X) I, \tag{B14}$$

and the last line follows from the fact that

$$\sum_j w_j 2^k \operatorname{Tr}(C_j^\dagger C_j) = \sum_j \mathop{\mathbb{E}}_{(a,s)} [p(j|a,s)] = 1. \tag{B15}$$

This completes the proof of Lemma 6. □

2. A lower bound for adaptive but non-concatenating strategies

In this section, we prove a (perhaps loose) lower bound of $\Omega(2^{(n-k)/3})$ for all adaptive and non-concatenating k-qubit ancilla-assisted Pauli channel estimation protocols, as stated in the following theorem.

Theorem 7. *For any adaptive, non-concatenating k-qubit ancilla-assisted protocols that give an estimate $\widehat{\boldsymbol{\lambda}}$ of the Pauli eigenvalues $\boldsymbol{\lambda}$ of an arbitrary unknown n-qubit Pauli channel such that*

$$|\widehat{\lambda}_a - \lambda_a| < \frac{1}{2}, \quad \forall a \in \mathbb{Z}_2^{2n} \tag{B16}$$

holds with high probability, the number of samples of Λ required is at least $\Omega(2^{(n-k)/3})$.

Our proof techniques generalize the methods of Huang *et al.* [8] for proving adaptive sample complexity lower bounds for Pauli expectation values estimation of unknown quantum states. Consider the following 4^n possible Pauli channels

$$\begin{cases} \Lambda_{\text{dep}}(\cdot) = \dfrac{1}{2^n} I \operatorname{Tr}(\cdot), \\ \Lambda_a(\cdot) = \dfrac{1}{2^n}\left(I \operatorname{Tr}(\cdot) + P_a \operatorname{Tr}(P_a(\cdot))\right), \ \forall a \in \{1, \cdots, 4^n - 1\}. \end{cases} \tag{B17}$$

Here Λ_{dep} is known as the completely deplorizing channel. If there exists an Pauli channel estimation protocol satisfying the requirement of Theorem 7, one can unambiguously identify each one of the 4^n possible Pauli channels appearing above with high probability, given sufficient number of samples of Λ. This in turn implies that one should be able to distinguish the following two equal-probable hypotheses with high success probability.

1. Given N copies of $\Lambda = \Lambda_{\text{dep}}$.
2. Given N copies of $\Lambda = \Lambda_a$ for a uniformly-randomly picked $a \in \{1, \cdots, 4^n - 1\}$.

For an adaptive but non-concatenating protocol, one has to choose an $2^n \times 2^k$ dimensional input state and a POVM measurement for the ith sample of Λ, where the choice may depend on previous measurement outcomes. Denote the measurement outcome of the ith round as o_i. We explicitly write the state and measurement at the ith round as $\rho^{o_{<i}}$ and $\{E_j^{o_{<i}}\}_j$ to emphasize their dependence on $o_{<i} \coloneqq [o_1, ..., o_{i-1}]$.Denote the measurement outcomes among all the N samples as $o_{1:N} \coloneqq [o_1, \cdots, o_N]$. The probability distribution of $o_{1:N}$ under the above two hypothesis can be expressed as

$$\begin{cases} \text{Hypothesis 1:} \quad p_1(o_{1:N}) = \displaystyle\prod_{i=1}^{N} \operatorname{Tr}\left(E_{o_i}^{o_{<i}} \Lambda_{\text{dep}} \otimes \mathbb{1}(\rho^{o_{<i}})\right), \\ \text{Hypothesis 2:} \quad p_2(o_{1:N}) = \displaystyle\mathop{\mathbb{E}}_{a \neq 0} \prod_{i=1}^{N} \operatorname{Tr}\left(E_{o_i}^{o_{<i}} \Lambda_a \otimes \mathbb{1}(\rho^{o_{<i}})\right). \end{cases} \tag{B18}$$

The ability to distinguish these two hypotheses is equivalent to the ability to distinguish p_1 from p_2. The maximal success probability of distinguishing two probability distributions is given by $\frac{1}{2}(1 + \operatorname{TV}(p_1, p_2))$ where TV stands for the total variance distance defined as follows

$$\operatorname{TV}(p_1, p_2) \coloneqq \sum_{\substack{o_{1:N} \text{ s.t.} \\ p_1(o_{1:N}) \geq p_2(o_{1:N})}} (p_1(o_{1:N}) - p_2(o_{1:N})). \tag{B19}$$

We will show in the following Lemma that $\mathrm{TV}(p_1,p_2)=\mathcal{O}(N2^{(k-n)/3})$, which immediately implies that one must have $N=\Omega(2^{(n-k)/3})$ in order to obtain high success probability in distinguishing p_1 from p_2. This then completes the proof of Theorem 7.

Lemma 8. $\mathrm{TV}(p_1,p_2)\leq 2N\left(\frac{2^k}{2^n-1}\right)^{1/3}$.

Proof. First notice that it suffices to consider pure state input and rank-1 POVM measurement. For the former, the probability distribution obtained from mixed state input can be viewed as a convex combination of distributions obtained from pure state input. Thanks to the joint convexity, mixed state input can not yield a larger total variance distance; For the latter, every POVM measurement can be viewed as a coarse-graining of some rank-1 POVM measurement. Because of the data-processing property, this coarse-graining would not yield a larger total variance distance.

In light of this observation, we can without loss of generality let the input state at the ith round be $|A^{o_{<i}}\rangle$ and let the POVM measurement be $\{w_{o_i}^{o_{<i}}2^{n+k}|B_{o_i}^{o_{<i}}\rangle\langle B_{o_i}^{o_{<i}}|\}_{o_i}$, conditioned on previous measurement outcomes $o_{<i}$, where $|A^{o_{<i}}\rangle, |B_{o_i}^{o_{<i}}\rangle\in\mathbb{C}^{2^n\times 2^k}$ are unit vectors, and $\sum_{o_i}w_{o_i}^{o_{<i}}=1$ by normalization. We also introduce the two $2^n\times 2^k$ matrices $A^{o_{<i}}$, $B_{o_i}^{o_{<i}}$ defined similarly as in Eq. (B5), and define $C_{o_i}^{o_{<i}}:=B_{o_i}^{o_{<i}}A^{o_{<i}\dagger}$ which is a $2^n\times 2^n$ matrix of rank less or equal to 2^k. With the above definitions, one can verify that p_1 and p_2 can be expressed as follows

$$\begin{aligned}
p_1(o_{1:N}) &= \prod_{i=1}^{N} w_{o_i}^{o_{<i}}2^{n+k}\left\langle B_{o_i}^{o_{<i}}\right|\Lambda_{\mathrm{dep}}\otimes\mathbb{1}\left(\left(|A^{o_{<i}}\rangle\langle A^{o_{<i}}|\right)\left|B_{o_i}^{o_{<i}}\right\rangle\right),\\
&= \prod_{i=1}^{N} w_{o_i}^{o_{<i}}2^{k}\,\mathrm{Tr}\left(C_{o_i}^{o_{<i}\dagger}C_{o_i}^{o_{<i}}\right)\\
p_2(o_{1:N}) &= \mathop{\mathbb{E}}_{a\neq 0}\prod_{i=1}^{N} w_{o_i}^{o_{<i}}2^{n+k}\left\langle B_{o_i}^{o_{<i}}\right|\Lambda_a\otimes\mathbb{1}\left(\left(|A^{o_{<i}}\rangle\langle A^{o_{<i}}|\right)\left|B_{o_i}^{o_{<i}}\right\rangle\right)\\
&= \mathop{\mathbb{E}}_{a\neq 0}\prod_{i=1}^{N} w_{o_i}^{o_{<i}}2^{k}\left(\mathrm{Tr}\left(C_{o_i}^{o_{<i}\dagger}C_{o_i}^{o_{<i}}\right)+\mathrm{Tr}\left(C_{o_i}^{o_{<i}\dagger}P_aC_{o_i}^{o_{<i}}P_a\right)\right).
\end{aligned}\tag{B20}$$

The total variance between p_1 and p_2 can then be bounded as

$$\begin{aligned}
\mathrm{TV}(p_1,p_2) &= \mathop{\mathbb{E}}_{a\neq 0}\sum_{\substack{o_{1:N}\ \text{s.t.}\\ p_1(o_{1:N})\geq p_2(o_{1:N})}}\left(\prod_{i=1}^{N} w_{o_i}^{o_{<i}}2^{k}\,\mathrm{Tr}\left(C_{o_i}^{o_{<i}\dagger}C_{o_i}^{o_{<i}}\right)\right)\left(1-\prod_{i=1}^{N}\left(1+\frac{\mathrm{Tr}\left(C_{o_i}^{o_{<i}\dagger}P_aC_{o_i}^{o_{<i}}P_a\right)}{\mathrm{Tr}\left(C_{o_i}^{o_{<i}\dagger}C_{o_i}^{o_{<i}}\right)}\right)\right)\\
&= \mathop{\mathbb{E}}_{a\neq 0}\sum_{\substack{o_{1:N}\ \text{s.t.}\\ p_1(o_{1:N})\geq p_2(o_{1:N})}} p_1(o_{1:N})\left(1-\prod_{i=1}^{N}\left(1+\frac{\mathrm{Tr}\left(C_{o_i}^{o_{<i}\dagger}P_aC_{o_i}^{o_{<i}}P_a\right)}{\mathrm{Tr}\left(C_{o_i}^{o_{<i}\dagger}C_{o_i}^{o_{<i}}\right)}\right)\right),
\end{aligned}\tag{B21}$$

In order to bound the R.H.S., we make use of a technique from Huang *et al.* [8]. Let C denote an arbitrary $2^n\times 2^n$ complex matrix of rank no more than 2^k, consider the following subset of n-qubit Pauli operators

$$G:=\left\{a\in\{1,\cdots,4^n-1\}\ :\ \left|\frac{\mathrm{Tr}\left(C^\dagger P_aCP_a\right)}{\mathrm{Tr}\left(C^\dagger C\right)}\right|\leq\left(\frac{2^k}{2^n-1}\right)^{1/3}\right\}.\tag{B22}$$

We claim that the size of G satisfies

$$|G|\geq\left(1-\left(\frac{2^k}{2^n-1}\right)^{1/3}\right)(4^n-1),\tag{B23}$$

which can be shown by contradiction: Suppose this does not hold, we would have

$$\begin{aligned}
\sum_{a=1}^{4^n-1}\left(\frac{\mathrm{Tr}\left(C^\dagger P_aCP_a\right)}{\mathrm{Tr}\left(C^\dagger C\right)}\right)^2 &\geq \left(\frac{2^k}{2^n-1}\right)^{2/3}\times(4^n-1-|G|)\\
&> \left(\frac{2^k}{2^n-1}\right)^{2/3}\times\left(\frac{2^k}{2^n-1}\right)^{1/3}(4^n-1)\\
&= 2^k(2^n+1).
\end{aligned}\tag{B24}$$

However, the L.H.S. of the above can be upper bounded as

$$\begin{aligned}\sum_{a=1}^{4^n-1}\left(\frac{\mathrm{Tr}\left(C^\dagger P_a C P_a\right)}{\mathrm{Tr}\left(C^\dagger C\right)}\right)^2 &\le \sum_{a=1}^{4^n-1}\frac{2^k\,\mathrm{Tr}(CC^\dagger P_a CC^\dagger P_a)}{\mathrm{Tr}^2\left(C^\dagger C\right)}\\ &= 2^k\frac{2^n\,\mathrm{Tr}^2\left(C^\dagger C\right)-\mathrm{Tr}(C^\dagger CC^\dagger C)}{\mathrm{Tr}^2\left(C^\dagger C\right)}\\ &\le 2^k 2^n,\end{aligned} \tag{B25}$$

where the first inequality follows from Cauchy-Schwarz (see Eq. (B12)), and the first equality evaluates the Pauli twirling (see Eq. (B13)). This leads to contradition, and hence proves the desired lower bound on $|G|$.

Now, we define another subset of n-qubit Pauli operators, conditioned on the measurement outcomes $o_{1:N}$, as follows

$$G^{(o_{1:N})} \coloneqq \left\{ a\in\{1,\cdots,4^n-1\} \;:\; \left|\frac{\mathrm{Tr}\left(C_{o_i}^{o_{<i}\dagger}P_a C_{o_i}^{o_{<i}}P_a\right)}{\mathrm{Tr}\left(C_{o_i}^{o_{<i}\dagger}C_{o_i}^{o_{<i}}\right)}\right| \le \left(\frac{2^k}{2^n-1}\right)^{1/3},\ \forall i=1,\cdots,N\right\}. \tag{B26}$$

By applying a "union bound" on Eq. (B22), we immediately have the following lower bound on the size of $G^{(o_{1:N})}$:

$$\left|G^{(o_{1:N})}\right| \ge \left(1-N\left(\frac{2^k}{2^n-1}\right)^{1/3}\right)(4^n-1). \tag{B27}$$

We are now ready to upper bound $\mathrm{TV}(p_1,p_2)$ from Eq. (B21). The strategy is to divide the sum over all Pauli operators into $G^{(o_{1:N})}$ and $\mathsf{P}^n\backslash G^{(o_{1:N})}$. All terms within the former group are small thanks to the definition of $G^{(o_{1:N})}$; Terms within the latter group could be large, but the total number of them are small. Combining these two gives a pretty good upper bound on $\mathrm{TV}(p_1,p_2)$. In math,

$$\begin{aligned}\mathrm{TV}(p_1,p_2) &= \frac{1}{4^n-1}\sum_{\substack{o_{1:N}\ \text{s.t.}\\ p_1(o_{1:N})\ge p_2(o_{1:N})}} p_1(o_{1:N})\left(\sum_{a\in G^{(o_{1:N})}} + \sum_{\substack{a\in\mathbb{Z}_2^{2n}\backslash G^{(o_{1:N})},\\ a\neq 0}}\right)\left(1-\prod_{i=1}^{N}\left(1+\frac{\mathrm{Tr}\left(C_{o_i}^{o_{<i}\dagger}P_a C_{o_i}^{o_{<i}}P_a\right)}{\mathrm{Tr}\left(C_{o_i}^{o_{<i}\dagger}C_{o_i}^{o_{<i}}\right)}\right)\right)\\ &\le \frac{1}{4^n-1}\sum_{\substack{o_{1:N}\ \text{s.t.}\\ p_1(o_{1:N})\ge p_2(o_{1:N})}} p_1(o_{1:N})\left(\left|G^{(o_{1:N})}\right|\times\left(1-\left(1-\left(\frac{2^k}{2^n-1}\right)^{1/3}\right)^N\right)+\left(4^n-1-\left|G^{(o_{1:N})}\right|\right)\right)\\ &\le \sum_{\substack{o_{1:N}\ \text{s.t.}\\ p_1(o_{1:N})\ge p_2(o_{1:N})}} p_1(o_{1:N})\left(N\left(\frac{2^k}{2^n-1}\right)^{1/3}+N\left(\frac{2^k}{2^n-1}\right)^{1/3}\right)\\ &\le 2N\left(\frac{2^k}{2^n-1}\right)^{1/3}.\end{aligned} \tag{B28}$$

The first inequality uses an additional fact that $\left|\mathrm{Tr}(C^\dagger P_a C P_a)\right| / \left|\mathrm{Tr}(C^\dagger C)\right| \le 1$ to bound the second sum, which follows from the Cauchy-Schwarz inequality. The second inequality uses the bounds $\left|G^{(o_{1:N})}\right| \le 4^n-1$ for the first sum and Eq. (B27) for the second sum, as well as the fact that $1-(1-x)^N \le Nx$ for all $0\le x\le 1$. (Note that $2^k/(2^n-1)\le 1$ only if $k\le n-1$, but our targeted upper bound trivially holds for $k=n$.) This completes the proof of Lemma 8.

□

3. A lower bound for the most general ancilla-free strategies

The lower bounds in the previous sections work for non-concatenating strategies, where one is not allowed to concatenate (or, coherently access) multiple copies of the unknwon channel before doing a single measurement. The concatenating strategies, as depicted in Fig. 1, are however a natural apparatus for many randomized benchmarking protocols, where one effectively concatenate a varying number of noise channels in order to measure a series of

exponentially-decaying values, and then extract the parameters of interest via fitting the decay rate. The original purpose of such concatenation in these protocols is to eliminate the effect of state-preparation-and-measurement (SPAM) error. Here, we want to understand whether concatenating strategies also provide a sample complexity advantage. The short answer is no. We will present a sample complexity lower bound of $\Omega(2^{n/3})$ for the most powerful (adaptive, concatenating) ancilla-free ($k = 0$) Pauli channel estimation schemes. This result justifies our claim that ancillary systems are indeed indispensable to overcome the exponential barrier in sample complexity.

To start with, we give a rigorous definition of the most general ancilla-free strategies that we are going to study

Definition 1. *Let Λ be an unknwon n-qubit Pauli channel. An adaptive, concatenating, ancilla-free ($k = 0$) estimation protocol is specified by the following parameters. Let N denote the total rounds of measurements. For the ith round, let $\rho^{o_{<i}}$ denote the input state, $\{E_{o_i}^{o_{<i}}\}_{o_i}$ denote the POVM measurement, $M^{o_{<i}}$ denote the length of concatenation, and $\{\mathcal{C}_k^{o_{<i}}\}_{k=1}^{M^{o_{<i}}}$ denote a set of processing channels. The ith measurement outcome is given by o_i with probability*

$$\Pr(o_i|o_{<i}) = \operatorname{Tr}\left[E_{o_i}^{o_{<i}}\Lambda(\mathcal{C}_{M^{o_{<i}}-1}^{o_{<i}}(\cdots\mathcal{C}_2^{o_{<i}}(\Lambda(\mathcal{C}_1^{o_{<i}}(\Lambda(\rho^{o_{<i}}))))\cdots))\right]. \tag{B29}$$

All the superscript $o_{<i} \coloneqq [o_1, \cdots, o_{i-1}]$ are used to emphasize the dependence on previous measurement outcomes. The protocol should produce an estimate of Λ via classical processing on the measurement outcomes $o_{1:N}$.

The problem we are interested in is still approximating the Pauli eigenvalues $\boldsymbol{\lambda}$ to small error in l_∞ distance. Note that, the parameter N in the above definition is not exactly the sample complexity but is the total number of measurements conducted. Since one is allowed to concatenate multiple copies of Λ in a single measurement, N is a lower bound for the sample complexity of Λ. Our result is summarized in the following theorem.

Theorem 9. *For any adaptive, concatenating, ancilla-free ($k = 0$) protocols that gives an estimate $\widehat{\boldsymbol{\lambda}}$ of the Pauli eigenvalues $\boldsymbol{\lambda}$ of an arbitrary unknown n-qubit Pauli channel Λ such that*

$$|\widehat{\lambda}_a - \lambda_a| < \frac{1}{2}, \quad \forall a \in \mathbb{Z}_2^{2n} \tag{B30}$$

holds with high probability, the rounds of measurements N (and hence the number of samples of Λ) required is at least $\Omega(2^{n/3})$.

Proof. The proof methods are similar to the proof of Theorem 7. Define the following Pauli channels.

$$\begin{cases} \Lambda_{\text{dep}}(\cdot) = \dfrac{1}{2^n} I \operatorname{Tr}(\cdot), \\ \Lambda_a(\cdot) = \dfrac{1}{2^n}\left(I\operatorname{Tr}(\cdot) + P_a \operatorname{Tr}(P_a(\cdot))\right), \ \forall a \in \{1, \cdots, 4^n - 1\}. \end{cases} \tag{B31}$$

We consider the problem of distinguishing the following two equal-probable hypotheses

1. Given N copies of $\Lambda = \Lambda_{\text{dep}}$.
2. Given N copies of $\Lambda = \Lambda_a$ for a uniformly-randomly picked $a \in \{1, \cdots, 4^n - 1\}$.

An estimation protocol satisfying our assumptions should be able to distinguish these two hypotheses with high probability. Let the probability distribution of the measurement outcomes $o_{1:N}$ under these two hypotheses be p_1 and p_2, respectively. The total variance distance $\text{TV}(p_1, p_2)$ must be at least $\Omega(1)$ for the distinguishing task to succeed with high probability. We will show in the following that $\text{TV}(p_1, p_2) = O(N2^{-n/3})$, which then gives the claimed lower bound of $N = \Omega(2^{n/3})$.

To start with, based on the same argument as in the proof of Lemma 8, it suffices to consider pure state input and rank-1 POVM measurements, so we replace $\rho^{o_{<i}}$ and $\{E_{o_i}^{o_{<i}}\}_{o_i}$ in Definition 1 with $|A^{o_{<i}}\rangle\langle A^{o_{<i}}|$ and $\{w_{o_i}^{o_{<i}} 2^n |B^{o_{<i}}\rangle\langle B^{o_{<i}}|\}_{o_i}$ respectively, where $|A^{o_{<i}}\rangle, |B_{o_i}^{o_{<i}}\rangle \in \mathbb{C}^{2^n}$ are unit vectors, and $\sum_{o_i} w_{o_i}^{o_{<i}} = 1$ by normalization.

Next, we calculate the distribution of $o_{1:N}$ under the two different hypotheses. The expression for p_1 can be easily obtained, as Λ_{dep} is simply the completely deplorizing channel. We have

$$\begin{aligned} p_1(o_{1:N}) &= \prod_{i=1}^N w_{o_i}^{o_{<i}} 2^n \langle B_{o_i}^{o_{<i}}| \Lambda_{\text{dep}}(\mathcal{C}_{M^{o_{<i}}-1}^{o_{<i}}(\cdots\mathcal{C}_2^{o_{<i}}(\Lambda_{\text{dep}}(\mathcal{C}_1^{o_{<i}}(\Lambda_{\text{dep}}(|A^{o_{<i}}\rangle\langle A^{o_{<i}}|)))))\cdots)) |B_{o_i}^{o_{<i}}\rangle \\ &= \prod_{i=1}^N w_{o_i}^{o_{<i}}. \end{aligned} \tag{B32}$$

The expression for p_2 is more complicated. We first define the following recursive expression

$$\xi_a^{o_{<i}}[m] \coloneqq \begin{cases} 2^{-n} \operatorname{Tr}\left(P_a \mathcal{C}_{m-1}(I + P_a \xi_a^{o_{<i}}[m-1])\right), & 2 \le m \le M^{o_{<i}}, \\ \langle A^{o_{<i}}| \, P_a \, |A^{o_{<i}}\rangle, & m = 1. \end{cases} \tag{B33}$$

The expression for p_2 can then be calculated as follows

$$\begin{aligned} p_2(o_{1:N}) &= \mathop{\mathbb{E}}_{a\neq 0} \prod_{i=1}^{N} w_{o_i}^{o_{<i}} 2^n \langle B_{o_i}^{o_{<i}}| \, \Lambda_a(\mathcal{C}_{M^{o_{<i}}-1}^{o_{<i}}(\cdots \mathcal{C}_2^{o_{<i}}(\Lambda_a(\mathcal{C}_1^{o_{<i}}(\Lambda_a(|A^{o_{<i}}\rangle\langle A^{o_{<i}}|)))) \cdots)) \, |B_{o_i}^{o_{<i}}\rangle \\ &= \mathop{\mathbb{E}}_{a\neq 0} \prod_{i=1}^{N} w_{o_i}^{o_{<i}} \langle B_{o_i}^{o_{<i}}| \, \Lambda_a(\mathcal{C}_{M^{o_{<i}}-1}^{o_{<i}}(\cdots \mathcal{C}_2^{o_{<i}}(\Lambda_a(\mathcal{C}_1^{o_{<i}}(I + P_a \xi_a^{o_{<i}}[1]))) \cdots)) \, |B_{o_i}^{o_{<i}}\rangle \\ &= \mathop{\mathbb{E}}_{a\neq 0} \prod_{i=1}^{N} w_{o_i}^{o_{<i}} \langle B_{o_i}^{o_{<i}}| \, \Lambda_a(\mathcal{C}_{M^{o_{<i}}-1}^{o_{<i}}(\cdots \mathcal{C}_2^{o_{<i}}(I + 2^{-n} P_a \operatorname{Tr}(P_a \mathcal{C}_1^{o_{<i}}(I + P_a \xi_a^{o_{<i}}[1])) \cdots)) \, |B_{o_i}^{o_{<i}}\rangle \\ &= \mathop{\mathbb{E}}_{a\neq 0} \prod_{i=1}^{N} w_{o_i}^{o_{<i}} \langle B_{o_i}^{o_{<i}}| \, \Lambda_a(\mathcal{C}_{M^{o_{<i}}-1}^{o_{<i}}(\cdots \mathcal{C}_2^{o_{<i}}(I + P_a \xi_a^{o_{<i}}[2]) \cdots)) \, |B_{o_i}^{o_{<i}}\rangle \\ &= \cdots \\ &= \mathop{\mathbb{E}}_{a\neq 0} \prod_{i=1}^{N} w_{o_i}^{o_{<i}} \langle B_{o_i}^{o_{<i}}| \, I + P_a \xi_a^{o_{<i}}[M^{o_{<i}}] \, |B_{o_i}^{o_{<i}}\rangle \\ &= \mathop{\mathbb{E}}_{a\neq 0} \prod_{i=1}^{N} w_{o_i}^{o_{<i}} (1 + \xi_a^{o_{<i}}[M^{o_{<i}}] \, \langle B_{o_i}^{o_{<i}}| \, P_a \, |B_{o_i}^{o_{<i}}\rangle). \end{aligned} \tag{B34}$$

The third line uses the fact that $\mathcal{C}_k^{o_{<i}}$ is trace-preserving. The total variance distance between p_1 and p_2 is then

$$\begin{aligned} \mathrm{TV}(p_1, p_2) &= \mathop{\mathbb{E}}_{a\neq 0} \sum_{\substack{o_{1:N} \text{ s.t.} \\ p_1(o_{1:N}) \ge p_2(o_{1:N})}} \left(\prod_{i=1}^{N} w_{o_i}^{o_{<i}}\right) \left(1 - \prod_{i=1}^{N} \left(1 + \xi_a^{o_{<i}}[M^{o_{<i}}] \, \langle B_{o_i}^{o_{<i}}| \, P_a \, |B_{o_i}^{o_{<i}}\rangle\right)\right) \\ &= \mathop{\mathbb{E}}_{a\neq 0} \sum_{\substack{o_{1:N} \text{ s.t.} \\ p_1(o_{1:N}) \ge p_2(o_{1:N})}} p_1(o_{1:N}) \left(1 - \prod_{i=1}^{N} \left(1 + \xi_a^{o_{<i}}[M^{o_{<i}}] \, \langle B_{o_i}^{o_{<i}}| \, P_a \, |B_{o_i}^{o_{<i}}\rangle\right)\right). \end{aligned} \tag{B35}$$

We now need a bound of $|\xi_a^{o_{<i}}[M^{o_{<i}}]| \le 1$, which can be shown by induction. We see $|\xi_a^{o_{<i}}[1]| \le 1$ by definition. Suppose $|\xi_a^{o_{<i}}[m-1]| \le 1$, we have

$$\begin{aligned} |\xi_a^{o_{<i}}[m]| &= \left|\operatorname{Tr}\left(P_a \mathcal{C}_{m-1}^{o_{<i}}\left(\frac{I + P_a \xi_a^{o_{<i}}[m-1]}{2^n}\right)\right)\right| \\ &\le \|P_a\|_\infty \operatorname{Tr}\left|\mathcal{C}_{m-1}^{o_{<i}}\left(\frac{I + P_a \xi_a^{o_{<i}}[m-1]}{2^n}\right)\right| \\ &= \operatorname{Tr}\left(\frac{I + P_a \xi_a^{o_{<i}}[m-1]}{2^n}\right) \\ &= 1. \end{aligned} \tag{B36}$$

The first line is by the defining recursive expression; The second line is by the tracial matrix Hölder inequality; The third line uses the fact that $\mathcal{C}_{m-1}^{o_{<i}}$ is a positive map, and that $2^{-n}(I + P_a \xi_a^{o_{<i}}[m-1])$ is positive semidefinite thanks to the induction hypothesis $|\xi_a^{o_{<i}}[m-1]| \le 1$. Thus we can remove the modulus within the trace, and also remove $\mathcal{C}_{m-1}^{o_{<i}}$ as it is trace-preserving. By induction, we've shown $|\xi_a^{o_{<i}}[M^{o_{<i}}]| \le 1$.

The remaining part of bounding $\mathrm{TV}(p_1, p_2)$ is basically the same as in the proof of Lemma 8. We repeat it here for completeness. Let $|B\rangle$ be any n-qubit pure state. Consider the following subset of n-qubit Pauli operators

$$G \coloneqq \left\{ a \in \{1, \cdots, 4^n - 1\} \ : \ |\langle B| \, P_a \, |B\rangle| \le \left(\frac{1}{2^n + 1}\right)^{1/3} \right\}. \tag{B37}$$

We claim that the size of G satisfies

$$|G| \geq \left(1 - \left(\frac{1}{2^n+1}\right)^{1/3}\right)(4^n - 1), \tag{B38}$$

which can be shown by contradiction: Suppose this does not hold, we would have

$$\begin{aligned}\sum_{a=1}^{4^n-1} \langle B| P_a |B\rangle^2 &\geq \left(\frac{1}{2^n+1}\right)^{2/3} \times (4^n - 1 - |G|) \\ &> \left(\frac{1}{2^n+1}\right)^{2/3} \times \left(\frac{1}{2^n+1}\right)^{1/3} (4^n - 1) \\ &= 2^n - 1.\end{aligned} \tag{B39}$$

However, the L.H.S. of the above can be calculated as

$$\sum_{a=1}^{4^n-1} \langle B| P_a |B\rangle^2 = \sum_{a=0}^{4^n-1} \langle B| P_a |B\rangle^2 - 1 = 2^n - 1. \tag{B40}$$

This leads to contradition, and hence proves the desired lower bound on $|G|$.

Now, we define another subset of n-qubit Pauli operators, conditioned on the measurement outcomes $o_{1:N}$, as follows

$$G^{(o_{1:N})} := \left\{ a \in \{1, \cdots, 4^n - 1\} \ : \ \left|\langle B_{o_i}^{o_{<i}}| P_a |B_{o_i}^{o_{<i}}\rangle\right| \leq \left(\frac{1}{2^n+1}\right)^{1/3}, \ \forall i = 1, \cdots, N \right\}. \tag{B41}$$

By applying a "union bound" on Eq. (B37), we immediately have the following lower bound on the size of $G^{(o_{1:N})}$:

$$\left|G^{(o_{1:N})}\right| \geq \left(1 - N\left(\frac{1}{2^n+1}\right)^{1/3}\right)(4^n - 1). \tag{B42}$$

We are now ready to upper bound $\mathrm{TV}(p_1, p_2)$ from Eq. (B35). The strategy is to divide the sum over all Pauli operators into $G^{(o_{1:N})}$ and $\mathsf{P}^n \backslash G^{(o_{1:N})}$. All terms within the former group are small thanks to the definition of $G^{(o_{1:N})}$; Terms within the latter group could be large, but the total number of them are small. Combining these two gives a pretty good upper bound on $\mathrm{TV}(p_1, p_2)$. In math,

$$\begin{aligned}\mathrm{TV}(p_1, p_2) &= \frac{1}{4^n - 1} \sum_{\substack{o_{1:N} \text{ s.t.} \\ p_1(o_{1:N}) \geq p_2(o_{1:N})}} p_1(o_{1:N}) \left(\sum_{a \in G^{(o_{1:N})}} + \sum_{\substack{a \in \mathbb{Z}_2^{2n} \backslash G^{(o_{1:N})}, \\ a \neq 0}} \right) \left(1 - \prod_{i=1}^{N} \left(1 + \xi_a^{o_{<i}}[M^{o_{<i}}] \langle B_{o_i}^{o_{<i}}| P_a |B_{o_i}^{o_{<i}}\rangle\right)\right) \\ &\leq \frac{1}{4^n - 1} \sum_{\substack{o_{1:N} \text{ s.t.} \\ p_1(o_{1:N}) \geq p_2(o_{1:N})}} p_1(o_{1:N}) \left(\left|G^{(o_{1:N})}\right| \times \left(1 - \left(1 - \left(\frac{1}{2^n+1}\right)^{1/3}\right)^N\right) + \left(4^n - 1 - \left|G^{(o_{1:N})}\right|\right) \right) \\ &\leq \sum_{\substack{o_{1:N} \text{ s.t.} \\ p_1(o_{1:N}) \geq p_2(o_{1:N})}} p_1(o_{1:N}) \left(N\left(\frac{1}{2^n+1}\right)^{1/3} + N\left(\frac{1}{2^n+1}\right)^{1/3} \right) \\ &\leq 2N\left(\frac{1}{2^n+1}\right)^{1/3}.\end{aligned} \tag{B43}$$

The first inequality applies the bound $|\xi_a^{o_{<i}}[M^{o_{<i}}]| \leq 1$. Besides, the first sum uses the bound from the definition of $G^{(o_{1:N})}$, and the second sum is bounded using $|\langle B_{o_i}^{o_{<i}}| P_a |B_{o_i}^{o_{<i}}\rangle| \leq 1$. The second inequality uses the bounds $\left|G^{(o_{1:N})}\right| \leq 4^n - 1$ for the first sum and Eq. (B42) for the second sum, as well as the fact that $1 - (1-x)^N \leq Nx$ for all $0 \leq x \leq 1$. Now we have obtained the claimed bound $\mathrm{TV}(p_1, p_2) = O(N 2^{-n/3})$, and hence complete the proof of Theorem 9. □

4. A lower bound for the most general entangled strategies

In this section, we prove a lower bound of $\Omega(n)$ for the fully entangled estimation strategies, which is the most general measurement strategies one can do to learn an unknown channel even with the help of quantum computers, see Fig. 1 (a). Note that, since we assume there is an unlimited amount of quantum memory, we can without loss of generality eliminate any intermediate measurements, and only conduct one joint measurement after sequentially processing all N samples of the channel.

Theorem 10. *For any fully entangled measurement protocols that give an estimate $\widehat{\boldsymbol{\lambda}}$ for the Pauli eigenvalues $\boldsymbol{\lambda}$ of an arbitrary unknown n-qubit Pauli channel Λ such that*

$$|\widehat{\lambda}_a - \lambda_a| < \frac{1}{2}, \quad \forall a \in \mathbb{Z}_2^{2n} \tag{B44}$$

holds with high probability, the number of samples of Λ required is at least $\Omega(n)$.

Proof. We first show that, by using a technique known as *teleportation stretching* [37, 38], any fully entangled measurement protocols for N copies of an arbitrary Pauli channel Λ can be simulated by a joint measurement on N copies of the Choi state J_Λ which is defined as

$$\begin{aligned} J_\Lambda &\coloneqq \Lambda \otimes \mathbb{1}(|\Psi^+\rangle\langle\Psi^+|) \\ &= \frac{1}{4^n} \sum_{a \in \mathbb{Z}_2^{2n}} \lambda_a P_a \otimes P_a^T. \end{aligned} \tag{B45}$$

This follows from the existence of a quantum channel $\mathcal{T}$ such that $\Lambda(\rho) = \mathcal{T}(\rho \otimes J_\Lambda)$ holds for all Pauli channels Λ. One possible construction of $\mathcal{T}$ is shown in Fig. 4 (see Ref. [53]). In word, one first applies a Bell measurement on the input state ρ and half of the Choi state J_Λ. Then, conditioned on the Bell measurement outcome $|\Psi_b\rangle$, one applies a Pauli correction P_b on the other half of J_Λ, which will then be equal to $\Lambda(\rho)$. Indeed, the post-measurement state conditioned on Bell measurement outcome b is

$$\begin{aligned} \rho_b &\propto \langle\Psi_b|_{AB} \, \rho^A \otimes J_\Lambda^{BC} \, |\Psi_b\rangle_{AB} \\ &\propto \sum_{a \in \mathbb{Z}_2^{2n}} \lambda_a \, \langle\Psi_b| \, \rho \otimes P_a \, |\Psi_b\rangle \otimes P_a^T \\ &= \frac{1}{2^n} \sum_{a \in \mathbb{Z}_2^{2n}} \lambda_a (-1)^{\langle a,b \rangle} \operatorname{Tr}(\rho P_a) P_a. \end{aligned} \tag{B46}$$

After applying the Pauli correction, the state becomes

$$P_b \rho_b P_b = \frac{1}{2^n} \sum_{a \in \mathbb{Z}_2^{2n}} \lambda_a \operatorname{Tr}(\rho P_a) P_a = \Lambda(\rho), \tag{B47}$$

which justify the relation $\Lambda(\rho) = \mathcal{T}(\rho \otimes J_\Lambda)$.

With the help of teleportation stretching, one can reduce any measurement protocols for N copies of Λ to a single POVM measurement on N copies of J_Λ, as shown in Fig. 5.

Now, recall the communication task defined in Sec. B 1, where Alice and Bob share the following "codebook"

$$(a, s) \in \{1, \cdots, 4^n - 1\} \times \{\pm 1\} \;\longrightarrow\; \Lambda_{(a,s)}(\cdot) = \frac{1}{2^n} \left(I \operatorname{Tr}(\cdot) + s P_a \operatorname{Tr}(P_a(\cdot)) \right), \tag{B48}$$

and Alice randomly picks one possible (a, s) and send N copies of $\Lambda_{(a,s)}$ to Bob. If there exists a fully entangled estimation protocol using N samples and satisfying the assumption of Theorem 10, Bob can determine Alice's choice of (a, s) with high probability. According to Fano's inequality, the mutual information between the random variable pair (a, s) and Bob's measurement result o has the following lower bound

$$I((a, s) : o) \geq \Omega(\log 2(4^n - 1)) = \Omega(n). \tag{B49}$$

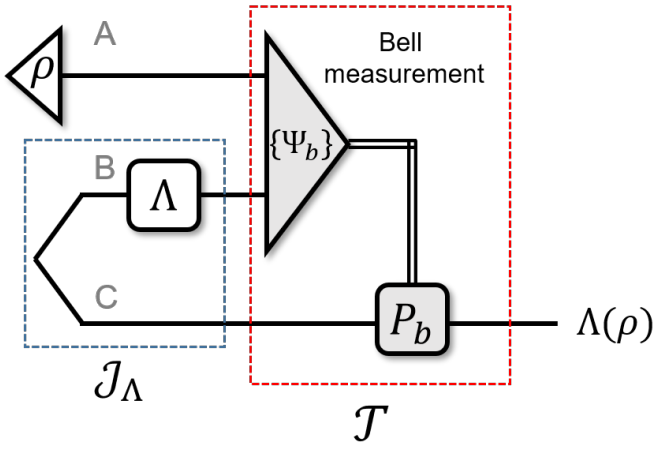

FIG. 4. Construction of the teleportation simulation channel $\mathcal{T}$.

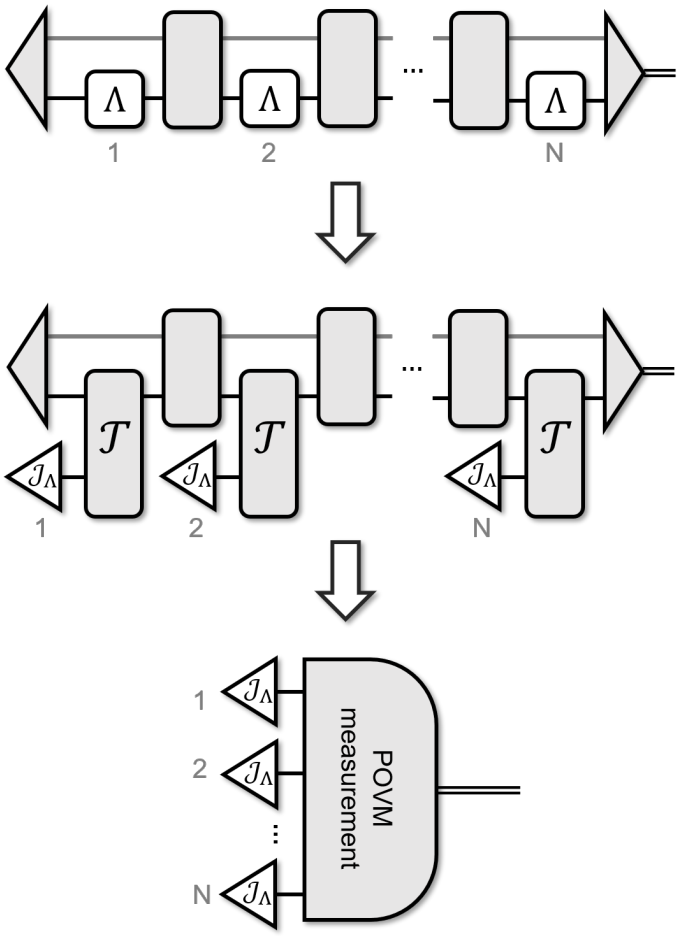

FIG. 5. Teleportation stretching [37, 38]: simulation of an arbitrary entangled measurement on Λ by a single measurement on J_Λ. One just needs to simulate every application of $\Lambda(\cdot)$ by $\mathcal{T}((\cdot)\otimes\rho)$. The whole measurement protocol then become a joint POVM measurement on N copies of J_Λ with no adaptivity.

One the other hand, since any measurement Bob conducts can be simulated by a measurement on N copies of $J_{\Lambda_{(a,s)}}$,

Holevo's theorem [41, 54] can be apply to $I((a,s):o)$. We have

$$\begin{aligned} I((a,s):o) &\le S\left(\mathop{\mathbb{E}}_{(a,s)} J_{\Lambda_{(a,s)}}^{\otimes N}\right) - \mathop{\mathbb{E}}_{(a,s)} S\left(J_{\Lambda_{(a,s)}}^{\otimes N}\right) \\ &= S\left(\mathop{\mathbb{E}}_{(a,s)} J_{\Lambda_{(a,s)}}^{\otimes N}\right) - N \mathop{\mathbb{E}}_{(a,s)} S\left(J_{\Lambda_{(a,s)}}\right) \\ &\le 2nN - (2n-1)N \\ &= N. \end{aligned} \tag{B50}$$

In the third line, the first term is a trivial upper bound for the von Neumann entropy on a 2^{2nN}-dimensional Hilbert space. The second term uses the observation that

$$J_{\Lambda_{(a,s)}} = \frac{1}{4^n}(I\otimes I + sP_a \otimes P_a^T) \tag{B51}$$

is a maximally mixed state on a 2^{2n-1}-dimensional Hilbert space, thus $S(J_{\Lambda_{(a,s)}}) = 2n-1$. This yields the lower bound $N = \Omega(n)$. □

Appendix C: SPAM-robust ancilla-assisted Pauli gate benchmarking protocols

In this section we present the proof of Theorem 4. To start with, we introduce the *Pauli-transfer-matrix* (PTM) representation to simplify notations. A linear operator O acting on a 2^n-dimensional Hilbert space can be viewed as a vector in a 4^n-dimensional Hilbert space. We denote this vectorization of O as $|O\rangle\!\rangle$ and the corresponding Hermitian conjugate as $\langle\!\langle O|$. The inner product within this space is the *Hilbert-Schmidt* product defined as $\langle\!\langle A|B\rangle\!\rangle := \mathrm{Tr}(A^\dagger B)$. The *normalized Pauli operators* $\{\sigma_a := P_a/\sqrt{2^n},\ a\in\mathbb{Z}_2^{2n}\}$ forms an orthonormal basis for this space. In the PTM representation, a superoperator (*i.e.*, quantum channel) becomes an operator acting on the 4^n-dimensional Hilbert space, sometimes called the Pauli transfer operator. Explicitly, we have $|\Lambda(\rho)\rangle\!\rangle = \Lambda^{\mathrm{PTM}}|\rho\rangle\!\rangle \equiv \Lambda|\rho\rangle\!\rangle$, where we use the same notation to denote a channel and its Pauli transfer operator, which should be clear from the context. Specifically, a general Pauli channel Λ has the follwing Pauli transfer operator

$$\Lambda = \sum_{a\in\mathbb{Z}_2^{2n}} \lambda_a |\sigma_a\rangle\!\rangle\langle\!\langle\sigma_a|,$$

where $\{\lambda_a\}_a$ are the Pauli eigenvalues. It is also obvious that the m-th power of Λ is

$$\Lambda^m = \sum_{a\in\mathbb{Z}_2^{2n}} \lambda_a^m |\sigma_a\rangle\!\rangle\langle\!\langle\sigma_a|.$$

Using the PTM representation, the constant A_a in Theorem 4 is defined as

$$A_a := \sum_{v\in\mathbb{Z}_2^{2n}} (-1)^{\langle a,v\rangle} \langle\!\langle\widetilde{\Psi}_v|\mathbb{1}\otimes(|\sigma_a\rangle\!\rangle\langle\!\langle\sigma_a|\Lambda_G)|\widetilde{\Psi}^+\rangle\!\rangle, \tag{C1}$$

where $|\widetilde{\Psi}^+\rangle\!\rangle$ is just the PTM representation for the density matrix of the (noisy) Bell state $\widetilde{\Psi}^+$. Same for $|\widetilde{\Psi}_b\rangle\!\rangle$. One can verify that $A_a = 1$ for the noiseless case (where there is no SPAM error and $\Lambda_G = \mathbb{1}$).

Proof of Theorem 4. The following proof is a generalization of [23, Proposition 5] and we borrow some of their presentations. Consider the probability that a specific sequence of Pauli gates $\{P_{a_t}\}_{t=0}^m$ is sampled (Line 4, Alg. 2) and the Bell measurement outcome is v (Line 5, Alg. 2),

$$\begin{aligned} \Pr(a_0,\cdots,a_m,v) &= \frac{1}{4^{n(m+1)}} \langle\!\langle\widetilde{\Psi}_v|\mathbb{1}\otimes(\mathcal{P}_{a_m}\Lambda_G\cdots\mathcal{P}_{a_1}\Lambda_G\mathcal{P}_{a_0}\Lambda_G)|\widetilde{\Psi}^+\rangle\!\rangle \\ &= \frac{1}{4^{n(m+1)}} \langle\!\langle\widetilde{\Psi}_v|\mathbb{1}\otimes\left(\prod_{t=m}^{0}\mathcal{P}_{a_t}\Lambda_G\right)|\widetilde{\Psi}^+\rangle\!\rangle, \end{aligned} \tag{C2}$$

Here we use the assumption that the noise on the ancilla is negligible except for the state preparation and measurement part. We can absorb the noise channel on the ancilla into the SPAM error that is independent of the concatenating length m. That is why we can have an $\mathbb{1}$ on the ancillary system.

The distribution can be rewritten as

$$\Pr(a_0,\cdots,a_m,v)=\frac{1}{4^{n(m+1)}}\langle\!\langle\widetilde{\Psi}_v|\mathbb{1}\otimes P_{a'_m}\left(\prod_{t=m-1}^{0}\mathcal{P}_{a'_t}\Lambda_G\mathcal{P}_{a'_t}\right)\Lambda_G|\widetilde{\Psi}^+\rangle\!\rangle, \tag{C3}$$

where we define $a'_t\coloneqq\sum_{k=0}^{t}a_k$ (bit-wise modulo 2 sum). Taking this change-of-variables and averaging over $\{a_0,\ \cdots,\ a_{m-1}\}$, we get

$$\begin{aligned}\Pr(a'_m,v)&=\frac{1}{4^n}\langle\!\langle\widetilde{\Psi}_v|\mathbb{1}\otimes\mathcal{P}_{a'_m}\left(\prod_{t=m-1}^{0}\mathop{\mathbb{E}}_{a'_t\in\mathbb{Z}_2^{2n}}\mathcal{P}_{a'_t}\Lambda_G\mathcal{P}_{a'_t}\right)\Lambda_G|\widetilde{\Psi}^+\rangle\!\rangle\\&=\frac{1}{4^n}\langle\!\langle\widetilde{\Psi}_v|\mathbb{1}\otimes\mathcal{P}_{a'_m}\Lambda^m\Lambda_G|\widetilde{\Psi}^+\rangle\!\rangle\\&=\frac{1}{4^n}\sum_{a\in\mathbb{Z}_2^{2n}}\lambda_a^m\langle\!\langle\widetilde{\Psi}_v|\mathbb{1}\otimes(\mathcal{P}_{a'_m}|\sigma_a\rangle\!\rangle\langle\!\langle\sigma_a|\Lambda_G)|\widetilde{\Psi}^+\rangle\!\rangle\\&=\frac{1}{4^n}\sum_{a\in\mathbb{Z}_2^{2n}}(-1)^{\langle a,a'_m\rangle}\lambda_a^m\langle\!\langle\widetilde{\Psi}_v|\mathbb{1}\otimes(|\sigma_a\rangle\!\rangle\langle\!\langle\sigma_a|\Lambda_G)|\widetilde{\Psi}^+\rangle\!\rangle.\end{aligned} \tag{C4}$$

Define $z\coloneqq v+a'_m$, the marginal distribution of z is

$$\begin{aligned}\Pr(z)&=\sum_{v\in\mathbb{Z}_2^{2n}}\Pr(v+z,v)\\&=\frac{1}{4^n}\sum_{a\in\mathbb{Z}_2^{2n}}(-1)^{\langle a,z\rangle}\lambda_a^m\sum_{v\in\mathbb{Z}_2^{2n}}(-1)^{\langle a,v\rangle}\langle\!\langle\widetilde{\Psi}_v|\mathbb{1}\otimes(|\sigma_a\rangle\!\rangle\langle\!\langle\sigma_a|\Lambda_G)|\widetilde{\Psi}^+\rangle\!\rangle\\&=\frac{1}{4^n}\sum_{a\in\mathbb{Z}_2^{2n}}(-1)^{\langle a,z\rangle}\lambda_a^m A_a.\end{aligned} \tag{C5}$$

Apply the inverse Walsh-Hadamard transform, we obtain

$$A_a\lambda_a^m=\sum_{z\in\mathbb{Z}_2^{2n}}(-1)^{\langle a,z\rangle}\Pr(z). \tag{C6}$$

Therefore, $(-1)^{\langle a,z\rangle}=(-1)^{\langle a,v\rangle+\sum_{t=0}^{m}\langle a,a_t\rangle}$ is an unbiased estimator for $A_a\lambda_a^m$. In other word,

$$\mathbb{E}\left[\widehat{F}_a^{(k)}(m)\right]\equiv\mathbb{E}\left[(-1)^{\langle a,z\rangle}\right]=A_a\lambda_a^m. \tag{C7}$$

This is exactly the claim of Theorem 4. □