Lower Bounds on Learning Pauli Channels

Omar Fawzi, Aadil Oufkir, Daniel Stilck Franca

Lower Bounds on Learning Pauli Channels

Omar Fawzi, Aadil Oufkir, and Daniel Stilck França

Univ Lyon, Inria, ENS Lyon, UCBL, LIP, Lyon, France

Abstract

Understanding the noise affecting a quantum device is of fundamental importance for scaling quantum technologies. A particularly important class of noise models is that of Pauli channels, as randomized compiling techniques can effectively bring any quantum channel to this form and are significantly more structured than general quantum channels. In this paper, we show fundamental lower bounds on the sample complexity for learning Pauli channels in diamond norm with unentangled measurements. We consider both adaptive and non-adaptive strategies. In the non-adaptive setting, we show a lower bound of $\Omega(2^{3n}\varepsilon^{-2})$ to learn an n-qubit Pauli channel. In particular, this shows that the recently introduced learning procedure by Flammia and Wallman (2020) is essentially optimal. In the adaptive setting, we show a lower bound of $\Omega(2^{2.5n}\varepsilon^{-2})$ for $\varepsilon = \mathcal{O}(2^{-n})$, and a lower bound of $\Omega(2^{2n}\varepsilon^{-2})$ for any $\varepsilon > 0$. This last lower bound even applies for arbitrarily many sequential uses of the channel, as long as they are only interspersed with other unital operations.

1 Introduction

In spite of their impressive progress over the last few years Arute et al. (2019); Ebadi et al. (2021); Scholl et al. (2021); Zhong et al. (2020), the scaling and effective employment of quantum technologies still face many challenges. One of the most significant ones is how to tame the noise affecting such devices. For that, more effective tools are required to characterize and learn noisy quantum channels Eisert et al. (2020). As the number of parameters required to describe a quantum channel scales exponentially in the size of the device, it is challenging to learn the noise beyond a few qubits.

A class of quantum channels that deserves particular attention is that of Pauli channels (Watrous, 2018, Sec. 4.1.2). The reasons for that are manifold. First, Pauli channels provide a simple and effective model of incoherent noise, admitting a representation in terms of a probability distribution corresponding to different Pauli errors and inheriting the rich structure of the Pauli matrices. Second, they are a physically relevant noise model and the noise affecting a device can always be mapped into a Pauli channel by using randomized compiling Wallman and Emerson (2016) techniques without incurring a loss in fidelity. These properties make the problem of Pauli tomography, i.e. learning a Pauli channel, particularly relevant. Finally, Pauli channel tomography is also known to be a problem for which quantum resources provide an advantage Chen, Zhou, Seif, and Jiang (2022).

Furthermore, reliable protocols to learn quantum channels face the additional hurdle that there might be errors both in the initial state preparation and measurements (SPAM errors). Thus, it is desirable to design protocols that are robust to such errors. And, of course, practical protocols should not rely on the preparation of complex states or measurements. A popular and widely used protocol to learn Pauli channels that fulfills these desiderata is that of randomized benchmarking and its variations Flammia and Wallman (2020); França and Hashagen (2018); Helsen, Roth, Onorati, Werner, and Eisert (2022); Helsen, Xue, Vandersypen, and Wehner (2019); Magesan, Gambetta, and Emerson (2012). Finally, Pauli noise model reflects experiments that are actually done in practice (see e.g., Harper, Flammia, and Wallman (2020) which includes an experimental implementation). It is thus natural to ask to what extent it is optimal or whether we could hope for better protocols to learn Pauli noise.

Contributions We provide lower bounds on the number of measurements or channel uses for learning a Pauli quantum channel in diamond norm using incoherent measurements and no auxiliary systems in both non-adaptive and adaptive settings (see Table 1 for a summary). Let $d = 2^n$ the dimension of the input and output of the unknown Pauli channel on n qubits and $\varepsilon > 0$ the precision parameter.

- **Non-adaptive setting:** We show that any non-adaptive learning algorithm of a Pauli channel should, at the worst case, use at least $\Omega(d^3/\varepsilon^2)$ measurements or a total number $\Omega(d^4/\varepsilon^6)$ of

Model	Lower bound	Upper bound
Non-adaptive, ℓ_1-distance	$N = \Omega(d^3/\varepsilon^2)$ or $\sum_{t=1}^{N} m_t = \Omega(d^4/\varepsilon^6)$ [this work]	$N = \tilde{\mathcal{O}}(d^3/\varepsilon^2)$ Flammia and Wallman (2020)
Non-adaptive, ℓ_∞-distance	$N = \Omega(1/\varepsilon^2)$ Flammia and O'Donnell (2021)	$N = \tilde{\mathcal{O}}(1/\varepsilon^2)$ Flammia and O'Donnell (2021)
Adaptive, ℓ_1-distance	$N = \Omega(d^2/\varepsilon^2)$ [this work]	$N = \tilde{\mathcal{O}}(d^3/\varepsilon^2)$ Flammia and Wallman (2020)
Adaptive, ℓ_1-distance $\varepsilon \leq 1/(20d)$, $\forall t \in [N] : m_t = 1$	$N = \Omega(d^{2.5}/\varepsilon^2)$ [this work]	$N = \tilde{\mathcal{O}}(d^3/\varepsilon^2)$ Flammia and Wallman (2020)

Table 1: Lower and upper bounds for Pauli channel tomography using incoherent measurements. N is the total number of steps or measurements. At each step $t \in [N]$, m_t denotes the total number of channel uses between the $(t-1)^{\text{th}}$ and t^{th} measurements.

channel uses. In particular, this shows that the randomized benchmarking algorithm of (Flammia & Wallman, 2020, Result 1) is almost optimal since the channels we consider in our construction have a spectral gap $\Delta \geq 1 - 4\varepsilon$ and thus the total number of channel uses is at most twice the number of measurements. This result is stated in Theorem 4.1. For the proof, we construct an ε-separated family of Pauli channels close to the maximally depolarizing channel and use it to encode a message from $[e^{\Omega(d^2)}]$. A learning algorithm can be used to decode this message with the same success probability. Hence, the encoder and decoder should share at least $\Omega(d^2)$ nats of information. On the other hand, after each step, we show that the correlation between the encoder and decoder can only increase by at most $\mathcal{O}(\varepsilon^2/d)$ nats if the channel is used at most 2 times. Moreover, if the channel is used $m \geq 3$ times, we show in this case that the correlation between the encoder and decoder can only increase by at most $\mathcal{O}(m\varepsilon^6/d^2)$ nats. Note that the naive upper bound on this correlation is $\mathcal{O}(\varepsilon^2)$, we obtain an improvement by a factor d or d^2/m by exploiting the randomness in the construction of the Pauli channel.

- **Adaptive setting:** We show that in general, any learning algorithm of a Pauli channel should use at least $\Omega(d^2/\varepsilon^2)$ measurements no matter how many times the channel is applied and intertwined with other unital operations before each measurement. For the proof, we can use the same construction to encode a message in $[e^{\Omega(d^2)}]$. In order to decode this message with high probability, a learner needs to share at least $\Omega(d^2)$ nats of information with the uniform encoder. Then, we need to show that, for a Pauli channel close to the maximally depolarizing channel, at each step, reapplying the channel $m \geq 1$ times even intertwined with unital operations can only add a noise and doesn't help to extract useful information: the amount of correlation between the encoder and decoder increases by at most $\mathcal{O}(\varepsilon^{2m})$ nats. This result is stated in Theorem 3.1. Furthermore, if the (adaptive) algorithm could only apply the Pauli channel once per step, it should use at least $\Omega(d^{2.5}/\varepsilon^2)$ measurements if $\varepsilon \leq 1/(20d)$. This result is stated in Theorem 5.1. The strategy of the proof is the same as in the non-adaptive case. When the learner can adapt its choices of input and measurement device depending on the previous observations, we expect that its correlations with the uniform encoder will increase by more than $\mathcal{O}(\varepsilon^2/d)$ nats per step. Besides the naive upper bound of $\mathcal{O}(\varepsilon^2)$ on this correlation, we show that if the learner uses the channel once per step, it can only increase its correlation with the encoder by at most $\mathcal{O}(k\varepsilon^4/d^3)$ nats at step k. For this, we change the previous construction and use normalized Gaussian random variables in the Pauli channel's coefficients. The Gaussian variables allow us to break the dependency between the probability of measurements at different steps by applying Gaussian integration by parts on an upper bound of the mutual information.

Related work Learning Pauli channels has been considered in different settings. Flammia and Wallman (2020) provides an algorithm for learning Pauli channels in ℓ_2-norm using $\tilde{\mathcal{O}}(d/\varepsilon^2)$ measurements. This implies an upper bound of $\tilde{\mathcal{O}}(d^3/\varepsilon^2)$ for learning Pauli channel in ℓ_1-norm. For completeness, we reproduce this argument in App. B. In this article, we address an open question posed in Flammia and Wallman (2020) about a lower bound for learning Pauli channels. In particular we show that the algorithm of Flammia and Wallman (2020) is optimal up to logarithmic factors. Moreover, learning a Pauli channel in ℓ_∞-norm was shown to be solved with $\tilde{\Theta}(1/\varepsilon^2)$ measurements in Flammia and O'Donnell (2021) and this is optimal up to logarithmic factors. The previous settings did not allow for ancillas. The work

of Chen, Zhou, et al. (2022) shows an exponential separation between allowing and not allowing ancilla for estimating the Pauli eigenvalues in ℓ_∞-norm. Using the Parseval–Plancherel identity, their upper bound can be translated to learning in ℓ_1-norm with an n-qubit ancilla assisted algorithm using $\tilde{\mathcal{O}}(d^2/\varepsilon^2)$ measurements. However, our lower bounds don't apply in this setting since we only consider ancilla-free strategies. We also note that Chen, Zhou, et al. (2022) shows a lower bound of $\Omega(d^{\frac{1}{3}}/\varepsilon^2)$ measurements to learn the eigenvalues of $\mathcal{P}$ in the adaptive setting up to ε in ℓ_∞-norm and $\Omega(d/\varepsilon^2)$ in the non-adaptive setting. However, this is a different figure of merit than the one we consider.

Other noteworthy protocols to learn quantum channels include gate set tomography Blume-Kohout et al. (2013) and techniques based on compressed sensing Roth et al. (2018). Although they apply to more general classes of channels, they do not offer quantitative or qualitative advantages over randomized benchmarking in the setting of Pauli channels. We refer the readers to the survey Montanaro and de Wolf (2013) for results on testing quantum channels and to Eisert et al. (2020) for quantum channel learning. The question of optimal quantum channel tomography remains open. In contrast, for the state tomography problem, it is known that the optimal copy complexity for incoherent strategies in both adaptive and non-adaptive settings is $\Theta(d^3/\varepsilon^2)$ Chen, Huang, Li, Liu, and Sellke (2022); Haah, Harrow, Ji, Wu, and Yu (2016). By reduction, optimal state tomography implies trivial upper bound of $\mathcal{O}(d^8/\varepsilon^2)$ and lower bound of $\Omega(d^3/\varepsilon^2)$ for channel tomography. However, if we add the Pauli structure to the channel, our lower bound along with the upper bound of Flammia and Wallman (2020) show that the optimal complexity is the same as state tomography complexity.

2 Preliminaries

Let $d = 2^n$ be the dimension of an n-qubit system. We use the notation $[d] := \{1, \ldots, d\}$. We adopt the bra-ket notation: a column vector is denoted $|\phi\rangle$ and its adjoint is denoted $\langle\phi| = |\phi\rangle^\dagger$. With this notation, $\langle\phi|\psi\rangle$ is the dot product of the vectors ϕ and ψ and, for a unit vector $|\phi\rangle \in \mathbf{S}^d$, $|\phi\rangle\langle\phi|$ is the rank-1 projector on the space spanned by the vector ϕ. The canonical basis $\{e_i\}_{i\in[d]}$ is denoted $\{|i\rangle\}_{i\in[d]} := \{|e_i\rangle\}_{i\in[d]}$. A quantum state is a positive semi-definite Hermitian matrix of trace 1. We will denote the identity matrix by $\mathbb{I}_d \in \mathbb{C}^{d\times d}$ and by $\mathrm{id}_d : \mathbb{C}^{d\times d} \to \mathbb{C}^{d\times d}$ the identity map. We will omit the the d subscript if the dimension is clear from context. A quantum channel is a map $\mathcal{N} : \mathbb{C}^{d\times d} \to \mathbb{C}^{d\times d}$ of the form $\mathcal{N}(\rho) = \sum_k A_k \rho A_k^\dagger$ where the Kraus operators $\{A_k\}_k$ satisfy $\sum_k A_k^\dagger A_k = \mathbb{I}$. A map $\mathcal{N}$ is a quantum channel if, and only if, it is:

- **completely positive:** for all $\rho \in \mathbb{C}^{d^2\times d^2}, \rho \succcurlyeq 0$, $\mathrm{id}_d \otimes \mathcal{N}(\rho) \succcurlyeq 0$ and
- **trace preserving:** for all $\rho \in \mathbb{C}^{d\times d}$, $\mathrm{tr}(\mathcal{N}(\rho)) = \mathrm{tr}(\rho)$.

If the quantum channel $\mathcal{N}$ satisfies further $\mathcal{N}(\mathbb{I}) = \mathbb{I}$, it is called *unital*.

We define the diamond distance between two quantum channels $\mathcal{N}$ and $\mathcal{M}$ as the diamond norm of their difference:

$$\|\mathcal{N} - \mathcal{M}\|_\diamond = \max_{\phi:\langle\phi|\phi\rangle=1} \|\mathrm{id}_d \otimes (\mathcal{N} - \mathcal{M})(|\phi\rangle\langle\phi|)\|_1$$

where the the Schatten 1-norm of a matrix M is defined as $\|M\|_1 = \mathrm{tr}\left(\sqrt{M^\dagger M}\right)$.

Pauli channels are a special quantum channel whose Kraus operators are weighted Pauli operators. Formally, a Pauli quantum channel $\mathcal{P}$ can be written as follows:

$$\mathcal{P}(\rho) = \sum_{P\in\{\mathbb{I},X,Y,Z\}^{\otimes n}} p(P) P\rho P \tag{1}$$

where the Pauli matrices $\mathbb{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, $Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$ and $Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ and $\{p(P)\}_{P\in\{\mathbb{I},X,Y,Z\}^{\otimes n}}$ is a probability distribution. Let $\mathbb{P}_n = \{\mathbb{I}, X, Y, Z\}^{\otimes n}$ be the set of Pauli operators. The elements of $\mathbb{P}_n$ either commute or anti commute. Let P and Q be two Pauli operators, we have $PQ = (-1)^{P.Q} QP$ where $P.Q = 0$ if $[P, Q] = 0$ and $P.Q = 1$ otherwise.

We consider the Pauli channel tomography problem which consists of learning a Pauli channel in the diamond norm. Given a precision parameter $\varepsilon > 0$, the goal is to construct a Pauli channel $\tilde{\mathcal{P}}$ satisfying with at least a probability 2/3:

$$\|\mathcal{P} - \tilde{\mathcal{P}}\|_\diamond \leq \varepsilon.$$

An algorithm $\mathcal{A}$ is 1/3-correct for this problem if it outputs a Pauli channel ε-close to $\mathcal{P}$ with a probability of error at most 1/3. We choose to learn in the diamond norm because it characterizes the minimal error probability to distinguish between two quantum channels when auxiliary systems are allowed Watrous (2018). Since the diamond norm between two Pauli channels is exactly twice the TV-distance between their corresponding probability distributions Magesan et al. (2012), approximating the Pauli channel $\mathcal{P}$ in diamond norm is equivalent to approximating the probability distribution p in TV-distance. The latter is defined for two probability distributions p and q on $[d]$ as follows:

$$\mathrm{TV}(p,q) = \frac{1}{2}\sum_{i=1}^{d} |p_i - q_i|.$$

The learner can only extract classical information from the unknown Pauli channel $\mathcal{P}$ by performing a measurement on the output state. Throughout the paper, we only consider unentangled or incoherent measurements. That is, the learner can only measure with an n-qubit measurement device and auxilliary qubits or measuring multiple copies at once is not allowed. This restriction is natural for the problem at hand, given that performing measurements on multiple copies requires a quantum memory.

More precisely, an n-qubit measurement is defined by a POVM (positive operator-valued measure) with a finite number of elements: this is a set of positive semi-definite matrices $\mathcal{M} = \{M_i\}_i$ acting on the Hilbert space $\mathbb{C}^{2^n}$ and satisfying $\sum_i M_i = \mathbb{I}$. Each element M_i in the POVM $\mathcal{M}$ is associated with the outcome i. The tuple $\{\mathrm{tr}(\rho M_i)\}_i$ is non-negative and sums to 1: it thus defines a probability. Born's rule Born (1926) says that the probability that the measurement on a quantum state ρ using the POVM $\mathcal{M}$ will output i is exactly $\mathrm{tr}(\rho M_i)$.

For an integer $t \geq 1$, we say that the learner is at step t if it has already performed $t-1$ measurements. With this definition, the total number of steps is exactly the total number of measurements. However, depending on the setting, the total number of channel uses could be different than the total number of steps. The goal of the paper is to show lower bounds on the total number of steps as well as the total number of the channel uses.

A simple example we can propose to see the effect of reusing the channel is the following test: $H_0 : \mathcal{P}(\rho) = \rho$ vs $H_1 : \mathcal{P}(\rho) = (1-\varepsilon)\rho + \varepsilon \mathrm{tr}(\rho)\frac{\mathbb{I}}{d}$. We can choose as input the rank one state $\rho = |0\rangle\langle 0|$. Under the null hypothesis H_0, the channel does not affect the state $|0\rangle\langle 0|$. On the other hand, under H_1, if we apply the channel $\mathcal{P}$ a number $m \in \mathbb{N}^*$ times the resulting quantum state is $\mathcal{P}^{(m)}(\rho) = (1-\varepsilon)^m |0\rangle\langle 0| + (1-(1-\varepsilon)^m)\frac{\mathbb{I}}{d}$. Hence, if we measure with the POVM $\mathcal{M} = \{|0\rangle\langle 0|, I - |0\rangle\langle 0|\}$ of outcomes 0 and 1 respectively, under H_0 we will always see 0 while under H_1, we will see 0 with probability roughly $(1-\varepsilon)^m$. Therefore, we can achieve a confidence δ with only *one measurement* but the channel is reused $\log(1/\delta)/\varepsilon$-times[1]. However, if we don't allow reusing the channel, then the number of measurements needed is approximately $\log(1/\delta)/\varepsilon$.

3 A general lower bound on the number of steps required for Pauli channel tomography

In this section, we consider the problem of learning a Pauli quantum channel using incoherent measurements. Unlike the usual state tomography problem for which at each step the learner can only choose the measurement device, for quantum channels, the learner has additional choices. First, in every setting, the learner can choose the input quantum state at each step. This choice can be done in an adaptive fashion: the input quantum state at a given step can be chosen depending on the previous observations (and of course the previous input states and POVMs). Second, the learner has the ability to reuse the Pauli quantum channel as much as it wants before performing the measurement. This is specific to quantum process tomography too since for state tomography using incoherent measurements, once a measurement is performed, the post-measurement quantum state is usually useless. Finally, the learner can intertwine arbitrary unital quantum channels and the unknown Pauli quantum channel before measuring the output of this (possibly long) sequence of quantum channels. We propose a lower bound on the number of steps required for the Pauli channel tomography problem in this general setting.

Recall that Pauli channel tomography problem is equivalent to learning the probability p in the TV-distance. Mainly, the learner would like to construct a probability distribution $\hat{p}$ on the set of Pauli operators $\mathbb{P}_n$ satisfying with at least a probability 2/3:

$$\mathrm{TV}(p,\hat{p}) \leq \varepsilon$$

[1] all the logs are taken in base e so the information is measured in "nats".

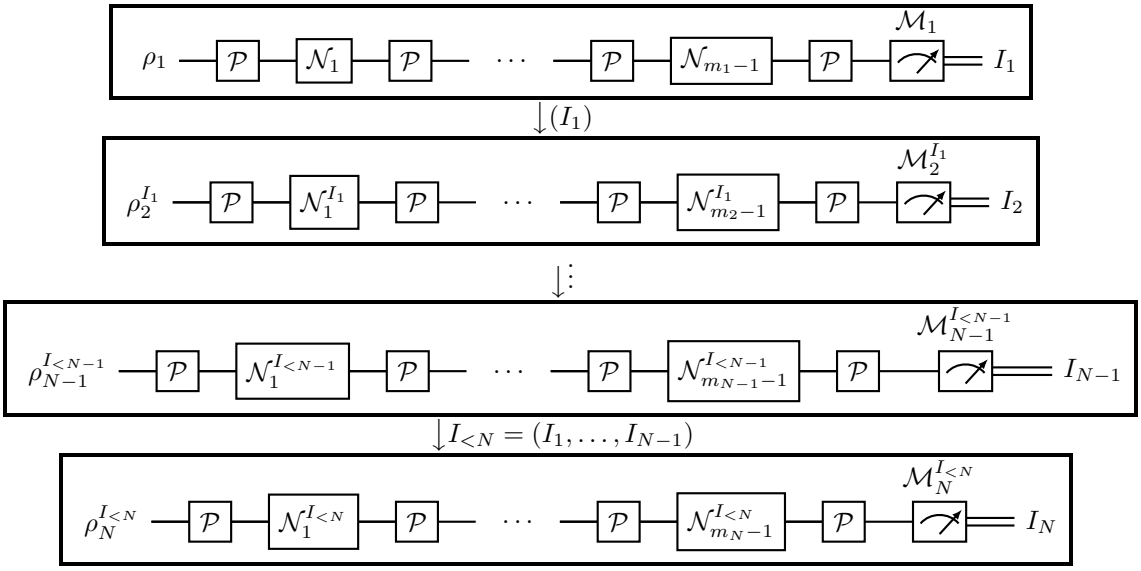

Figure 1: Illustration of an adaptive strategy for learning Pauli channel.

with as few steps as possible.

Let N be a sufficient number of steps to learn $\mathcal{P}$ as defined in Eq. (1). At step $t \in [N]$, the learner has the ability to choose an input quantum state ρ_t, the number $m_t \geq 1$ of uses of the quantum channel $\mathcal{P}$, the unital quantum channels applied in between $\mathcal{N}_1, \ldots, \mathcal{N}_{m_t-1}$ and the POVM $\mathcal{M}_t$ for measuring the output quantum state ρ_t^{output}:

$$\rho_t^{\text{output}} = \underbrace{\mathcal{P} \circ \mathcal{N}_{m_t-1} \circ \mathcal{P} \circ \cdots \circ \mathcal{P} \circ \mathcal{N}_1 \circ \mathcal{P}(\rho_t)}_{m_t \text{ times}}.$$

All these elements can be chosen adaptively: the choice of $m_t, \rho_t, \mathcal{N}_1, \ldots, \mathcal{N}_{m_t-1}$ and $\mathcal{M}_t$ can depend on the previous observations $I_1, \ldots, I_{t-1}$ (see Fig.1 for an illustration). However, to not overload the expressions we don't add the subscript $I_1, \ldots, I_{t-1}$ on $m_t, \rho_t, \mathcal{N}_1, \ldots, \mathcal{N}_{m_t-1}$ or $\mathcal{M}_t$. By Born's rule, performing a measurement on the output quantum state ρ_t^{output} using the POVM $\mathcal{M}_t = \{M_i^t\}_{i \in \mathcal{I}}$ is equivalent to sampling from the probability distribution

$$x \sim \{\text{tr}(\rho_t^{\text{output}} M_i^t)\}_{i \in \mathcal{I}}.$$

Note that unital operations cannot be used to prepare a new state and thus have a free step. In fact, applying a unital operation after a noisy Pauli channel cannot prepare a rank-1 state for example. We propose the following lower bound on the number of steps N.

Theorem 3.1. *The problem of Pauli channel tomography using incoherent measurements requires a number of steps satisfying:*

$$N = \Omega\left(\frac{d^2}{\varepsilon^2}\right).$$

This Theorem shows that no matter how often the learner reuses the quantum Pauli channel intertwined with other unital quantum channels on each step, the global number of steps should be exponential in the number of qubits. This can be explained by the fact that a Pauli channel adds noise to the input state, so reapplying it makes the input state more noisy and can't help to extract more information. Although, as we remark later, this lower bound is weaker in the dependency on the dimension d compared to the non-adaptive case, it has the particularity of not depending on the number of uses of the Pauli channel.

Proof. We will break down the proof into several steps.

Construction of the family $\mathcal{F}$ We start by describing a general construction of a big family $\mathcal{F} = \{\mathcal{P}_x\}_{x \in [\![1,M]\!]}$ constituted of quantum Pauli channels satisfying for all $x \neq y \in [\![1, M]\!] : \text{TV}(p_x, p_y) \geq \varepsilon$, we

say that the family $\mathcal{F}$ is ε-separated. These quantum channels have the form for $x \in [\![1, M]\!]$:

$$\mathcal{P}_x(\rho) = \sum_{P \in \{\mathbb{I},X,Y,Z\}^{\otimes n}} p_x(P) P\rho P = \sum_{P \in \{\mathbb{I},X,Y,Z\}^{\otimes n}} \left(\frac{1 + 4\alpha_x(P)\varepsilon}{d^2}\right) P\rho P \tag{2}$$

where $\alpha_x(P) = \pm 1$ to be chosen randomly so that $\alpha_x(P) = -\alpha_x(\sigma(P))$ for some matching σ of $\{\mathbb{I}, X, Y, Z\}^{\otimes n}$ [2]. Suppose that we have already constructed an ε-separated family of Pauli quantum channels $\mathcal{F} = \{\mathcal{P}_x\}_x$ of cardinality M. We show that we can add another element to this family as long as $M < e^{cd^2}$ for some sufficiently small constant c. For this, we choose $\alpha(P) = -\alpha(\sigma(P)) = \pm 1$ with probability $1/2$ each. This α leads to a quantum channel $\mathcal{P}(\rho) = \sum_{P \in \{\mathbb{I},X,Y,Z\}^{\otimes n}} \left(\frac{1+4\alpha(P)\varepsilon}{d^2}\right) P\rho P$. Then, we control the probability that the corresponding Pauli quantum channel isn't ε-far from the family $\mathcal{F}$. By the union bound and Chernoff-Hoeffding inequality Hoeffding (1963):

$$\begin{aligned}
\mathbb{P}\left(\exists \mathcal{P}_x \in \mathcal{F} : \mathrm{TV}(p, p_x) < \varepsilon\right) &\leq \sum_{x=1}^{M} \mathbb{P}\left(\sum_{P \in \mathbb{P}_n} |p(P) - p_x(P)| < 2\varepsilon\right) \\
&= \sum_{x=1}^{M} \mathbb{P}\left(\sum_{P \in \mathbb{P}_n} 4|\alpha(P) - \alpha_x(P)| < 2d^2\right) \\
&= \sum_{x=1}^{M} \mathbb{P}\left(\sum_{P \in \mathbb{P}_n} \mathbb{1}_{\alpha(P) \neq \alpha_x(P)} < \frac{d^2}{4}\right) \\
&= \sum_{x=1}^{M} \mathbb{P}\left(\frac{2}{d^2} \sum_{P \in \mathbb{P}_n/\sigma} \mathbb{1}_{\alpha(P) \neq \alpha_x(P)} - \mathbb{E}\left(\mathbb{1}_{\alpha(P) \neq \alpha_x(P)}\right) < -\frac{1}{4}\right) \\
&= \sum_{x=1}^{M} \exp(-2(d^2/2)(1/4)^2) = M \exp(-d^2/16)
\end{aligned}$$

which is strictly smaller than 1 if $M < e^{d^2/16}$. So far, we have proven the following lemma:

Lemma 3.2. *There exists an ε-separated family of quantum Pauli channels of the form 2 and size at least $e^{d^2/16}$.*

Hence, we can use this family to encode a message $X \sim \mathrm{Unif}\,[\![1, M]\!]$ to a quantum Pauli channel $\mathcal{P} = \mathcal{P}_X$ in the family constructed above. The decoder receives this unknown quantum Pauli channel, chooses its inputs states and performs incoherent measurements possibly after many uses of the channel intertwined with arbitrary unital quantum channels, and learns it to within a precision $\varepsilon/2$. It thus produces a Pauli quantum channel $\hat{\mathcal{P}}$ corresponding to a probability distribution $\hat{p}$ satisfying, with a probability at least $2/3$, $\mathrm{TV}(\hat{p}, p_X) \leq \varepsilon/2$. Since the family of probability distributions $\{p_x\}_{x \in [M]}$ is ε-separated, there is only one $\hat{X}$ such that $\mathrm{TV}(\hat{p}, p_{\hat{X}}) \leq \varepsilon/2$. Therefore a $1/3$-correct algorithm can decode with a probability of failure at most $1/3$. By Fano's inequality, the encoder and decoder should share at least $\Omega(\log(M)) = \Omega(d^2)$ nats of information.

Lemma 3.3 (Fano (1961))**.** *The mutual information between the index of the actual channel X and the estimated index $\hat{X}$ is at least*

$$\mathcal{I}(X : \hat{X}) \geq 2/3 \log(M) - \log(2) = \Omega(d^2).$$

Then we show that no algorithm can extract more than $\mathcal{O}(\varepsilon^2)$ nats of information at each step. For this, recall that X is the uniform random variable on the set $[\![1, M]\!]$ representing the encoder and denote by $I_1, \ldots, I_N$ the observations of the decoder or the $1/3$-correct algorithm. The Data-Processing inequality implies:

$$\mathcal{I}(X : \hat{X}) \leq \mathcal{I}(X : I_1, \ldots, I_N).$$

[2] in order to have $\sum_{P \in \mathbb{P}_n} \alpha_x(P) = 0$ and thus a quantum channel for $\varepsilon \leq 1/4$.

Moreover, if we denote by $I_{\leq k-1} := (I_1, \ldots, I_{k-1})$ for all $1 \leq k \leq N$, the chain rule of mutual information guves:

$$\mathcal{I}(X : I_1, \ldots, I_N) = \sum_{k=1}^{N} \mathcal{I}(X : I_k | I_{\leq k-1})$$

where $\mathcal{I}(X : I_k | I_{\leq k-1})$ denotes the conditional mutual information between X and I_k giving $I_{\leq k-1}$. We claim that every conditional mutual information $\mathcal{I}(X : I_k | I_{\leq k-1})$ can be upper bounded by $\mathcal{O}(\varepsilon^2)$. To prove this claim, we prove first a general upper bound on the conditional mutual information.

At step $t \in [N]$, the 1/3-correct algorithm used by the decoder chooses the input state ρ_t, uses the unknown quantum Pauli channel $\mathcal{P}$ $m_t \geq 1$ times, eventually intertwines the $\mathcal{P}$ with unital quantum channels $\mathcal{N}_1^t, \mathcal{N}_2^t, \ldots, \mathcal{N}_{m_t-1}^t$ and finally measures the output with a POVM $\mathcal{M}_t = \{\lambda_i^t |\phi_i^t\rangle\langle\phi_i^t|\}_{i \in \mathcal{I}_t}$ where $\langle\phi_i^t|\phi_i^t\rangle = 1$ and $\sum_i \lambda_i^t |\phi_i^t\rangle\langle\phi_i^t| = I$. Note that this implies $\sum_i \lambda_i^t = d$. Observe that we can always reduce the measurement with a general POVM $\mathcal{M}$ to the measurement with such a POVM by taking the projectors on the eigenvectors of each element of the POVM $\mathcal{M}$ weighted by the corresponding eigenvalues. We denote by $\mathcal{P}^{m_t}(\rho_t) = \underbrace{\mathcal{P} \circ \mathcal{N}_{m_t-1}^t \circ \mathcal{P} \ldots \mathcal{P} \circ \mathcal{N}_1^t \circ \mathcal{P}(\rho_t)}_{m_t \text{ times}}$ the quantum channel applied to the input quantum state ρ_t. We denote by q the joint distribution of $(X, I_1, \ldots, I_N)$:

$$q(x, i_1, \ldots, i_N) = \frac{1}{M} \prod_{t=1}^{N} \lambda_{i_t}^t \left\langle \phi_{i_t}^t \right| \mathcal{P}_x^{m_t}(\rho_t) \left| \phi_{i_t}^t \right\rangle.$$

We use the usual notation of marginals by ignoring the indices on which we marginalize. For instance, for all adaptive algorithms, for all $1 \leq k \leq N$, we have:

$$\begin{aligned}
q_{\leq k}(x, i_1, \ldots, i_k) &= \sum_{i_{k+1}, \ldots, i_N} \frac{1}{M} \prod_{t=1}^{N} \lambda_{i_t}^t \left\langle \phi_{i_t}^t \right| \mathcal{P}_x^{m_t}(\rho_t) \left| \phi_{i_t}^t \right\rangle \\
&= \frac{1}{M} \prod_{t=1}^{k} \lambda_{i_t}^t \left\langle \phi_{i_t}^t \right| \mathcal{P}_x^{m_t}(\rho_t) \left| \phi_{i_t}^t \right\rangle \prod_{t=k+1}^{N} \sum_{i_t} \lambda_{i_t}^t \left\langle \phi_{i_t}^t \right| \mathcal{P}_x^{m_t}(\rho_t) \left| \phi_{i_t}^t \right\rangle \\
&= \frac{1}{M} \prod_{t=1}^{k} \lambda_{i_t}^t \left\langle \phi_{i_t}^t \right| \mathcal{P}_x^{m_t}(\rho_t) \left| \phi_{i_t}^t \right\rangle \prod_{t=k+1}^{N} \operatorname{tr}(\mathcal{P}_x^{m_t}(\rho_t)) \\
&= \frac{1}{M} \prod_{t=1}^{k} \lambda_{i_t}^t \left\langle \phi_{i_t}^t \right| \mathcal{P}_x^{m_t}(\rho_t) \left| \phi_{i_t}^t \right\rangle.
\end{aligned}$$

We sometimes abuse the notation and use q instead of $q_{\leq k}$ when it is clear from the context. In order to simplify the expressions, we introduce the notation $u_{i_k}^{k,x} = \left\langle \phi_{i_k}^k \right| d\mathcal{P}_x^{m_k}(\rho_k) - \mathbb{I} \left| \phi_{i_k}^k \right\rangle$. Note that for adaptive strategies the vectors $\left|\phi_{i_k}^k\right\rangle = \left|\phi_{i_k}^t(i_{<k})\right\rangle$ and the states $\rho_k = \rho_k(i_{<k})$ depend on the previous observations $i_{<k} = (i_1, \ldots, i_{k-1})$ for all $k \in [N]$. Then the general upper bound on the conditional mutual information is:

Lemma 3.4. *Let $1 \leq k \leq N$ and $u_{i_k}^{k,x} = \left\langle \phi_{i_k}^k(i_{<k}) \right| d\mathcal{P}_x^{m_k}(\rho_k(i_{<k})) - \mathbb{I} \left| \phi_{i_k}^k(i_{<k}) \right\rangle$. We have for adaptive strategies:*

$$\mathcal{I}(X : I_k | I_{\leq k-1}) \leq 3\mathbb{E}_x \mathbb{E}_{i \sim q_{\leq k-1}} \left[\sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,x})^2 \right].$$

Moreover, for non-adaptive strategies $u_{i_k}^{k,x} = \left\langle \phi_{i_k}^k \right| d\mathcal{P}_x^{m_k}(\rho_k) - \mathbb{I} \left| \phi_{i_k}^k \right\rangle$ and:

$$\mathcal{I}(X : I_k | I_{\leq k-1}) \leq 3\mathbb{E}_x \left[\sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,x})^2 \right].$$

Proof. We can remark that, for all $1 \leq k \leq N$, $q(x, i_{\leq k}) = \lambda_{i_k}^k \left(\frac{1+u_{i_k}^{k,x}}{d}\right) q(x, i_{\leq k-1})$ thus

$$\begin{aligned}
\frac{q(x, i_k|i_{\leq k-1})}{q(x|i_{\leq k-1})q(i_k|i_{\leq k-1})} &= \frac{q(x, i_{\leq k})q(i_{\leq k-1})}{q(x, i_{\leq k-1})q(i_{\leq k})} = \frac{\lambda_{i_k}^k \left(\frac{1+u_{i_k}^{k,x}}{d}\right) q(x, i_{\leq k-1})q(i_{\leq k-1})}{q(x, i_{\leq k-1}) \sum_y q(y, i_{\leq k})} \\
&= \frac{\lambda_{i_k}^k \left(\frac{1+u_{i_k}^{k,x}}{d}\right) q(i_{\leq k-1})}{\sum_y q(y, i_{\leq k})} = \frac{\lambda_{i_k}^k \left(\frac{1+u_{i_k}^{k,x}}{d}\right) q(i_{\leq k-1})}{\sum_y q(y, i_{\leq k-1}) \lambda_{i_k}^k \left(\frac{1+u_{i_k}^{k,y}}{d}\right)} \\
&= \frac{(1+u_{i_k}^{k,x})q(i_{\leq k-1})}{\sum_y q(y, i_{\leq k-1})(1+u_{i_k}^{k,y})} = \frac{(1+u_{i_k}^{k,x})}{\sum_y q(y|i_{\leq k-1})(1+u_{i_k}^{k,y})}.
\end{aligned}$$

Therefore by Jensen's inequality:

$$\begin{aligned}
\mathcal{I}(X : I_k|I_{\leq k-1}) &= \mathbb{E}\left(\log\left(\frac{q(x, i_k|i_{\leq k-1})}{q(x|i_{\leq k-1})q(i_k|i_{\leq k-1})}\right)\right) \\
&= \mathbb{E}\left(\log\left(\frac{(1+u_{i_k}^{k,x})}{\sum_y q(y|i_{\leq k-1})(1+u_{i_k}^{k,y})}\right)\right) \\
&\leq \mathbb{E}\left(\log(1+u_{i_k}^{k,x}) - \sum_y q(y|i_{\leq k-1}) \log(1+u_{i_k}^{k,y})\right) \\
&= \mathbb{E}\left(\log(1+u_{i_k}^{k,x})\right) - \sum_y \mathbb{E}\left(q(y|i_{\leq k-1}) \log(1+u_{i_k}^{k,y})\right).
\end{aligned}$$

The first term can be upper bounded using the inequality $\log(1+x) \leq x$ verified for all $x \in (-1, +\infty)$:

$$\begin{aligned}
\mathbb{E}\left(\log(1+u_{i_k}^{k,x})\right) &= \mathbb{E}_{x,i\sim q} \log(1+u_{i_k}^{k,x}) \leq \mathbb{E}_{x,i\sim q} u_{i_k}^{k,x} = \mathbb{E}_{x,i\sim q_{\leq k}} u_{i_k}^{k,x} \\
&= \mathbb{E}_{x,i\sim q_{\leq k-1}} \sum_{i_k} \frac{\lambda_{i_k}^k}{d}(1+u_{i_k}^{k,x})u_{i_k}^{k,x} = \mathbb{E}_{x,i\sim q_{\leq k-1}} \sum_{i_k} \frac{\lambda_{i_k}^k}{d}(u_{i_k}^{k,x})^2
\end{aligned}$$

because $\sum_{i_k} \frac{\lambda_{i_k}^k}{d} u_{i_k}^{k,x} = \mathrm{tr}(d\mathcal{P}_x^{m_t}(\rho_t) - \mathbb{I}) = 0$.The second term can be upper bounded using the inequality $-\log(1+x) \leq -x + x^2/2$ verified for all $x \in (-1/2, +\infty)$:

$$\begin{aligned}
&\mathbb{E}\left(-\sum_y q(y|i_{\leq k-1}) \log(1+u_{i_k}^{k,y})\right) = -\sum_y \mathbb{E}_{x,i\sim q} q(y|i_{\leq k-1}) \log(1+u_{i_k}^{k,y}) \\
&= -\sum_y \mathbb{E}_{x,i\sim q_{\leq k-1}} q(y|i_{\leq k-1}) \sum_{i_k} \frac{\lambda_{i_k}^k}{d}(1+u_{i_k}^{k,x}) \log(1+u_{i_k}^{k,y}) \\
&\leq \sum_y \mathbb{E}_{x,i\sim q_{\leq k-1}} q(y|i_{\leq k-1}) \sum_{i_k} \frac{\lambda_{i_k}^k}{d}(1+u_{i_k}^{k,x})(-u_{i_k}^{k,y} + (u_{i_k}^{k,y})^2/2) \\
&= \sum_y \mathbb{E}_{x,i\sim q_{\leq k-1}} q(y|i_{\leq k-1}) \sum_{i_k} \frac{\lambda_{i_k}^k}{d}(-u_{i_k}^{k,y} - u_{i_k}^{k,y}u_{i_k}^{k,x} + (u_{i_k}^{k,y})^2/2 + u_{i_k}^{k,x}(u_{i_k}^{k,y})^2/2) \\
&\leq \sum_y \mathbb{E}_{x,i\sim q_{\leq k-1}} q(y|i_{\leq k-1}) \sum_{i_k} \frac{\lambda_{i_k}^k}{d}((u_{i_k}^{k,x})^2 + (u_{i_k}^{k,y})^2) \\
&= 2\sum_y \mathbb{E}_{x,i\sim q_{\leq k-1}} q(y|i_{\leq k-1}) \sum_{i_k} \frac{\lambda_{i_k}^k}{d}(u_{i_k}^{k,x})^2 = 2\mathbb{E}_{x,i\sim q_{\leq k-1}} \sum_{i_k} \frac{\lambda_{i_k}^k}{d}(u_{i_k}^{k,x})^2.
\end{aligned}$$

Since the conditional mutual is upper bounded by the sum of these two terms, the upper bound on the conditional mutual information follows. □

The following Lemma permits to conclude the upper bound on the conditional mutual information and thus the upper bound on the mutual information.

Lemma 3.5. *Let $m \geq 1$, $\mathcal{N}_1, \ldots, \mathcal{N}_{m-1}$ be unital quantum channels and $\mathcal{P}$ be a Pauli quantum channel in the family $\mathcal{F}$. We have for all quantum states ρ and vectors $|\phi\rangle \in \mathbf{S}^d$:*

$$|\langle\phi| d\mathcal{P}\mathcal{N}_{m-1}\mathcal{P}\ldots\mathcal{P}\mathcal{N}_1\mathcal{P}(\rho)|\phi\rangle - 1| \leq (4\varepsilon)^m.$$

Proof. For $x \in [\![1, M]\!]$, we define the map $\mathcal{M}_x$ verifying the following equality:

$$\mathcal{M}_x(\rho) = \mathcal{P}_x(\rho) - \mathrm{tr}(\rho)\frac{\mathbb{I}}{d} = \sum_{P\in\{\mathbb{I},X,Y,Z\}^{\otimes n}} \frac{4\alpha_x(P)\varepsilon}{d^2} P\rho P,$$

where we have used the fact (see Lemma A.2) that for all ρ:

$$\sum_{P\in\{\mathbb{I},X,Y,Z\}^{\otimes n}} P\rho P = d\mathrm{tr}(\rho)\mathbb{I}.$$

Note that $\mathrm{tr}(\mathcal{M}_x(\rho)) = \mathrm{tr}(\mathcal{P}_x(\rho)) - \mathrm{tr}(\rho)\mathrm{tr}(\frac{\mathbb{I}}{d}) = \mathrm{tr}(\rho) - \mathrm{tr}(\rho) = 0$. Applying a unital quantum channel $\mathcal{N}$ between two quantum channels $\mathcal{P}_x$ can be seen as :

$$\begin{aligned}\mathcal{P}_x\mathcal{N}\mathcal{P}_x(\rho) &= \mathcal{P}_x\mathcal{N}\left(\mathrm{tr}(\rho)\frac{\mathbb{I}}{d} + \mathcal{M}_x(\rho)\right) = \mathcal{P}_x\left(\mathrm{tr}(\rho)\frac{\mathbb{I}}{d} + \mathcal{N}\mathcal{M}_x(\rho)\right)\\ &= \mathrm{tr}(\rho)\frac{\mathbb{I}}{d} + \mathcal{M}_x(\frac{\mathbb{I}}{d} + \mathcal{N}\mathcal{M}_x(\rho)) = \mathrm{tr}(\rho)\frac{\mathbb{I}}{d} + \mathcal{M}_x\mathcal{N}\mathcal{M}_x(\rho)\end{aligned}$$

because $\mathrm{tr}(\mathcal{N}\mathcal{M}_x(\rho)) = \mathrm{tr}(\mathcal{M}_x(\rho)) = 0$ and

$$\mathcal{M}_x(I) = \sum_{P\in\{\mathbb{I},X,Y,Z\}^{\otimes n}} \frac{4\alpha_x(P)\varepsilon}{d^2}\mathbb{I} = \sum_{P\in\{\mathbb{I},X,Y,Z\}^{\otimes n}/\sigma} \frac{4\alpha_x(P)\varepsilon}{d^2}\mathbb{I} + \frac{4\alpha_x(\sigma(P))\varepsilon}{d^2}\mathbb{I} = 0.$$

By induction, we generalize this equality to m applications of the Pauli channel $\mathcal{P}_x$:

$$\underbrace{\mathcal{P}_x\mathcal{N}_{m-1}\mathcal{P}_x\ldots\mathcal{P}_x\mathcal{N}_1\mathcal{P}_x(\rho)}_{m \text{ times}} = \mathrm{tr}(\rho)\frac{\mathbb{I}}{d} + \underbrace{\mathcal{M}_x\mathcal{N}_{m-1}\mathcal{M}_x\ldots\mathcal{M}_x\mathcal{N}_1\mathcal{M}_x(\rho)}_{m \text{ times}}$$

Therefore

$$\begin{aligned}\langle\phi| d\mathcal{P}\mathcal{N}_{m-1}\mathcal{P}\ldots\mathcal{P}\mathcal{N}_1\mathcal{P}(\rho)|\phi\rangle &= \langle\phi| I + d\mathcal{M}\mathcal{N}_{m-1}\mathcal{M}\ldots\mathcal{M}\mathcal{N}_1\mathcal{M}(\rho)|\phi\rangle\\ &= 1 + d\langle\phi|\mathcal{M}\mathcal{N}_{m-1}\mathcal{M}\ldots\mathcal{M}\mathcal{N}_1\mathcal{M}(\rho)|\phi\rangle.\end{aligned}$$

On the other hand, for all vectors $|\phi\rangle \in \mathbf{S}^d$ and Hermitian matrices $X = \sum_i \lambda_i |\phi_i\rangle\langle\phi_i|$ we have: $|\langle\phi| X |\phi\rangle| = |\sum_i \lambda_i |\langle\phi|\phi_i\rangle|^2| \leq \sum_i |\lambda_i||\langle\phi|\phi_i\rangle|^2 = \langle\phi| |X| |\phi\rangle$ therefore using Lemma A.2:

$$\begin{aligned}|\langle\phi|\mathcal{M}(X)|\phi\rangle| &= \left|\langle\phi| \sum_{P\in\mathbb{P}_n} \frac{4\alpha(P)\varepsilon}{d^2} PXP |\phi\rangle\right| \leq \frac{4\varepsilon}{d^2}\sum_{P\in\mathbb{P}_n} |\langle\phi| PXP |\phi\rangle|\\ &\leq \frac{4\varepsilon}{d^2}\sum_{P\in\mathbb{P}_n}\langle\phi| P|X|P|\phi\rangle = \frac{4\varepsilon}{d^2}\langle\phi| d\mathrm{tr}|X|\mathbb{I}|\phi\rangle = \frac{4\varepsilon}{d}\mathrm{tr}|X|,\end{aligned}$$

moreover we can also obtain:

$$\begin{aligned}\mathrm{tr}|\mathcal{M}(X)| &= \left\|\sum_{P\in\mathbb{P}_n} \frac{4\alpha(P)\varepsilon}{d^2} PXP\right\|_1 \leq \sum_{P\in\mathbb{P}_n}\frac{4\varepsilon}{d^2}\|PXP\|_1\\ &= \sum_{P\in\mathbb{P}_n}\frac{4\varepsilon}{d^2}\mathrm{tr}|X| = 4\varepsilon\mathrm{tr}|X|,\end{aligned}$$

and for a quantum channel $\mathcal{N}_j$:

$$\begin{aligned}\mathrm{tr}|\mathcal{N}_j(X)| = \|\mathcal{N}_j(X)\|_1 &= \left\|\sum_i \lambda_i\mathcal{N}_j(|\phi_i\rangle\langle\phi_i|)\right\|_1\\ &\leq \sum_i \|\lambda_i\mathcal{N}_j(|\phi_i\rangle\langle\phi_i|)\|_1 = \sum_i |\lambda_i| = \mathrm{tr}|X|.\end{aligned}$$

Therefore by induction we can prove:

$$\begin{aligned}
|\langle\phi| d\mathcal{P}\mathcal{N}_{m-1}\mathcal{P}\dots\mathcal{P}\mathcal{N}_1\mathcal{P}(\rho)|\phi\rangle - 1| &= d|\langle\phi|\mathcal{M}\mathcal{N}_{m-1}\mathcal{M}\dots\mathcal{M}\mathcal{N}_1\mathcal{M}(\rho)|\phi\rangle| \\
&\leq d\frac{4\varepsilon}{d}\mathrm{tr}|\mathcal{N}_{m-1}\mathcal{M}\dots\mathcal{M}\mathcal{N}_1\mathcal{M}(\rho)| \\
&= 4\varepsilon\mathrm{tr}|\mathcal{M}\dots\mathcal{M}\mathcal{N}_1\mathcal{M}(\rho)| \\
&\leq (4\varepsilon)^2\mathrm{tr}|\mathcal{N}_{m-2}\dots\mathcal{M}\mathcal{N}_1\mathcal{M}(\rho)| \\
&\leq (4\varepsilon)^m . \qquad (3)
\end{aligned}$$

□

Now we can finally upper bound the mutual information between X and $(I_1, \dots, I_N)$:

Lemma 3.6. *The mutual information can be upper bounded as follows:*

$$\mathcal{I}(X : I_1, \dots, I_N) = \mathcal{O}(N\varepsilon^2).$$

Proof. For all $1 \leq t \leq N$, we remark that $u_{i_t}^{t,x} = \langle\phi_{i_t}^t| d\mathcal{P}_x^{m_t}(\rho_t) - \mathbb{I}|\phi_{i_t}^t\rangle = (\langle\phi| d\mathcal{P}_x\mathcal{N}_{m_t-1}\mathcal{P}_x\dots\mathcal{P}_x\mathcal{N}_1\mathcal{P}(\rho_t)|\phi\rangle - 1)$, so by Lemma 3.4 and Lemma 3.5:

$$\mathcal{I}(X : I_t|I_{\leq t-1}) \leq 3\mathbb{E}_{x,i\sim q_{\leq t-1}}\sum_{i_t}\frac{\lambda_{i_t}^t}{d}(u_{i_t}^{t,x})^2 \leq 3\mathbb{E}_{x,i\sim q_{\leq k-1}}\sum_{i_t}\frac{\lambda_{i_t}^t}{d}16\varepsilon^2 = 48\varepsilon^2$$

because $\sum_{i_t}\lambda_{i_t}^t = d$. Finally:

$$\mathcal{I}(X : I_1, \dots, I_N) = \sum_{t=1}^N \mathcal{I}(X : I_t|I_{\leq t-1}) = \mathcal{O}(N\varepsilon^2).$$

Using Lemma 3.3 and Lemma 3.6 we obtain:

$$\Omega(d^2) \leq \mathcal{I} \leq \mathcal{O}(N\varepsilon^2),$$

which yields the lower bound $N = \Omega(d^2/\varepsilon^2)$. □

□

To assess a lower bound, we need to compare it with upper bounds. The algorithm of Flammia and Wallman (2020) implies an upper bound of $\mathcal{O}\left(\frac{d^3\log(d)}{\varepsilon^2}\right)$ (see App. B for a self contained proof), so there is a gap between our lower bound and this upper bound. However, note that the algorithm of Flammia and Wallman (2020) (and in fact most channel learning protocols we are aware of) use non-adaptive strategies. We will now show that indeed Flammia and Wallman (2020) is optimal if we restrict to non-adaptive protocols.

4 Optimal Pauli channel tomography with non-adaptive strategies

The main difference between non-adaptive and adaptive strategies is that the former should choose the set of inputs, number of repetition, unital channels applied in between and the measurement devices before starting the learning procedure so that they cannot depend on the actual observations of the algorithm. Concretely, besides fixing the total number of steps N and the total number of channels uses at each step $\{m_t\}_{t\in[N]}$, the non-adaptive algorithm is asked to choose also the inputs $\{\rho_t\}_{t\in[N]}$, the unital channels $\{\{\mathcal{N}_j\}_{j\in[m_t-1]}\}_{t\in[N]}$ and the POVMs $\{\mathcal{M}_t\}_{t\in[N]}$ which we suppose without loss of generality have the form $\mathcal{M}_t = \{\lambda_i^t|\phi_i^t\rangle\langle\phi_i^t|\}_{i\in\mathcal{I}_t}$ where $\langle\phi_i^t|\phi_i^t\rangle = 1$ and $\sum_{i\in\mathcal{I}_t}\lambda_i^t = d$. The output state at step $t \in [N]$ has the form:

$$\rho_t^{\text{output}} = \underbrace{\mathcal{P}\circ\mathcal{N}_{m_t-1}\circ\mathcal{P}\circ\dots\circ\mathcal{P}\circ\mathcal{N}_1\circ\mathcal{P}(\rho_t)}_{m_t \text{ times}}.$$

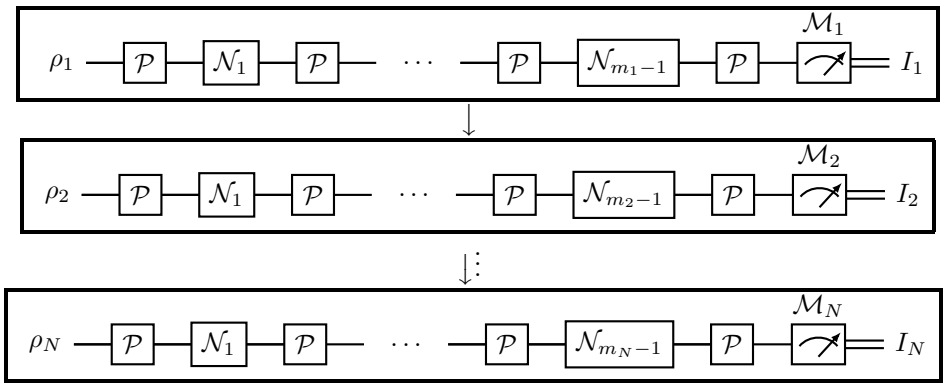

Figure 2: Illustration of a non-adaptive strategy for learning Pauli channel.

Hence, when the algorithm performs a measurement on the output state ρ_t^{output} using the POVM $\mathcal{M}_t = \{\lambda_i^t |\phi_i^t\rangle\langle\phi_i^t|\}_{i\in\mathcal{I}_t}$, it observes $i_t \in \mathcal{I}_t$ with probability:

$$\operatorname{tr}\left(\lambda_{i_t}^t \left|\phi_{i_t}^t\right\rangle\!\left\langle\phi_{i_t}^t\right| \rho_t^{\text{output}}\right) = \lambda_{i_t}^t \left\langle\phi_{i_t}^t\right| \mathcal{P}\circ\mathcal{N}_{m_t-1}\circ\mathcal{P}\circ\cdots\circ\mathcal{P}\circ\mathcal{N}_1\circ\mathcal{P}(\rho_t)\left|\phi_{i_t}^t\right\rangle.$$

We refer to Fig.2 for an illustration. We prove the following lower bound on the total number of measurements and steps:

Theorem 4.1. *The problem of Pauli channel tomography using non-adaptive incoherent measurements requires a total number of channel uses verifying:*

$$\sum_{t=1}^{N} m_t = \Omega\left(\frac{d^4}{\varepsilon^6}\right)$$

or a total number of steps verifying:

$$N = \Omega\left(\frac{d^3}{\varepsilon^2}\right).$$

At a first sight we can think that this Theorem is not comparable to Theorem 3.1 since we give lower bounds on different parameters. However, if we ask the algorithm to only apply the channel once per step, we obtain an improved lower bound on the number of steps required for Pauli channel tomography using non-adaptive strategies. Moreover, it shows that the upper bound of Flammia and Wallman (2020) is almost optimal especially if we know that the additional uses of channels at each step are only required to make the algorithm resilient to errors in SPAM. Finally, the optimal complexity $\Theta\left(\frac{d^3}{\varepsilon^2}\right)$ for Pauli channel tomography is quite surprising: We are ultimately interested in learning a classical distribution on $\mathbb{P}_n \simeq [d^2]$ in TV-distance which requires a complexity of $\Theta\left(\frac{d^2}{\varepsilon^2}\right)$ in the usual sampling access model, so our model is strictly weaker than the usual sampling access model. Furthermore, the quantum state tomography problem has also an optimal copy complexity of $\Theta\left(\frac{d^3}{\varepsilon^2}\right)$: this shows that adding an additional structure to the channel can make the optimal complexity of channel tomography smaller.

Proof. The construction on the family $\mathcal{F}$ is similar to the construction in the proof of Theorem 3.1. We only need to add some constraints about the concentration of the means $\frac{1}{M}\sum_{x=1}^{M} g(\alpha_x)$ around their expectations for every function $g \in \mathcal{G}$. To see what are the functions we need to consider in the set $\mathcal{G}$, let us simplify the mutual information between X and $I_1, \ldots, I_N$ in the non-adaptive setting. Recall from Lemma 3.4 that the mutual information can be upper bounded as follows:

$$\mathcal{I}(X : I_1, \ldots, I_N) = \sum_{t=1}^{N} \mathcal{I}(X : I_t | I_{\leq t-1}) \leq 3\sum_{t=1}^{N} \mathbb{E}_{x,i\sim q_{\leq t-1}} \sum_{i_t} \frac{\lambda_{i_t}^t}{d} (u_{i_t}^{t,x})^2.$$

Since now we consider non-adaptive algorithms, this upper bound can be simplified:

$$3\mathbb{E}_{x,i\sim q_{\leq t-1}} \sum_{i_t} \frac{\lambda_{i_t}^t}{d} (u_{i_t}^{t,x})^2 = 3\frac{1}{M}\sum_{x=1}^{M}\sum_{i_t\in\mathcal{I}_t} \frac{\lambda_{i_t}^t}{d} (u_{i_t}^{t,x})^2.$$

We remark that we only need to approximate

$$\frac{1}{M}\sum_{x=1}^{M}\sum_{t=1}^{N}\sum_{i\in\mathcal{I}_t}\frac{\lambda_i^t}{d}\left(\left\langle\phi_i^t\right|d\mathcal{P}_x^{m_t}(\rho_t)\left|\phi_i^t\right\rangle-1\right)^2.$$

Note that $\left(\langle\phi|\, d\mathcal{P}^{m_t}(\rho_t)\,|\phi\rangle-1\right)^2\in[0,(4\varepsilon)^2]$ for every $|\phi\rangle\in\mathbf{S}^d$ and $\varepsilon\leq 1/4$ (see (3)). Also, we have for all $t\in[N]$, $\sum_{i\in\mathcal{I}_t}\frac{\lambda_i^t}{d}=1$ so

$$\sum_{t=1}^{N}\sum_{i\in\mathcal{I}_t}\frac{\lambda_i^t}{d}\left(\left\langle\phi_i^t\right|d\mathcal{P}_x^{m_t}(\rho_t)\left|\phi_i^t\right\rangle-1\right)^2\in[0,16N\varepsilon^2].$$

Therefore by Hoeffding's inequality Hoeffding (1963) for $s=\sqrt{\frac{(16N\varepsilon^2)^2\log(10)}{2M}}$

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{M}\sum_{x=1}^{M}\sum_{t=1}^{N}\sum_{i\in\mathcal{I}_t}\frac{\lambda_i^t}{d}\left(\left\langle\phi_i^t\right|d\mathcal{P}_x^{m_t}(\rho_t)\left|\phi_i^t\right\rangle-1\right)^2-\mathbb{E}_\alpha\sum_{t=1}^{N}\sum_{i\in\mathcal{I}_t}\frac{\lambda_i^t}{d}\left(\left\langle\phi_i^t\right|d\mathcal{P}_\alpha^{m_t}(\rho_t)\left|\phi_i^t\right\rangle-1\right)^2\right|>s\right)\\&\leq\exp\left(-\frac{2Ms^2}{(16N\varepsilon^2)^2}\right)=\exp\left(-\frac{2M\frac{(16N\varepsilon^2)^2\log(10)}{2M}}{(16N\varepsilon^2)^2}\right)=\frac{1}{10}.\end{aligned}$$

By a union bound, this error probability $1/10$ can be absorbed in the error probability of the construction by choosing a small enough constant c in the cardinality of the family $M=\exp(cd^2)$. To recapitulate, we have proven so far that we can construct the family of quantum Pauli channels $\mathcal{F}$ so that the mutual information satisfies:

$$\Omega(d^2)\leq\mathcal{I}\leq\mathcal{I}(X:I_1,\ldots,I_N)\leq 3\sum_t\sum_{i_t\in\mathcal{I}_t}\frac{\lambda_{i_t}^t}{d}\mathbb{E}_\alpha\left(\left\langle\phi_{i_t}^t\right|d\mathcal{P}_\alpha^{m_t}(\rho_t)\left|\phi_{i_t}^t\right\rangle-1\right)^2+52N\varepsilon^2\exp(-\Omega(d^2)).\tag{4}$$

We claim that the RHS can be upper bounded for $m_t=1$ as follows:

Lemma 4.2. *For all $t\in[N]$, for all unit vectors $|\phi\rangle\in\mathbf{S}^d$:*

$$\mathbb{E}_\alpha\left(\langle\phi|\, d\mathcal{P}_\alpha(\rho_t)\,|\phi\rangle-1\right)^2\leq\frac{32\varepsilon^2}{d}.$$

If the claim is true, the inequalities 4 imply using the fact that for all $t\leq N$, $\sum_{i_t\in\mathcal{I}_t}\lambda_{i_t}^t=d$:

$$\Omega(d^2)\leq\mathcal{I}\leq 3\sum_{t=1}^{N}\sum_{i_t\in\mathcal{I}_t}\frac{\lambda_{i_t}^t}{d}\frac{\varepsilon^2}{d}+52N\varepsilon^2\exp(-\Omega(d^2))\leq\mathcal{O}\left(N\frac{\varepsilon^2}{d}\right)$$

and the lower bound of $N=\Omega(d^3/\varepsilon^2)$ yields for strategies using only one channel per step.

Proof. Let $t\in[N]$ and $|\phi\rangle\in\mathbf{S}^d$. We have:

$$\begin{aligned}&\mathbb{E}_\alpha(\langle\phi|\, d\mathcal{P}_\alpha(\rho_t)\,|\phi\rangle-1)^2=\mathbb{E}_\alpha\left(\sum_{P\in\mathbb{P}_n}\frac{4\alpha(P)\varepsilon}{d}\langle\phi|\,P\rho_tP^\dagger\,|\phi\rangle\right)^2\\&=\mathbb{E}_\alpha\sum_{P,Q\in\mathbb{P}_n}\frac{16\alpha(P)\alpha(Q)\varepsilon^2}{d^2}\langle\phi|\,P\rho_tP^\dagger\,|\phi\rangle\langle\phi|\,Q\rho_tQ^\dagger\,|\phi\rangle\\&=\sum_{P\in\mathbb{P}_n}\frac{16\varepsilon^2}{d^2}(\langle\phi|\,P\rho_tP^\dagger\,|\phi\rangle\langle\phi|\,P\rho_tP^\dagger\,|\phi\rangle-\langle\phi|\,P\rho_tP^\dagger\,|\phi\rangle\langle\phi|\,\sigma(P)\rho_t\sigma(P)^\dagger\,|\phi\rangle)\\&\leq\sum_{P\in\mathbb{P}_n}\frac{32\varepsilon^2}{d^2}\langle\phi|\,P\rho_tP^\dagger\,|\phi\rangle^2\leq\sum_{P\in\mathbb{P}_n}\frac{32\varepsilon^2}{d^2}\langle\phi|\,P\rho_t^2P^\dagger\,|\phi\rangle=\frac{32\varepsilon^2}{d^2}\langle\phi|\,d\mathrm{tr}(\rho_t^2)\mathbb{I}\,|\phi\rangle\leq\frac{32\varepsilon^2}{d},\end{aligned}$$

where we used the Cauchy-Schwartz inequality. □

Now, if we allow multiple uses of the channel at each step, we obtain the following upper bound depending on the number $m \geq 2$ of channel uses:

Lemma 4.3. *For all $t \in [N]$, $m \geq 2$ and unit vectors $|\phi\rangle \in \mathbf{S}^d$:*

$$\mathbb{E}_\alpha \left(\langle \phi | \, d\mathcal{P}_\alpha^m(\rho_t) \, |\phi\rangle - 1 \right)^2 \leq 4m \frac{(4\varepsilon)^{2m}}{d^{\min\{2,m-1\}}}.$$

Proof. Recall that for a Pauli channel $\mathcal{P}_\alpha$, we can define $\mathcal{M}_\alpha = \mathcal{P}_\alpha - \mathrm{tr}(\cdot)\frac{\mathbb{I}}{d}$ so that after m applications of the Pauli channel $\mathcal{P}_\alpha$ intertwined by the unital quantum channels $\mathcal{N}_1, \ldots, \mathcal{N}_{m-1}$, we have the following identity:

$$\underbrace{\mathcal{P}_\alpha \mathcal{N}_{m-1} \mathcal{P}_\alpha \ldots \mathcal{P}_\alpha \mathcal{N}_1 \mathcal{P}_\alpha(\rho)}_{m \text{ times}} = \mathrm{tr}(\rho)\frac{\mathbb{I}}{d} + \underbrace{\mathcal{M}_\alpha \mathcal{N}_{m-1} \mathcal{M}_\alpha \ldots \mathcal{M}_\alpha \mathcal{N}_1 \mathcal{M}_\alpha(\rho)}_{m \text{ times}}.$$

The definition of $\mathcal{P}_\alpha$ implies:

$$\mathcal{M}_\alpha(\rho) = \mathcal{P}_\alpha(\rho) - \mathrm{tr}(\rho)\frac{\mathbb{I}}{d} = \sum_{P \in \mathbb{P}_n} \frac{4\alpha(P)\varepsilon}{d^2} P\rho P = \sum_{P \in \mathbb{P}_n} \frac{4\alpha(P)\varepsilon}{d^2} \mathcal{N}_P(\rho)$$

where we use the notation for the unital quantum channel $\mathcal{N}_P(\rho) = P\rho P$ for all $P \in \mathbb{P}_n$. So, using the notation $\mathcal{N}_{P_m, m-1, \ldots, 1, P_1} = \mathcal{N}_{P_m} \mathcal{N}_{m-1} \mathcal{N}_{P_{m-1}} \ldots \mathcal{N}_{P_2} \mathcal{N}_1 \mathcal{N}_{P_1}$, we can develop the quantity we want to upper bound as follows:

$$\begin{aligned}
&\mathbb{E}_\alpha \left(\langle \phi | \, d\mathcal{P}_\alpha^m(\rho) \, |\phi\rangle - 1 \right)^2 = d^2 \mathbb{E}_\alpha \left(\langle \phi | \, \mathcal{M}_\alpha \mathcal{N}_{m-1} \mathcal{M}_\alpha \ldots \mathcal{M}_\alpha \mathcal{N}_1 \mathcal{M}_\alpha(\rho) \, |\phi\rangle \right)^2 \\
&= d^2 \mathbb{E}_\alpha \left(\sum_{P_1, \ldots, P_m} \frac{4\alpha(P_1)\varepsilon}{d^2} \cdots \frac{4\alpha(P_m)\varepsilon}{d^2} \langle \phi | \, \mathcal{N}_{P_m} \mathcal{N}_{m-1} \mathcal{N}_{P_{m-1}} \ldots \mathcal{N}_{P_2} \mathcal{N}_1 \mathcal{N}_{P_1}(\rho) \, |\phi\rangle \right)^2 \\
&= \frac{(4\varepsilon)^{2m}}{d^{4m-2}} \sum_{P, Q \in \mathbb{P}_n} \mathbb{E}_\alpha \left(\alpha(P_1) \cdots \alpha(P_m) \alpha(Q_1) \cdots \alpha(Q_m) \right) \langle \phi | \, \mathcal{N}_{P_m, m-1, \ldots, 1, P_1}(\rho) \, |\phi\rangle \, \langle \phi | \, \mathcal{N}_{Q_m, m-1, \ldots, 1, Q_1}(\rho) \, |\phi\rangle .
\end{aligned}$$

If $Q_1, \sigma(Q_1) \notin \{P_1, \ldots, P_m, Q_2, \ldots, Q_m\}$ the expected value $\mathbb{E}_\alpha \left(\alpha(P_1) \cdots \alpha(P_m) \alpha(Q_1) \cdots \alpha(Q_m) \right)$ is 0, otherwise, we can upper bound each term inside the sum by 1 and we count the number of these terms. Moreover we can gain two factors d by using the properties of Pauli group for $m \geq 3$. For example, suppose that $Q_1 = P_1$, we have $\sum_{P \in \mathbb{P}_n} \mathcal{N}_P(\rho) = \sum_{P \in \mathbb{P}_n} P\rho P = d\mathrm{tr}(\rho)\mathbb{I}$ hence for $m \geq 3$:

$$\begin{aligned}
&\frac{(4\varepsilon)^{2m}}{d^{4m-2}} \sum_{P, Q : Q_1 = P_1} \mathbb{E}_\alpha \left(\alpha(P_1) \cdots \alpha(P_m) \alpha(Q_1) \cdots \alpha(Q_m) \right) \langle \phi | \, \mathcal{N}_{P_m, m-1, \ldots, 1, P_1}(\rho) \, |\phi\rangle \, \langle \phi | \, \mathcal{N}_{Q_m, m-1, \ldots, 1, Q_1}(\rho) \, |\phi\rangle \\
&\leq \frac{(4\varepsilon)^{2m}}{d^{4m-2}} \sum_{P, Q : Q_1 = P_1} \langle \phi | \, \mathcal{N}_{P_m, m-1, \ldots, 1, P_1}(\rho) \, |\phi\rangle \, \langle \phi | \, \mathcal{N}_{Q_m, m-1, \ldots, 1, Q_1}(\rho) \, |\phi\rangle \\
&= \frac{(4\varepsilon)^{2m}}{d^{4m-2}} \sum_{P, Q_{<m} : Q_1 = P_1} \langle \phi | \sum_{P_m} \mathcal{N}_{P_m, m-1, \ldots, 1, P_1}(\rho) \, |\phi\rangle \, \langle \phi | \sum_{Q_m} \mathcal{N}_{Q_m, m-1, \ldots, 1, Q_1}(\rho) \, |\phi\rangle \\
&= \frac{(4\varepsilon)^{2m}}{d^{4m-2}} \sum_{P, Q_{<m} : Q_1 = P_1} \langle \phi | \, d\mathrm{tr}(\mathcal{N}_{m-1, \ldots, 1, P_1}(\rho))\mathbb{I} \, |\phi\rangle \, \langle \phi | \, d\mathrm{tr}(\mathcal{N}_{m-1, \ldots, 1, Q_1}(\rho))\mathbb{I} \, |\phi\rangle \\
&= \frac{(4\varepsilon)^{2m}}{d^{4m-2}} \sum_{P, Q_{<m} : Q_1 = P_1} d^2 = \frac{(4\varepsilon)^{2m}}{d^{4m-2}} (d^2)^{2m-3} d^2 = \frac{(4\varepsilon)^{2m}}{d^2}.
\end{aligned}$$

Since we have $2(2m-1)$ possibilities for $Q_1, \sigma(Q_1) \in \{P_1, \ldots, P_m, Q_2, \ldots, Q_m\}$, we conclude that:

$$\mathbb{E}_\alpha \left(\langle \phi | \, d\mathcal{P}_\alpha^m(\rho) \, |\phi\rangle - 1 \right)^2 \leq 2(2m-1)\frac{(4\varepsilon)^{2m}}{d^2} \leq 4m\frac{(4\varepsilon)^{2m}}{d^2}.$$

Now, if $m = 2$, we can have $Q_1 = Q_2$ and therefore we can't gain a factor d when summing over Q_2. In this case, we obtain instead an upper bound:

$$\mathbb{E}_\alpha \left(\langle \phi | \, d\mathcal{P}_\alpha^2(\rho) \, |\phi\rangle - 1 \right)^2 \leq 6\frac{(4\varepsilon)^4}{d}.$$

□

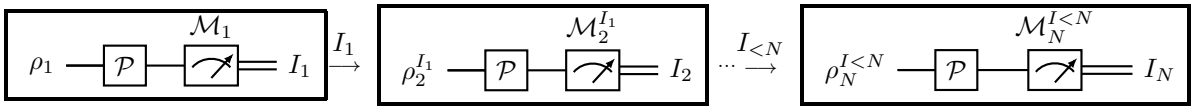

Figure 3: Illustration of an adaptive strategy for learning Pauli channel using one channel per step.

Using the inequalities 4 and the fact that for all $t \in [N]$, $\sum_{i_t \in \mathcal{I}_t} \lambda_{i_t}^t = d$, we deduce:

$$2^{10} \sum_{t:m_t \leq 2} \frac{\varepsilon^2}{d} + 4 \sum_{t:m_t \geq 3} m_t \frac{(4\varepsilon)^{2m_t}}{d^2} = \Omega(d^2).$$

Therefore we have either $\sum_{t:m_t \leq 2} \frac{\varepsilon^2}{d} = \Omega(d^2)$ or $4 \sum_{t:m_t \geq 3} m_t \frac{(4\varepsilon)^{2m_t}}{d^2} = \Omega(d^2)$. Finally, we have either $N = \Omega\left(\frac{d^3}{\varepsilon^2}\right)$ or $\sum_{t=1}^N m_t = \Omega\left(\frac{d^4}{\varepsilon^6}\right)$.

□

This proof relies crucially on the non-adaptiveness of the strategy. This can be seen clearly when simplifying the upper bound of the conditional mutual information in Lemma 3.4. For an adaptive strategy, this upper bound contains large products for which the expectation (under α) can only upper bounded by $\mathcal{O}(\varepsilon^2)$ which implies a lower bound on N similar to Theorem 3.1. In the next section, we explore how to overcome this difficulty in some regime of ε and improve the general lower bound $N = \Omega(d^2/\varepsilon^2)$.

5 A lower bound for Pauli channel tomography with adaptive strategies in the high precision regime

In this section, we improve the general lower bound of quantum Pauli channel tomography in Theorem 3.1 for adaptive strategies with one use of the channel each step. In the adaptive setting, a learner could adapt its choices depending on the previous observations. It can prepare a large set of inputs and measurements and thus it potentially has more power to extract information much earlier than its non-adaptive counterpart. With this intuition, we expect that lower bounds for adaptive strategies should be harder to establish. Moreover, in this section, we only consider one use of the channel for each step, i.e., $m_t = 1$. After observing $I_1, \ldots, I_t$ at steps 1 to t, the learner would choose an input $\rho_{t+1}^{\leq t} := \rho_{t+1}^{I_1,\ldots,I_t}$ and a measurement device represented by a POVM $\mathcal{M}_{t+1}^{\leq t} := \left\{ \lambda_{i_{t+1}}^{I_1,\ldots,I_t} \left| \phi_{i_{t+1}}^{I_1,\ldots,I_t} \right\rangle \left\langle \phi_{i_{t+1}}^{I_1,\ldots,I_t} \right| \right\}_{i_{t+1} \in \mathcal{I}_{t+1}^{I_1,\ldots,I_t}}$ where the rank one matrices are projectors and the coefficients sum to 1. So, the adaptive algorithm extracts classical information at step $t+1$ from the unknown Pauli quantum channel $\mathcal{P}$ by first applying $\mathcal{P}$ to the input $\rho_{t+1}^{I_1,\ldots,I_t}$ and then performing a measurement using the POVM $\mathcal{M}_{t+1}^{I_1,\ldots,I_t}$ (see Fig.3 for an illustration). In this case, it observes $i_{t+1} \in \mathcal{I}_{t+1}^{I_1,\ldots,I_t}$ with a probability given by Born's rule:

$$\operatorname{tr}\left(\rho_{t+1}^{I_1,\ldots,I_t} \lambda_{i_{t+1}}^{I_1,\ldots,I_t} \left| \phi_{i_{t+1}}^{I_1,\ldots,I_t} \right\rangle \left\langle \phi_{i_{t+1}}^{I_1,\ldots,I_t} \right| \right) = \lambda_{i_{t+1}}^{I_1,\ldots,I_t} \left\langle \phi_{i_{t+1}}^{I_1,\ldots,I_t} \right| \rho_{t+1}^{I_1,\ldots,I_t} \left| \phi_{i_{t+1}}^{I_1,\ldots,I_t} \right\rangle.$$

We prove the following lower bound on the number of steps. Note that because of the assumption $m_t = 1$ for all steps t, the number of steps is the same as the number of channel uses.

Theorem 5.1. *Let $\varepsilon \leq 1/(20d)$. Adaptive strategies for the problem of Pauli channel tomography using incoherent measurements require a number of steps satisfying:*

$$N = \Omega\left(\frac{d^{5/2}}{\varepsilon^2}\right).$$

Furthermore, any adaptive strategy that uses $\mathcal{O}(d^2/\varepsilon^2)$ memory requires a number of steps N satisfying

$$N = \Omega\left(\frac{d^3}{\varepsilon^2}\right). \tag{5}$$

In this Theorem, we show that we can improve on the general lower bound of Theorem 3.1 by an exponential factor of number of qubits if the precision parameter ε is small enough. However, this

lower bound could be as well not optimal so it remains either to improve it to match the non-adaptive upper bound of Flammia and Wallman (2020) or to propose an adaptive algorithm with a number of steps matching this lower bound. With the same proof, we can generalize this lower bound to adaptive algorithms with limited memory. Any strategy that adapts on at most $\left\lceil \frac{H}{\varepsilon^2} \right\rceil$ previous observations for the problem of Pauli channel tomography using incoherent measurements requires a number of steps $N = \Omega\left(\min\left\{\frac{d^4}{\sqrt{H}\varepsilon^2}, \frac{d^5}{H\varepsilon^2}, \frac{d^3}{\varepsilon^2}\right\}\right)$. For instance, if the algorithm can only adapt its input state and measurement device on the previous $\left\lceil \frac{d^2}{\varepsilon^2} \right\rceil$ observations then it requires $N = \Omega\left(\frac{d^3}{\varepsilon^2}\right)$ steps to correctly approximate the unknown Pauli channel. The remaining of this Section is reserved to the proof of this Theorem.

Construction of the family $\mathcal{F}$ We start by constructing a family of Pauli quantum channels that is $\Omega(\varepsilon)$-separated. The elements of this family would have the following form, for all $x \in \mathcal{F} = [\![1, M]\!]$:

$$\mathcal{P}_x(\rho) = \sum_{P \in \mathbb{P}_n} \frac{1 + 2\tilde{\alpha}_x(P)\varepsilon d/\|\alpha_x\|_2}{d^2} P\rho P = \sum_{P \in \mathbb{P}_n} p_x(P) P\rho P$$

where $\tilde{\alpha}(P) = \alpha(P) - \frac{1}{d^2}\sum_{Q \in \mathbb{P}_n} \alpha(Q)$ and $\{\alpha(P)\}_P$ are d^2 random variables i.i.d. as $\mathcal{N}(0,1)$ and $p_x(P) = \frac{1+2\tilde{\alpha}_x(P)\varepsilon d/\|\alpha_x\|_2}{d^2}$. It is not difficult to check that $\{p_x\}_x$ are valid probabilities for $\varepsilon \leq 1/4d$. Indeed, for all $P \in \mathbb{P}_n$ we have $|\tilde{\alpha}(P)| \leq 2\|\alpha\|_2$ so for $\varepsilon \leq 1/4d$ we have $1 + 2\tilde{\alpha}_x(P)\varepsilon d/\|\alpha_x\|_2 \in [0, 2]$ thus $p_x(P) \in [0, 2/d^2] \subset [0, 1)$ for $d \geq 2$. Moreover, we claim that:

Lemma 5.2. *Let β be an independent and identically distributed copy of α. We have:*

$$\mathbb{P}\left(\mathrm{TV}(p_\alpha, p_\beta) < \varepsilon/5\right) \leq e^{-\Omega(d^2)}.$$

If this claim is true, then a union bound permits to show the existence of the family with the property $M = \exp(\Omega(d^2))$. Let us prove first a lower bound on the expected TV distance between p_α and p_β.

Lemma 5.3. *Let β be an independent and identically distributed copy of α. We have:*

$$\mathbb{E}\left(\mathrm{TV}(p_\alpha, p_\beta)\right) \geq \frac{7\varepsilon}{20}.$$

Proof. We start by writing:

$$\mathrm{TV}(p_\alpha, p_\beta) = \frac{\varepsilon}{d}\sum_{P \in \mathbb{P}_n}\left|\frac{\tilde{\alpha}(P)}{\|\alpha\|_2} - \frac{\tilde{\beta}(P)}{\|\beta\|_2}\right| \geq \frac{\varepsilon}{d}\sum_{P \in \mathbb{P}_n}\left|\frac{\alpha(P)}{\|\alpha\|_2} - \frac{\beta(P)}{\|\beta\|_2}\right| - \frac{\varepsilon}{d}\left|\frac{\sum_{Q \in \mathbb{P}_n}\alpha(Q)}{\|\alpha\|_2} - \frac{\sum_{Q \in \mathbb{P}_n}\beta(Q)}{\|\beta\|_2}\right|$$

where we use the triangle inequality. We can upper bound the expectation of the second difference using the fact that $\|\alpha\|_2$ is independent of $\{\alpha(P)/\|\alpha\|_2\}_P$:

$$\mathbb{E}\left(\frac{\varepsilon}{d}\left|\frac{\sum_{Q \in \mathbb{P}_n}\alpha(Q)}{\|\alpha\|_2} - \frac{\sum_{Q \in \mathbb{P}_n}\beta(Q)}{\|\beta\|_2}\right|\right) \leq 2\mathbb{E}\left(\frac{\varepsilon}{d}\left|\frac{\sum_{P \in \mathbb{P}_n}\alpha(P)}{\|\alpha\|_2}\right|\right) = \frac{2\varepsilon}{d}\frac{\mathbb{E}\left(\left|\sum_{P \in \mathbb{P}_n}\alpha(P)\right|\right)}{\mathbb{E}\left(\|\alpha\|_2\right)} \leq \frac{4\varepsilon}{d}.$$

Indeed, by the Cauchy-Schwartz inequality:

$$\mathbb{E}\left(\left|\sum_{P \in \mathbb{P}_n}\alpha(P)\right|\right) \leq \sqrt{\mathbb{E}\left(\left(\sum_{P \in \mathbb{P}_n}\alpha(P)\right)^2\right)} = d$$

and by the Hölder's inequality:

$$\mathbb{E}\left(\|\alpha\|_2\right) \geq \sqrt{\frac{\mathbb{E}\left(\|\alpha\|_2^2\right)^3}{\mathbb{E}\left(\|\alpha\|_2^4\right)}} = \sqrt{\frac{(d^2)^3}{d^2(d^2-1) + 3d^2}} \geq \frac{d}{2}.$$

We move to lower bound the expectation of the first difference using Hölder's inequality:

$$\mathbb{E}\left(\frac{\varepsilon}{d}\sum_{P \in \mathbb{P}_n}\left|\frac{\alpha(P)}{\|\alpha\|_2} - \frac{\beta(P)}{\|\beta\|_2}\right|\right) = \varepsilon d\,\mathbb{E}\left(\left|\frac{\alpha(P)}{\|\alpha\|_2} - \frac{\beta(P)}{\|\beta\|_2}\right|\right) \geq \varepsilon d\sqrt{\frac{\mathbb{E}\left(\left|\frac{\alpha(P)}{\|\alpha\|_2} - \frac{\beta(P)}{\|\beta\|_2}\right|^2\right)^3}{\mathbb{E}\left(\left|\frac{\alpha(P)}{\|\alpha\|_2} - \frac{\beta(P)}{\|\beta\|_2}\right|^4\right)}}$$

$$\geq \varepsilon d\sqrt{\frac{8/d^6}{48/d^4}} \geq \frac{2\varepsilon}{5}.$$

because we can compute the numerator using the fact that $\|\alpha\|_2$ is independent of $\{\alpha(P)/\|\alpha\|_2\}_P$:

$$\mathbb{E}\left(\left|\frac{\alpha(P)}{\|\alpha\|_2}-\frac{\beta(P)}{\|\beta\|_2}\right|^2\right)=2\mathbb{E}\left(\frac{\alpha(P)^2}{\|\alpha\|_2^2}\right)-2\mathbb{E}\left(\frac{\alpha(P)\beta(P)}{\|\alpha\|_2\|\beta\|_2}\right)=2\frac{\mathbb{E}\left(\alpha(P)^2\right)}{\mathbb{E}\left(\|\alpha\|_2^2\right)}=\frac{2}{d^2}.$$

Moreover, we can bound the denominator by Hölder's inequality and the fact that $\|\alpha\|_2$ is independent of $\{\alpha(P)/\|\alpha\|_2\}_P$:

$$\mathbb{E}\left(\left|\frac{\alpha(P)}{\|\alpha\|_2}-\frac{\beta(P)}{\|\beta\|_2}\right|^4\right)\leq 16\mathbb{E}\left(\frac{\alpha(P)^4}{\|\alpha\|_2^4}\right)=16\frac{\mathbb{E}\left(\alpha(P)^4\right)}{\mathbb{E}\left(\|\alpha\|_2^4\right)}=\frac{48}{d^2(d^2-1)+3d^2}\leq\frac{48}{d^4}.$$

Therefore the expected value of the TV-distance satisfies:

$$\mathbb{E}\left(\mathrm{TV}(p_\alpha,p_\beta)\right)\geq\frac{2\varepsilon}{5}-\frac{4\varepsilon}{d}\geq\frac{7\varepsilon}{20}.$$

□

Once we have a lower bound on the expected value of $\mathrm{TV}(p_\alpha,p_\beta)$, we can proceed to prove Lemma 5.2.

Proof. We want to show that the function $\mathrm{TV}(p_\alpha,p_\beta)$ concentrates around its mean. Let (α,γ) and (β,δ) be two couples of standard Gaussian vectors. By the reverse triangle inequality we have:

$$\begin{aligned}|\,\mathrm{TV}(p_\alpha,p_\gamma)-\mathrm{TV}(p_\beta,p_\delta)|&\leq|\,\mathrm{TV}(p_\alpha,p_\gamma)-\mathrm{TV}(p_\gamma,p_\beta)|+|\,\mathrm{TV}(p_\gamma,p_\beta)-\mathrm{TV}(p_\beta,p_\delta)|\\&\leq\mathrm{TV}(p_\alpha,p_\beta)+\mathrm{TV}(p_\gamma,p_\delta).\end{aligned}$$

On the set $E^2=\{(\alpha,\beta):\|\alpha\|_2,\|\beta\|_2>d/4\}$ we have

$$\begin{aligned}\mathrm{TV}(p_\alpha,p_\beta)&\leq\frac{\varepsilon}{d}\sum_{P\in\mathbb{P}_n}\left|\frac{\alpha(P)}{\|\alpha\|_2}-\frac{\beta(P)}{\|\beta\|_2}\right|+\frac{\varepsilon}{d}\left|\sum_{P\in\mathbb{P}_n}\frac{\alpha(P)}{\|\alpha\|_2}-\frac{\beta(P)}{\|\beta\|_2}\right|\leq\frac{2\varepsilon}{d}\sum_{P\in\mathbb{P}_n}\left|\frac{\alpha(P)}{\|\alpha\|_2}-\frac{\beta(P)}{\|\beta\|_2}\right|\\&\leq\frac{2\varepsilon}{d}\sum_{P\in\mathbb{P}_n}\left|\frac{\alpha(P)}{\|\alpha\|_2}-\frac{\alpha(P)}{\|\beta\|_2}\right|+\frac{2\varepsilon}{d}\sum_{P\in\mathbb{P}_n}\left|\frac{\alpha(P)}{\|\beta\|_2}-\frac{\beta(P)}{\|\beta\|_2}\right|\\&\leq\frac{2\varepsilon}{d}\sum_{P\in\mathbb{P}_n}|\alpha(P)|\left|\frac{\|\alpha\|_2-\|\beta\|_2}{\|\alpha\|_2\|\beta\|_2}\right|+\frac{2\varepsilon}{d}\sum_{P\in\mathbb{P}_n}\left|\frac{\alpha(P)-\beta(P)}{\|\beta\|_2}\right|\\&=\frac{2\varepsilon}{d}\|\alpha\|_1\left|\frac{\|\alpha\|_2-\|\beta\|_2}{\|\alpha\|_2\|\beta\|_2}\right|+\frac{2\varepsilon}{d}\left|\frac{\|\alpha-\beta\|_1}{\|\beta\|_2}\right|\\&\leq 2\varepsilon\|\alpha\|_2\frac{\|\alpha-\beta\|_2}{\|\alpha\|_2\|\beta\|_2}+2\varepsilon\frac{\|\alpha-\beta\|_2}{\|\beta\|_2}\\&\leq 2\varepsilon\frac{4}{d}\|\alpha-\beta\|_2+2\varepsilon\|\alpha-\beta\|_2\frac{4}{d}=\frac{16\varepsilon}{d}\|\alpha-\beta\|_2.\end{aligned}$$

Here we used that $\|\alpha\|_1\leq d\|\alpha\|_2$, as α is a vector with d^2 entries, and our assumption on the norms in the last line. Hence, on the set E^4, by using the inequality $x+y\leq\sqrt{2}\sqrt{x^2+y^2}$:

$$\begin{aligned}|\,\mathrm{TV}(p_\alpha,p_\gamma)-\mathrm{TV}(p_\beta,p_\delta)|&\leq\mathrm{TV}(p_\alpha,p_\beta)+\mathrm{TV}(p_\gamma,p_\delta)\leq\frac{16\varepsilon}{d}\|\alpha-\beta\|_2+\frac{16\varepsilon}{d}\|\gamma-\delta\|_2\\&\leq\frac{16\sqrt{2}\varepsilon}{d}\sqrt{\|\alpha-\beta\|_2^2+\|\gamma-\delta\|_2^2}=:L\|(\alpha,\gamma)-(\beta,\delta)\|_2.\end{aligned}$$

Moreover, the function $\mathrm{TV}(p_\alpha,p_\beta)$ can be extended to an L-Lipschitz function on the whole set $\mathbb{R}^{d^2}\times\mathbb{R}^{d^2}$ using the following definition for every $(\alpha,\beta)\in\mathbb{R}^{d^2}\times\mathbb{R}^{d^2}$ (Kirszbraun theorem, see e.g. Mattila (1999)):

$$f(\alpha,\beta)=\inf_{(\gamma,\delta)\in E^2}\{\mathrm{TV}(p_\gamma,p_\delta)+L\|(\alpha,\beta)-(\gamma,\delta)\|_2\}.$$

We can control the expected value of f using the lower bound on the expected value of $\mathrm{TV}(p_\alpha,p_\beta)$ (Lemma 5.3) as follows:

$$\mathbb{E}\left(f\right)=\mathbb{E}\left(f1_{E^2}\right)+\mathbb{E}\left(f1_{(E^2)^c}\right)\geq\mathbb{E}\left(f1_{E^2}\right)\geq\frac{7\varepsilon}{20}-8\varepsilon\exp(-\Omega(d^2))\geq\frac{3\varepsilon}{10}$$

because

$$|\mathbb{E}\left(f(\alpha,\beta)1_{E^2}\right)-\mathbb{E}\left(\mathrm{TV}(p_\alpha,p_\beta)\right)|=\mathbb{E}\left(\mathrm{TV}(p_\alpha,p_\beta)1_{(E^2)^c}\right)\leq 8\varepsilon\mathbb{P}\left(E^c\right)\leq 8\varepsilon\exp(-\Omega(d^2))$$

where we have used the fact that $\mathrm{TV}(p_\alpha,p_\beta)\leq 4\varepsilon$ and $\mathbb{P}\left(E^c\right)=\mathbb{P}\left(\|\alpha\|_2\leq d/4\right)\leq\exp(-\Omega(d^2))$. Indeed, we can apply the concentration of Lipschitz functions of Gaussian random variables (Wainwright, 2019, Theorem 2.26) for the function $\alpha\to\|\alpha\|_2$ which is 1-Lipschitz by the triangle inequality:

$$|\|\alpha\|_2-\|\beta\|_2|\leq\|\alpha-\beta\|_2$$

and its expectation satisfies $\mathbb{E}\left(\|\alpha\|_2\right)\geq d/2$, thus:

$$\mathbb{P}\left(E^c\right)=\mathbb{P}\left(\|\alpha\|_2\leq d/4\right)=\mathbb{P}\left(\|\alpha\|_2-\mathbb{E}\left(\|\alpha\|_2\right)\leq -d/4\right)\leq\exp(-\Omega(d^2)).$$

We proceed with the same strategy for the function f which is L-Lipschitz where $L=\frac{16\sqrt{2}\varepsilon}{d}$. By the concentration of Lipschitz functions of Gaussian random variables (Wainwright, 2019, Theorem 2.26), we obtain for all $s\geq 0$:

$$\mathbb{P}\left(|f-\mathbb{E}\left(f\right)|>s\right)\leq\exp\left(-\frac{cd^2s^2}{\varepsilon^2}\right)$$

with $c>0$ a constant. Then, we can deduce the upper bound on the probability:

$$\begin{aligned}
\mathbb{P}\left(\mathrm{TV}(p_\alpha,p_\beta)<\varepsilon/5\right)&=\mathbb{P}\left(\mathrm{TV}(p_\alpha,p_\beta)<\varepsilon/5,(\alpha,\beta)\in E^2\right)+\mathbb{P}\left(\mathrm{TV}(p_\alpha,p_\beta)<\varepsilon/5,(\alpha,\beta)\notin E^2\right)\\
&\leq\mathbb{P}\left(f(\alpha,\beta)<\varepsilon/5,(\alpha,\beta)\in E^2\right)+\mathbb{P}\left((\alpha,\beta)\notin E^2\right)\\
&\leq\mathbb{P}\left(f(\alpha,\beta)-\mathbb{E}\left(f\right)<\varepsilon/5-3\varepsilon/10,(\alpha,\beta)\in E^2\right)+2\mathbb{P}\left(\alpha\notin E\right)\\
&\leq\mathbb{P}\left(f(\alpha,\beta)-\mathbb{E}\left(f\right)<-\varepsilon/10\right)+2\mathbb{P}\left(\alpha\notin E\right)\\
&\leq 3\exp(-\Omega(d^2))\leq\exp(-\Omega(d^2)).
\end{aligned}$$

□

Hence we construct an $\varepsilon/5$-separated family $\mathcal{F}$ of cardinal $\Omega(d^2)$. By changing $\varepsilon\leftrightarrow 5\varepsilon$ in the definition of $\{\mathcal{P}_x\}_{x\in\mathcal{F}}$, the family becomes ε-separated for $\varepsilon\leq 1/(20d)$.

Once the family $\mathcal{F}$ is constructed, we can use it to encode a message in $[\![1,M]\!]$ to a quantum Pauli channel $\mathcal{P}=\mathcal{P}_x$ in the family $\mathcal{F}$. The decoder receives this unknown quantum Pauli channel, chooses its inputs states and performs adaptive incoherent measurements and learns it. Therefore a $1/3$-correct algorithm can decode with a probability of failure at most $1/3$ by finding the closest quantum Pauli channel in the family $\mathcal{F}$ to the channel approximated by the algorithm. By Fano's inequality, the encoder and decoder should share at least $\Omega(\log(M))=\Omega(d^2)$ nats of information.

Lemma 5.4 (Fano (1961))**.** *The mutual information between the encoder and the decoder is at least*

$$\mathcal{I}\geq 2/3\log(M)-\log(2)\geq\Omega(d^2).$$

Upper bound on the mutual information Since we have a lower bound on the mutual information, it remains to prove an upper bound depending on the number of steps N and the precision ε. For this, let us denote by X the uniform random variable on the set $[\![1,M]\!]$ representing the encoder and $I_1,\ldots,I_N$ the observations of the decoder or the $1/3$-correct algorithm. The Data-Processing inequality implies:

$$\mathcal{I}\leq\mathcal{I}(X:I_1,\ldots,I_N).$$

By upper bounding the mutual information between X and $I_1,\ldots,I_N$ and using a contradiction argument, we prove Theorem 5.1 which we recall:

Theorem. *Let $\varepsilon\leq 1/(20d)$. Adaptive strategies for the problem of Pauli channel tomography using incoherent measurements requires a number of steps satisfying:*

$$N=\Omega\left(\frac{d^{5/2}}{\varepsilon^2}\right).$$

Proof. Recall that we can write the mutual information as: $\mathcal{I}(X : I_1, \dots, I_N) = \sum_{k=1}^{N} \mathcal{I}(X : I_k | I_{\leq k-1})$. Fix $k \in [N]$, by Lemma 3.4, we can upper bound the conditional mutual information:

$$\mathcal{I}(X : I_k | I_{\leq k-1}) \leq 3\mathbb{E}_x \mathbb{E}_{i \sim q_{\leq k-1}} \left[\sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,x})^2 \right]$$

where we use the notation:

$$\begin{aligned} u_{i_k}^{k,x} &= \langle \phi_{i_k}^k | \left(d\mathcal{P}_x(\rho_k) - \mathbb{I} \right) | \phi_{i_k}^k \rangle = \langle \phi_{i_k}^k | \left(\sum_{P \in \mathbb{P}_n} \frac{2\tilde{\alpha}_x(P)\varepsilon}{\|\alpha_x\|_2} P\rho_k P \right) | \phi_{i_k}^k \rangle \\ &= \sum_{P \in \mathbb{P}_n} \frac{2\alpha_x(P)\varepsilon}{\|\alpha_x\|_2} \langle \phi_{i_k}^k | P\rho_k P | \phi_{i_k}^k \rangle - \sum_{P,Q \in \mathbb{P}_n} \frac{2\alpha_x(Q)\varepsilon}{d^2\|\alpha_x\|_2} \langle \phi_{i_k}^k | P\rho_k P | \phi_{i_k}^k \rangle \\ &= \sum_{P \in \mathbb{P}_n} \frac{2\alpha_x(P)\varepsilon}{\|\alpha_x\|_2} \langle \phi_{i_k}^k | P\rho_k P | \phi_{i_k}^k \rangle - \sum_{P \in \mathbb{P}_n} \frac{2\alpha_x(P)\varepsilon}{d\|\alpha_x\|_2}. \end{aligned}$$

Note that for adaptive strategies the vectors $|\phi_{i_k}^k\rangle = |\phi_{i_k}^k(i_1, \dots, i_{k-1})\rangle$ and the states $\rho_k = \rho_k(i_1, \dots, i_{k-1})$ depend on the previous observations $(i_1, \dots, i_{k-1})$ for all $k \in [N]$. Similarly, we denote:

$$\begin{aligned} u_{i_k}^{k,\alpha} &= \sum_{P \in \mathbb{P}_n} \frac{2\alpha(P)\varepsilon}{\|\alpha\|_2} \langle \phi_{i_k}^k | P\rho_k P | \phi_{i_k}^k \rangle - \sum_{P \in \mathbb{P}_n} \frac{2\alpha(P)\varepsilon}{d\|\alpha\|_2} \\ &= \frac{2}{\|\alpha\|_2} \sum_{P \in \mathbb{P}_n} \alpha(P)\varepsilon \langle \phi_{i_k}^k | P(\rho_k - \mathbb{I}/d)P | \phi_{i_k}^k \rangle. \end{aligned}$$

We have $\sum_{i_k} \lambda_{i_k}^k u_{i_k}^{k,x} = \operatorname{tr}\left(d\mathcal{P}_x(\rho_k) - \mathbb{I} \right) = 0$ as

$$\sum_{i_k} \lambda_{i_k}^k u_{i_k}^{k,\alpha} = \sum_{P \in \mathbb{P}_n} \frac{2\alpha_x(P)\varepsilon}{\|\alpha_x\|_2} \operatorname{tr}(P\rho_k P) - \sum_{P \in \mathbb{P}_n} \frac{2\alpha_x(P)\varepsilon}{\|\alpha_x\|_2} = \sum_{P \in \mathbb{P}_n} \frac{2\alpha_x(P)\varepsilon}{\|\alpha_x\|_2} - \sum_{P \in \mathbb{P}_n} \frac{2\alpha_x(P)\varepsilon}{\|\alpha_x\|_2} = 0.$$

Note that the cardinal of the constructed family $M = |\mathcal{F}|$ is of order $\exp(\Omega(d^2))$, so every mean in x can be approximated by the expected value for α following the distribution explained in the construction, the difference will be, by Hoeffding's inequality, of order $\exp(-\Omega(d^2))$ so negligible (see the proof of Proposition C.1 for a detailed justification of this claim, note that the error probability can be absorbed in the total error probability by adding these inequalities in the construction of the family $\mathcal{F}$). Therefore we obtain the upper bound on the mutual information:

$$\begin{aligned} \sum_{k=1}^{N} \mathcal{I}(X : I_k | I_{\leq k-1}) &\leq 3 \sum_{k=1}^{N} \mathbb{E}_{x, i \sim q_{\leq k-1}} \sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,x})^2 \\ &= 3 \sum_{k=1}^{N} \frac{1}{M} \sum_{x=1}^{M} \sum_{i_1, \dots, i_{k-1}} \left(\prod_{t=1}^{k-1} \lambda_{i_t}^t \left(\frac{1 + u_{i_t}^{t,x}}{d} \right) \right) \sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,x})^2 \\ &\leq 3 \sum_{k=1}^{N} \mathbb{E}_\alpha \left[\sum_{i_1, \dots, i_{k-1}} \left(\prod_{t=1}^{k-1} \lambda_{i_t}^t \left(\frac{1 + u_{i_t}^{t,\alpha}}{d} \right) \right) \sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,\alpha})^2 \right] + 3N\varepsilon^2 \exp(-Cd^2) \\ &= 3 \sum_{k=1}^{N} \mathbb{E}_{\leq k} \mathbb{E}_\alpha \left[\left(\prod_{t=1}^{k-1} \left(1 + u_{i_t}^{t,\alpha} \right) \right) (u_{i_k}^{k,\alpha})^2 \right] + 3N\varepsilon^2 \exp(-Cd^2) \end{aligned} \tag{6}$$

where we use the notation $\mathbb{E}_{\leq k}[X(i_1, \dots, i_k)] := \frac{1}{d^k} \sum_{i_1, \dots, i_k} \prod_{t=1}^{k} \lambda_{i_k}^k X(i_1, \dots, i_k)$. Observe that for non-adaptive strategies, we can simplify these large products using the fact $u_{i_t}^{t,\alpha}$ does not depend on $(i_1, i_2, \dots, i_{t-1})$. We obtain in this case an upper bound on the mutual information:

$$\mathcal{I}(X : I_1, \dots, I_N) \leq 3 \sum_{k=1}^{N} \mathbb{E}_k \mathbb{E}_\alpha \left[(u_{i_k}^{k,\alpha})^2 \right] + 3N\varepsilon^2 \exp(-Cd^2).$$

For this expression, using methods similar to the proof of Theorem 4.1, one can obtain a bound of the form $\mathbb{E}_k\mathbb{E}_\alpha\left[(u_{i_k}^{k,\alpha})^2\right] = \mathcal{O}(\frac{\varepsilon^2}{d})$ which would lead to a lower bound of $\Omega(\frac{d^3}{\varepsilon^2})$ as in Theorem 4.1. However, for adaptive strategies, we can not simplify the terms $(1 + u_{i_t}^{t,\alpha})$ for $t < k$ because $(u_{i_k}^{k,\alpha})^2$ depends on the previous observations $(i_1, \dots, i_{k-1})$. For this reason, we use Gaussian integration by parts (see Theorem A.3) to break the dependency between the variables in the last expectation. Recall that for all t, i_t, $\tilde{\rho}_t = \rho_t - \mathbb{I}/d$ and:

$$u_{i_t}^{t,\alpha} = \frac{2}{\|\alpha\|_2} \sum_{P \in \mathbb{P}_n} \alpha(P)\varepsilon \left\langle \phi_{i_t}^t \right| P\tilde{\rho}_t P \left| \phi_{i_t}^t \right\rangle .$$

Using the fact that $\|\alpha\|_2$ is independent of $\{\alpha(P)/\|\alpha\|_2\}_P$, we can write:

$$\begin{aligned}
&\mathbb{E}\left(\|\alpha\|_2^2\right)\mathbb{E}\left(\left(\prod_{t=1}^{k-1}\left(1+u_{i_t}^{t,\alpha}\right)\right)(u_{i_k}^{k,\alpha})^2\right) \\
&= 2\varepsilon \sum_{P\in\mathbb{P}_n} \left\langle \phi_{i_k}^k \right| P\tilde{\rho}_k P \left| \phi_{i_k}^k \right\rangle \mathbb{E}\left(\|\alpha\|_2^2\right)\mathbb{E}\left(\frac{\alpha(P)}{\|\alpha\|_2}\left(u_{i_k}^{k,\alpha}\right)\prod_{t=1}^{k-1}\left(1+u_{i_t}^{t,\alpha}\right)\right) \\
&= 2\varepsilon \sum_{P\in\mathbb{P}_n} \left\langle \phi_{i_k}^k \right| P\tilde{\rho}_k P \left| \phi_{i_k}^k \right\rangle \mathbb{E}\left(\alpha(P)\left(\|\alpha\|_2 u_{i_k}^{k,\alpha}\right)\prod_{t=1}^{k-1}\left(1+u_{i_t}^{t,\alpha}\right)\right) \\
&= 2\varepsilon \sum_{P\in\mathbb{P}_n} \left\langle \phi_{i_k}^k \right| P\tilde{\rho}_k P \left| \phi_{i_k}^k \right\rangle \mathbb{E}\left(\alpha(P)F(\alpha)\right),
\end{aligned}$$

where $F(\alpha) = \left(\|\alpha\|_2 u_{i_k}^{k,\alpha}\right)\prod_{t=1}^{k-1}\left(1+u_{i_t}^{t,\alpha}\right)$. Gaussian integration by parts (see Theorem A.3) implies:

$$\begin{aligned}
&\mathbb{E}\left(\alpha(P)F(\alpha)\right) = \mathbb{E}\left(\partial_P F(\alpha)\right) \\
&= 2\varepsilon \left\langle \phi_{i_k}^k \right| P\tilde{\rho}_k P \left| \phi_{i_k}^k \right\rangle \mathbb{E}\left(\prod_{t=1}^{k-1}\left(1+u_{i_t}^{t,\alpha}\right)\right) + \sum_{s=1}^{k-1}\mathbb{E}\left(\|\alpha\|_2 u_{i_k}^{k,\alpha}\partial_P u_{i_s}^{s,\alpha}\prod_{t\in[k-1]\setminus s}\left(1+u_{i_t}^{t,\alpha}\right)\right)
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\|\alpha\|_2\partial_P u_{i_s}^{s,\alpha} &= 2\frac{\left\langle \phi_{i_s}^s \right| P\tilde{\rho}_s P \left| \phi_{i_s}^s \right\rangle \varepsilon\|\alpha\|_2 - \partial_P\|\alpha\|_2 \sum_{P\in\mathbb{P}_n}\alpha(P)\left\langle \phi_{i_s}^s \right| P\tilde{\rho}_s P \left| \phi_{i_s}^s \right\rangle \varepsilon}{\|\alpha\|_2} \\
&= 2\left\langle \phi_{i_s}^s \right| P\tilde{\rho}_s P \left| \phi_{i_s}^s \right\rangle \varepsilon - \frac{1}{\|\alpha\|_2}\alpha(P)u_{i_s}^{s,\alpha}
\end{aligned}$$

hence:

$$\begin{aligned}
&\mathbb{E}_{<k}\mathbb{E}\left(\left(\prod_{t=1}^{k-1}\left(1+u_{i_t}^{t,\alpha}\right)\right)(u_{i_k}^{k,\alpha})^2\right) = \mathbb{E}_{<k}\frac{2\varepsilon}{d^2}\sum_{P\in\mathbb{P}_n}\left\langle \phi_{i_k}^k \right| P\tilde{\rho}_k P \left| \phi_{i_k}^k \right\rangle \mathbb{E}\left(\alpha(P)F(\alpha)\right) \\
&= \mathbb{E}_{<k}\frac{4\varepsilon^2}{d^2}\sum_{P\in\mathbb{P}_n}\left\langle \phi_{i_k}^k \right| P\tilde{\rho}_k P \left| \phi_{i_k}^k \right\rangle^2 \mathbb{E}\left(\prod_{t=1}^{k-1}\left(1+u_{i_t}^{t,\alpha}\right)\right) \qquad \text{(L1)} \\
&+ \mathbb{E}_{<k}\frac{4\varepsilon^2}{d^2}\sum_{P\in\mathbb{P}_n}\sum_{s=1}^{k-1}\left\langle \phi_{i_k}^k \right| P\tilde{\rho}_k P \left| \phi_{i_k}^k \right\rangle \left\langle \phi_{i_s}^s \right| P\tilde{\rho}_s P \left| \phi_{i_s}^s \right\rangle \mathbb{E}\left(u_{i_k}^{k,\alpha}\prod_{t\in[k-1]\setminus s}\left(1+u_{i_t}^{t,\alpha}\right)\right) \qquad \text{(L2)} \\
&- \mathbb{E}_{<k}\frac{2\varepsilon}{d^2}\sum_{P\in\mathbb{P}_n}\left\langle \phi_{i_k}^k \right| P\tilde{\rho}_k P \left| \phi_{i_k}^k \right\rangle \sum_{s=1}^{k-1}\mathbb{E}\left(\frac{\alpha(P)}{\|\alpha\|_2}u_{i_k}^{k,\alpha}u_{i_s}^{s,\alpha}\prod_{t\in[k-1]\setminus s}\left(1+u_{i_t}^{t,\alpha}\right)\right). \qquad \text{(L3)}
\end{aligned}$$

We analyze the latter expressions line by line. Our goal is to upper bound these terms better with some expression improving the naive upper bound $\mathcal{O}(\varepsilon^2)$ on the conditional mutual information. Let us start by line (L1), we have

$$\sum_{P\in\mathbb{P}_n}\frac{1}{d}\sum_{i_k}\lambda_{i_k}^k\left\langle \phi_{i_k}^k \right| P\tilde{\rho}_k P \left| \phi_{i_k}^k \right\rangle^2 \le \sum_{P\in\mathbb{P}_n}\frac{1}{d}\cdot\operatorname{tr}(P\tilde{\rho}_k^2 P) = \sum_{P\in\mathbb{P}_n}\frac{1}{d}\cdot\operatorname{tr}(\tilde{\rho}_k^2) \le d,$$

so using $\frac{1}{d}\sum_{i_t}\lambda_{i_k}^k(1+u_{i_t}^{t,\alpha})=1$ we can upper bound the line (L1) as follows:

$$(\text{L1}) \leq \mathbb{E}_{\leq k-1}\frac{4\varepsilon^2}{d}\mathbb{E}\left(\prod_{t=1}^{k-1}\left(1+u_{i_t}^{t,\alpha}\right)\right)=\frac{4\varepsilon^2}{d}.$$

This upper bound has the same order as for non-adaptive strategies. So we expect that the contribution of line (L1) will not affect much the overall upper bound on the conditional mutual information. Next we move to line (L3), first we show a useful inequality:

Lemma 5.5. *Let $t\in[N]$. Recall that $u_{i_t}^{t,\alpha}=\frac{2}{\|\alpha\|_2}\sum_{P\in\mathbb{P}_n}\alpha(P)\varepsilon\left\langle\phi_{i_t}^t\right|P\tilde{\rho}_tP\left|\phi_{i_t}^t\right\rangle$. We have:*

$$\frac{1}{d}\sum_{i_t}\lambda_{i_t}^t(u_{i_t}^{t,\alpha})^2\leq 16\varepsilon^2.$$

Observe that if we apply this upper bound directly on the expression of the conditional mutual information (Lemma 3.4) we obtain an upper bound $\mathcal{I}=\mathcal{O}(N\varepsilon^2)$ which leads to a lower bound $N=\Omega(d^2/\varepsilon^2)$ similar to Theorem 3.1. Still this Lemma will be useful for controlling intermediate expressions appearing for the upper bound of line (L3).

Proof. We use the fact that every M can be written as $M=\sum_{R\in\mathbb{P}_n}\frac{\mathrm{tr}(MR)}{d}R$ and Lemma A.1

$$\begin{aligned}
\sum_{i_t}\lambda_{i_t}^t(u_{i_t}^{t,\alpha})^2 &= \frac{4\varepsilon^2}{\|\alpha\|_2^2}\sum_{i_t}\lambda_{i_t}^t\left\langle\phi_{i_t}^t\right|\left(\sum_{P\in\mathbb{P}_n}\alpha(P)P\tilde{\rho}_tP\right)\left|\phi_{i_t}^t\right\rangle^2\leq\frac{4\varepsilon^2}{\|\alpha\|_2^2}\mathrm{tr}\left(\sum_{P\in\mathbb{P}_n}\alpha(P)P\tilde{\rho}_tP\right)^2\\
&=\frac{4\varepsilon^2}{\|\alpha\|_2^2}\mathrm{tr}\left(\sum_{P\in\mathbb{P}_n}\alpha(P)\frac{1}{d}\sum_{R\in\mathbb{P}_n}\mathrm{tr}(R\tilde{\rho}_t)PRP\right)^2=\frac{4\varepsilon^2}{\|\alpha\|_2^2}\mathrm{tr}\left(\sum_{P\in\mathbb{P}_n}\alpha(P)\frac{1}{d}\sum_{R\in\mathbb{P}_n}\mathrm{tr}(R\tilde{\rho}_t)(-1)^{R.P}R\right)^2\\
&=\frac{4\varepsilon^2}{\|\alpha\|_2^2}\sum_{P,P',R,R'\in\mathbb{P}_n}\alpha(P)\alpha(P')\frac{1}{d^2}\mathrm{tr}(R\tilde{\rho}_t)\mathrm{tr}(R'\tilde{\rho}_t)(-1)^{R.P}(-1)^{R'.P'}\mathrm{tr}(RR')\\
&=\frac{4\varepsilon^2}{\|\alpha\|_2^2}\sum_{P,P',R\in\mathbb{P}_n}\alpha(P)\alpha(P')\frac{1}{d}\cdot\mathrm{tr}(R\tilde{\rho}_t)^2(-1)^{R.P}(-1)^{R.P'}\\
&=\frac{4\varepsilon^2}{\|\alpha\|_2^2}\sum_{R\in\mathbb{P}_n}\left(\sum_{P\in\mathbb{P}_n}\alpha(P)(-1)^{R.P}\right)^2\frac{1}{d}\cdot\mathrm{tr}(R\tilde{\rho}_t)^2\leq\frac{16\varepsilon^2}{\|\alpha\|_2^2}\sum_{P,P',R\in\mathbb{P}_n}\alpha(P)\alpha(P')\frac{1}{d}(-1)^{R.(PP')}\\
&=\frac{16\varepsilon^2}{\|\alpha\|_2^2}\sum_{P,P'\in\mathbb{P}_n}\alpha(P)\alpha(P')\cdot d\cdot\mathbb{1}_{PP'=\mathbb{I}}=\frac{16d\varepsilon^2}{\|\alpha\|_2^2}\sum_{P=P'\in\mathbb{P}_n}\alpha(P)^2=16d\varepsilon^2.
\end{aligned}$$

In the previous inequality, we used that for all $R\in\mathbb{P}_n$: we can write $R=\sum_i\lambda_i|\phi_i\rangle\langle\phi_i|$ where the eigenvalues $\{\lambda_i\}_i$ have absolute values 1, then by the triangle inequality:

$$\begin{aligned}
|\mathrm{tr}(R\tilde{\rho})| &= \left|\sum_i\lambda_i\mathrm{tr}(|\phi_i\rangle\langle\phi_i|\,\tilde{\rho})\right|\leq\sum_i|\lambda_i\mathrm{tr}(|\phi_i\rangle\langle\phi_i|\,\tilde{\rho})|\leq\sum_i|\mathrm{tr}(|\phi_i\rangle\langle\phi_i|\,(\rho-\mathbb{I}/d))|\\
&\leq\sum_i|\mathrm{tr}(|\phi_i\rangle\langle\phi_i|\,\rho)|+\sum_i|\mathrm{tr}(|\phi_i\rangle\langle\phi_i|\,\mathbb{I}/d))|=\mathrm{tr}(\rho)+\mathrm{tr}(\mathbb{I}/d)=2.
\end{aligned}$$

□

Observe that the condition $\varepsilon\leq 1/(4d)$ implies that for all $t\in[N]$ and $i_t\in\mathcal{I}_t$ we have $1+u_{i_t}^{t,\alpha}\geq 1/16$

and recall that $\mathbb{E}_k(1+u_{i_t}^{t,\alpha}) = \frac{1}{d}\sum_{i_t}\lambda_{i_t}^t(1+u_{i_t}^{t,\alpha}) = 1$. Therefore, (L3) can be controlled as follows:

$$
\begin{aligned}
(\text{L3}) &= -\mathbb{E}_{\leq k}\frac{1}{d^2}\sum_{s=1}^{k-1}\mathbb{E}\left(\left(u_{i_k}^{k,\alpha}\right)^2 u_{i_s}^{s,\alpha}\prod_{t\in[k-1]\setminus s}(1+u_{i_t}^{t,\alpha})\right)\\
&= \frac{1}{d^2}\mathbb{E}_{\leq k}\left(-\mathbb{E}\left(\left(\sum_{s<k}u_{i_s}^{s,\alpha}(u_{i_k}^{k,\alpha})^2\right)\prod_{t\in[k-1]\setminus s}(1+u_{i_t}^{t,\alpha})\right)\right)\\
&\leq \frac{1}{d^2}\mathbb{E}_{\leq k}\left(\mathbb{E}\left(\left(\left|\sum_{s<k}u_{i_s}^{s,\alpha}\right|(u_{i_k}^{k,\alpha})^2\right)\prod_{t\in[k-1]\setminus s}(1+u_{i_t}^{t,\alpha})\right)\right)\\
&\leq \frac{16\varepsilon^2}{d^2}\mathbb{E}_{<k}\left(\mathbb{E}\left(\left(\left|\sum_{s<k}\frac{u_{i_s}^{s,\alpha}}{(1+u_{i_s}^{s,\alpha})}\right|\right)\prod_{t<k}(1+u_{i_t}^{t,\alpha})\right)\right) && \text{(Lemma. 5.5)}\\
&\leq \frac{16\varepsilon^2}{d^2}\sqrt{\mathbb{E}_\alpha\mathbb{E}_{<k}\left|\sum_{s<k}\frac{u_{i_s}^{s,\alpha}}{(1+u_{i_s}^{s,\alpha})}\right|^2\prod_{t<k}(1+u_{i_t}^{t,\alpha})}\sqrt{\mathbb{E}_\alpha\mathbb{E}_{<k}\prod_{t<k}(1+u_{i_t}^{t,\alpha})} && \text{(Cauchy-Schwartz)}\\
&\leq \frac{16\varepsilon^2}{d^2}\sqrt{\mathbb{E}_\alpha\mathbb{E}_{<k}\sum_{s,r<k}\frac{u_{i_s}^{s,\alpha}}{(1+u_{i_s}^{s,\alpha})}\cdot\frac{u_{i_r}^{r,\alpha}}{(1+u_{i_r}^{r,\alpha})}\prod_{t<k}(1+u_{i_t}^{t,\alpha})} && \left(\mathbb{E}_{<k}\prod_{t<k}(1+u_{i_t}^{t,\alpha})=1\right)\\
&\leq \frac{16\varepsilon^2}{d^2}\sqrt{\mathbb{E}_\alpha\mathbb{E}_{<k}\sum_{s<k}\frac{(u_{i_s}^{s,\alpha})^2}{(1+u_{i_s}^{s,\alpha})^2}\prod_{t<k}(1+u_{i_t}^{t,\alpha})} && \left(\mathbb{E}_{\leq\max\{s,r\}}\left(u_{i_s}^{s,\alpha}u_{i_r}^{r,\alpha}\right)=0 \text{ if } s\neq r\right)\\
&\leq \frac{64\varepsilon^2}{d^2}\sqrt{\mathbb{E}_\alpha\mathbb{E}_{<k}\sum_{s<k}(u_{i_s}^{s,\alpha})^2\prod_{t\in[k-1]\setminus s}(1+u_{i_t}^{t,\alpha})} \leq \sqrt{k}\frac{256\varepsilon^3}{d^2}, && (1+u_{i_t}^{t,\alpha}\geq 1/16 \Leftarrow \varepsilon\leq 1/4d)
\end{aligned}
$$

where we use also Lemma 5.5 for the last inequality. Indeed, we can simplify the expectation as follows

$$
\begin{aligned}
\mathbb{E}_{<k}\sum_{s<k}(u_{i_s}^{s,\alpha})^2\prod_{t\in[k-1]\setminus s}(1+u_{i_t}^{t,\alpha}) &= \sum_{s<k}\mathbb{E}_{<k}(u_{i_s}^{s,\alpha})^2\prod_{t\in[k-1]\setminus s}(1+u_{i_t}^{t,\alpha})\\
&= \sum_{s<k}\mathbb{E}_{\leq s}(u_{i_s}^{s,\alpha})^2\prod_{t\in[s-1]}(1+u_{i_t}^{t,\alpha})\\
&\leq \sum_{s<k}\mathbb{E}_{\leq s-1}16\varepsilon^2\prod_{t\in[s-1]}(1+u_{i_t}^{t,\alpha}) && \text{(Lemma 5.5)}\\
&= \sum_{s<k}16\varepsilon^2 \leq 16\varepsilon^2 k.
\end{aligned}
$$

Finally, we control the line (L2) which is more involved. Let us adopt the notation for $s,k\in[N]$:

$$
\begin{aligned}
M_{s,k} &= \sum_{P\in\mathbb{P}_n}P\tilde{\rho}_kP\left\langle\phi_{i_s}^s\right|P\tilde{\rho}_sP\left|\phi_{i_s}^s\right\rangle\\
&= \frac{1}{d^2}\sum_{P,Q,R\in\mathbb{P}_n}\mathrm{tr}(\tilde{\rho}_kQ)\mathrm{tr}(\tilde{\rho}_sR)PQP\left\langle\phi_{i_s}^s\right|PRP\left|\phi_{i_s}^s\right\rangle\\
&= \frac{1}{d^2}\sum_{P,Q,R\in\mathbb{P}_n}\mathrm{tr}(\tilde{\rho}_kQ)\mathrm{tr}(\tilde{\rho}_sR)(-1)^{P.(QR)}Q\left\langle\phi_{i_s}^s\right|R\left|\phi_{i_s}^s\right\rangle\\
&= \sum_{Q\in\mathbb{P}_n}\mathrm{tr}(\tilde{\rho}_kQ)\mathrm{tr}(\tilde{\rho}_sQ)\left\langle\phi_{i_s}^s\right|Q\left|\phi_{i_s}^s\right\rangle Q && \text{(Lemma A.1)}
\end{aligned}
$$

so that we can write $\sum_{P\in\mathbb{P}_n}\left\langle\phi_{i_k}^k\right|P\tilde{\rho}_kP\left|\phi_{i_k}^k\right\rangle\left\langle\phi_{i_s}^s\right|P\tilde{\rho}_sP\left|\phi_{i_s}^s\right\rangle = \mathrm{tr}(\left|\phi_{i_k}^k\right\rangle\left\langle\phi_{i_k}^k\right|M_{s,k})$. Also we use the

notation $\Psi_k = \mathbb{E}_{\leq k}\mathbb{E}\left(\left(u_{i_k}^{k,\alpha}\right)^2 \prod_{t<k}\left(1+u_{i_t}^{t,\alpha}\right)\right)$ so that we have the (in)equalities:

$$
\begin{aligned}
\text{(L2)} &= \frac{4\varepsilon^2}{d^2}\mathbb{E}_{\leq k}\mathbb{E}\left(\sum_{s=1}^{k-1}\operatorname{tr}(\left|\phi_{i_k}^k\right\rangle\!\left\langle\phi_{i_k}^k\right| M_{s,k})u_{i_k}^{k,\alpha}\prod_{t\in[k-1]\setminus s}\left(1+u_{i_t}^{t,\alpha}\right)\right)\\
&= \frac{4\varepsilon^2}{d^2}\mathbb{E}_{\leq k}\mathbb{E}\left(\operatorname{tr}\left(\left|\phi_{i_k}^k\right\rangle\!\left\langle\phi_{i_k}^k\right|\sum_{s=1}^{k-1}\frac{M_{s,k}}{(1+u_{i_s}^{s,\alpha})}\right)u_{i_k}^{k,\alpha}\prod_{t\leq k-1}\left(1+u_{i_t}^{t,\alpha}\right)\right)\\
&\leq \frac{4\varepsilon^2}{d^2}\sqrt{\mathbb{E}_{\leq k}\mathbb{E}\left(\left(\operatorname{tr}\left(\left|\phi_{i_k}^k\right\rangle\!\left\langle\phi_{i_k}^k\right|\sum_{s=1}^{k-1}\frac{M_{s,k}}{(1+u_{i_s}^{s,\alpha})}\right)\right)^2\prod_{t\leq k-1}\left(1+u_{i_t}^{t,\alpha}\right)\right)}\sqrt{\Psi_k} \quad \text{(Cauchy-Schwartz)}\\
&\leq \frac{4\varepsilon^2}{d^2}\sqrt{\mathbb{E}_{\leq k}\mathbb{E}\left(\operatorname{tr}\left(\left|\phi_{i_k}^k\right\rangle\!\left\langle\phi_{i_k}^k\right|\left(\sum_{s=1}^{k-1}\frac{M_{s,k}}{(1+u_{i_s}^{s,\alpha})}\right)^2\right)\prod_{t\leq k-1}\left(1+u_{i_t}^{t,\alpha}\right)\right)}\sqrt{\Psi_k} \quad \text{(Cauchy-Schwartz)}\\
&= \frac{4\varepsilon^2}{d^2}\sqrt{\frac{1}{d}\mathbb{E}_{\leq k-1}\mathbb{E}\left(\operatorname{tr}\left(\left(\sum_{s=1}^{k-1}\frac{M_{s,k}}{(1+u_{i_s}^{s,\alpha})}\right)^2\right)\prod_{t\leq k-1}\left(1+u_{i_t}^{t,\alpha}\right)\right)}\sqrt{\Psi_k}.
\end{aligned}
$$

From the definition of $M_{s,k}$ we can write that for $s,t<k$

$$\operatorname{tr}(M_{s,k}M_{t,k}) = d\sum_{Q\in\mathbb{P}_n}\operatorname{tr}(\tilde{\rho}_k Q)^2\operatorname{tr}(\tilde{\rho}_s Q)\operatorname{tr}(\tilde{\rho}_t Q)\left\langle\phi_{i_s}^s\right| Q\left|\phi_{i_s}^s\right\rangle\left\langle\phi_{i_t}^t\right| Q\left|\phi_{i_t}^t\right\rangle.$$

Hence

$$
\begin{aligned}
\operatorname{tr}\left(\sum_{s=1}^{k-1}\frac{M_{s,k}}{(1+u_{i_s}^{s,\alpha})}\right)^2 &= \sum_{s,t<k} d\sum_{Q\in\mathbb{P}_n}\operatorname{tr}(\tilde{\rho}_k Q)^2\frac{\operatorname{tr}(\tilde{\rho}_s Q)\operatorname{tr}(\tilde{\rho}_t Q)\left\langle\phi_{i_s}^s\right| Q\left|\phi_{i_s}^s\right\rangle\left\langle\phi_{i_t}^t\right| Q\left|\phi_{i_t}^t\right\rangle}{(1+u_{i_s}^{s,\alpha})(1+u_{i_t}^{t,\alpha})}\\
&= d\sum_{Q\in\mathbb{P}_n}\operatorname{tr}(\tilde{\rho}_k Q)^2\left(\sum_{s<k}\frac{\operatorname{tr}(\tilde{\rho}_s Q)\left\langle\phi_{i_s}^s\right| Q\left|\phi_{i_s}^s\right\rangle}{(1+u_{i_s}^{s,\alpha})}\right)^2\\
&\leq 4d\sum_{Q\in\mathbb{P}_n}\left(\sum_{s<k}\frac{\operatorname{tr}(\tilde{\rho}_s Q)\left\langle\phi_{i_s}^s\right| Q\left|\phi_{i_s}^s\right\rangle}{(1+u_{i_s}^{s,\alpha})}\right)^2
\end{aligned}
$$

note that this step is crucial because ρ_k depends on $(i_1,\ldots,i_{k-1})$ so we need to avoid it in order to simplify with the expectations $\mathbb{E}_t$ for $t<k$. When we want to simplify the expectation $\mathbb{E}_{\leq k-1}\mathbb{E}\left(\operatorname{tr}\left(\left(\sum_{s=1}^{k-1}\frac{M_{s,k}}{(1+u_{i_s}^{s,\alpha})}\right)^2\right)\prod_{t\leq k-1}\left(1+u_{i_t}^{t,\alpha}\right)\right)$ and we expand the square of the latter expression, we can see that if $s_1<s_2$ (or $s_1>s_2$), we'll get 0 because we can simplify the terms $(1+u_{i_t}^{t,\alpha})$ in the product for $t>s_2$, the term $(1+u_{i_{s_2}}^{s_2,\alpha})$ is simplified with the denominator so we can take safely the expectation under $\mathbb{E}_{s_2}$:

$$\mathbb{E}_{s_2}\operatorname{tr}(\tilde{\rho}_{s_2}Q)\left\langle\phi_{i_{s_2}}^{s_2}\right| Q\left|\phi_{i_{s_2}}^{s_2}\right\rangle = \frac{1}{d}\sum_{i_{s_2}}\operatorname{tr}(\tilde{\rho}_{s_2}Q)\lambda_{i_{s_2}}^{s_2}\left\langle\phi_{i_{s_2}}^{s_2}\right| Q\left|\phi_{i_{s_2}}^{s_2}\right\rangle = \operatorname{tr}(\tilde{\rho}_{s_2}Q)\operatorname{tr}(Q) = 0$$

because $\mathrm{tr}(Q) = 0$ unless $Q = \mathbb{I}$ for which $\mathrm{tr}(\tilde{\rho}_{s_2} Q) = \mathrm{tr}(\tilde{\rho}_{s_2}) = \mathrm{tr}(\rho_{s_2} - \mathbb{I}/d) = 0$. Therefore

$$
\begin{aligned}
\text{(L2)} &\leq \frac{4\varepsilon^2}{d^2} \sqrt{\frac{1}{d}\mathbb{E}_{\leq k-1}\mathbb{E}\left(\mathrm{tr}\left(\sum_{s=1}^{k-1} \frac{M_{s,k}}{(1+u_{i_s}^{s,\alpha})}\right)^2 \prod_{t\leq k-1}\left(1+u_{i_t}^{t,\alpha}\right)\right)}\sqrt{\Psi_k} \\
&\leq \frac{4\varepsilon^2}{d^2} \sqrt{\frac{1}{d}\mathbb{E}_{\leq k-1}\mathbb{E}\left(4d\sum_{Q\in\mathbb{P}_n}\left(\sum_{s<k}\frac{\mathrm{tr}(\tilde{\rho}_s Q)\left\langle \phi_{i_s}^s\right| Q \left|\phi_{i_s}^s\right\rangle}{(1+u_{i_s}^{s,\alpha})}\right)^2 \prod_{t\leq k-1}\left(1+u_{i_t}^{t,\alpha}\right)\right)}\sqrt{\Psi_k} \\
&\leq \frac{4\varepsilon^2}{d^2} \sqrt{\sum_{s<k} 16 \sum_{Q\in\mathbb{P}_n}\mathbb{E}_{\leq s}\mathrm{tr}(\tilde{\rho}_s Q)^2 \left\langle \phi_{i_s}^s\right| Q \left|\phi_{i_s}^s\right\rangle^2 \mathbb{E}\left(\prod_{t\leq s-1}\left(1+u_{i_t}^{t,\alpha}\right)\right)}\sqrt{\Psi_k} \qquad \left((1+u_{i_s}^{s,\alpha}) \geq \frac{1}{16}\right) \\
&\leq \frac{4\varepsilon^2}{d^2} \sqrt{\sum_{s<k} 64\mathbb{E}_{\leq s}\sum_{Q\in\mathbb{P}_n}\left\langle \phi_{i_s}^s\right| Q \left|\phi_{i_s}^s\right\rangle\left\langle \phi_{i_s}^s\right| Q \left|\phi_{i_s}^s\right\rangle \mathbb{E}\left(\prod_{t\leq s-1}\left(1+u_{i_t}^{t,\alpha}\right)\right)}\sqrt{\Psi_k} \qquad (|\mathrm{tr}(\tilde{\rho}_s Q)| \leq 2) \\
&= \frac{4\varepsilon^2}{d^2} \sqrt{\sum_{s<k} 64\mathbb{E}_{\leq s}\left\langle \phi_{i_s}^s\right| d\mathrm{tr}(\left|\phi_{i_s}^s\right\rangle\left\langle \phi_{i_s}^s\right|)\mathbb{I}\left|\phi_{i_s}^s\right\rangle \mathbb{E}\left(\prod_{t\leq s-1}\left(1+u_{i_t}^{t,\alpha}\right)\right)}\sqrt{\Psi_k} \qquad \text{(Lemma A.2)} \\
&\leq \frac{32\varepsilon^2}{d\sqrt{d}}\sqrt{k}\sqrt{\Psi_k}.
\end{aligned}
$$

We have proven so far, for all $k \leq N$:

$$
\Psi_k \leq \frac{4\varepsilon^2}{d} + 256\sqrt{k}\frac{\varepsilon^3}{d^2} + \frac{32\varepsilon^2}{d\sqrt{d}}\sqrt{k}\sqrt{\Psi_k}. \tag{7}
$$

The first term of the upper bound can be seen as a non-adaptive contribution. The second one can be thought as a geometric mean of the first and third terms. The last term represents essentially the contribution of the adaptivity. Our final stage of the proof is to use these recurrence inequalities to prove the lower bound by a contradiction argument.

Recall that $\Psi_k = \mathbb{E}_{\leq k}\mathbb{E}\left(\left(u_{i_k}^{k,\alpha}\right)^2 \prod_{t<k}\left(1+u_{i_t}^{t,\alpha}\right)\right)$ and $\sum_{k=1}^N \mathcal{I}(X : I_k|I_{<k}) \leq 3\sum_{k=1}^N \Psi_k + 3N\varepsilon^2\exp(-Cd^2)$. We suppose that $N \leq c\frac{d^{5/2}}{\varepsilon^2}$ for sufficiently small $c > 0$. We know that from Lemma 5.4 and Lemma 3.4

$$
c_0 d^2 \leq \mathcal{I}(X:Y) \leq 3\sum_{k\leq N}\Psi_k + 3N\varepsilon^2\exp(-Cd^2).
$$

So $\sum_{k\leq N}\Psi_k \geq c'd^2$ (for example $c' = c_0/4$), on the other hand the inequality (7) implies:

$$
\begin{aligned}
\sum_k \Psi_k &\leq \sum_k \frac{4\varepsilon^2}{d} + 256\frac{\varepsilon^3}{d^2}\sqrt{k} + 32\frac{\varepsilon^2\sqrt{k}}{d\sqrt{d}}\sqrt{\Psi_k} \\
&\leq 4\frac{N\varepsilon^2}{d} + 256\frac{N\varepsilon^3}{d^2}\sqrt{N} + 32\sum_k \frac{\varepsilon^2\sqrt{k}}{d\sqrt{d}}\sqrt{\Psi_k} \\
&\leq 4\frac{N\varepsilon^2}{\sqrt{c'}d^2}\sqrt{\sum_{k\leq N}\Psi_k} + 256\frac{N\varepsilon^2}{d^2}\sqrt{cd^{5/2}} + 32\frac{\varepsilon^2}{d\sqrt{d}}\sqrt{\sum_k k}\sqrt{\sum_k \Psi_k} \qquad \text{(Cauchy-Schwartz)} \\
&\leq \left(\frac{8}{\sqrt{c'}d} + \frac{512}{\sqrt{c'}d^{1/4}} + 32\right)\frac{\varepsilon^2}{d\sqrt{d}}\sqrt{\sum_k k}\sqrt{\sum_k \Psi_k} \\
&\leq C'\frac{\varepsilon^2}{d\sqrt{d}}N\sqrt{\sum_k \Psi_k}
\end{aligned}
$$

where C' is a universal constant. Therefore:

$$
\sum_k \Psi_k \leq C'^2\left(\frac{N^2\varepsilon^4}{d^3}\right) \leq C'^2 c^2 d^2
$$

Hence

$$c_0 d^2 \leq \mathcal{I}(X:Y) \leq \sum_{k \leq N} 3\Psi_k + 3N\varepsilon^2 \exp(-Cd^2) \leq 6C'^2 c^2 d^2$$

which gives the contradiction for $c \ll \sqrt{c_0}/C'$. Finally we deduce $N = \Omega(d^{5/2}/\varepsilon^2)$ and we conclude the proof of Theorem 5.1.

If the adaptive algorithm has a memory of $\mathcal{O}(H/\varepsilon^2)$, the previous inequalities imply for all $1 \leq k \leq N$:

$$\begin{aligned}\sum_k \Psi_k &\leq \sum_k \frac{4\varepsilon^2}{d} + 256\sqrt{H}\frac{\varepsilon^2}{d^2} + \frac{32\varepsilon}{d\sqrt{d}}\sqrt{H}\sqrt{\Psi_k} \\ &\leq \sum_k \frac{4\varepsilon^2}{d} + 256\sqrt{H}\frac{\varepsilon^2}{d^2} + \frac{(32\varepsilon)^2 H}{d^3} + \frac{\Psi_k}{2}\end{aligned}$$

where we use AM-GM inequality, hence we deduce:

$$\begin{aligned}c_0 d^2 \leq \mathcal{I}(X:Y) &\leq \sum_k 3\Psi_k + 3N\varepsilon^2 \exp(-Cd^2) \\ &\leq \frac{30\varepsilon^2 N}{d} + 6 \cdot 256\sqrt{H}\frac{\varepsilon^2 N}{d^2} + \frac{6 \cdot (32\varepsilon)^2 HN}{d^3}\end{aligned}$$

and finally we obtain:

$$N = \Omega\left(\min\left\{\frac{d^4}{\sqrt{H}\varepsilon^2}, \frac{d^5}{H\varepsilon^2}, \frac{d^3}{\varepsilon^2}\right\}\right).$$

For $H = \mathcal{O}(d^2)$, this gives Eq. (5). □

6 Conclusion and open problems

We have provided lower bounds for Pauli channel tomography in the diamond norm using independent strategies for both adaptive and non-adaptive strategies. In particular, we have shown that the number of measurements should be at least $\Omega(d^3/\varepsilon^2)$ in the non-adaptive setting and $\Omega(d^{2.5}/\varepsilon^2)$ in the adaptive setting. We would like to finish with three interesting directions. Finding the optimal complexity of Pauli channel tomography using adaptive incoherent measurements remains an open question. We conjecture this complexity to be $\Theta(d^3/\varepsilon^2)$ since we remark that in many situations the adaptive strategies cannot overcome the non-adaptive ones. Furthermore, we already obtained a $\Theta(d^3/\varepsilon^2)$ bound for adaptive strategies in the high precision and sub-exponential memory regime, further evidence of this bound. Moreover, since Chen, Zhou, et al. (2022) established the optimal complexity for estimating the eigenvalues of a Pauli channel in the l_∞-norm using ancilla-assisted non-adaptive independent strategies, it would be interesting to find the optimal complexity to learn a Pauli channel in the diamond norm when the algorithm can use k-qubit ancilla for $k \leq n$. Finally, it should be noted that all of the channel constructions used in this work have a very large spectral gap, i.e. are very noisy. It would be interesting to study the sample complexity of Pauli channel tomography in terms of the spectral gap as well.

Acknowledgments

Aadil Oufkir thanks Guillaume Aubrun for helpful discussions. This work is part of HQI initiative (www.hqi.fr) and is supported by France 2030 under the French National Research Agency award number "ANR-22-PNCQ-0002". We also acknowledge support from the European Research Council (ERC Grant AlgoQIP, Agreement No. 851716).

References

Arute, F., Arya, K., Babbush, R., Bacon, D., Bardin, J. C., Barends, R., ... Martinis, J. M. (2019). Quantum supremacy using a programmable superconducting processor. *Nature*, *574*(7779), 505–510. Retrieved from https://doi.org/10.1038/s41586-019-1666-5 DOI: 10. 1038/s41586-019-1666-5

Bandyopadhyay, S., Boykin, P. O., Roychowdhury, V., & Vatan, F. (2002). A new proof for the existence of mutually unbiased bases. *Algorithmica*, *34*(4), 512–528.

Blume-Kohout, R., Gamble, J. K., Nielsen, E., Mizrahi, J., Sterk, J. D., & Maunz, P. (2013). *Robust, self-consistent, closed-form tomography of quantum logic gates on a trapped ion qubit.* arXiv. Retrieved from https://arxiv.org/abs/1310.4492 DOI: 10.48550/ARXIV.1310.4492

Born, M. (1926, December). Zur Quantenmechanik der Stoßvorgänge. *Zeitschrift fur Physik*, *37*(12), 863-867. DOI: 10.1007/BF01397477

Chen, S., Huang, B., Li, J., Liu, A., & Sellke, M. (2022). Tight bounds for state tomography with incoherent measurements. *arXiv preprint arXiv:2206.05265*.

Chen, S., Zhou, S., Seif, A., & Jiang, L. (2022). Quantum advantages for pauli channel estimation. *Physical Review A*, *105*(3), 032435.

Ebadi, S., Wang, T. T., Levine, H., Keesling, A., Semeghini, G., Omran, A., ... Lukin, M. D. (2021, jul). Quantum phases of matter on a 256-atom programmable quantum simulator. *Nature*, *595*(7866), 227–232. DOI: 10.1038/s41586-021-03582-4

Eisert, J., Hangleiter, D., Walk, N., Roth, I., Markham, D., Parekh, R., ... Kashefi, E. (2020, jun). Quantum certification and benchmarking. *Nat. Rev. Phys.*, *2*(7), 382–390. DOI: 10.1038/s42254-020-0186-4

Fano, R. M. (1961). Transmission of information: A statistical theory of communications. *American Journal of Physics*, *29*(11), 793–794.

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

Flammia, S. T., & Wallman, J. J. (2020). Efficient estimation of pauli channels. *ACM Transactions on Quantum Computing*, *1*(1), 1–32.

França, D. S., & Hashagen, A. (2018). Approximate randomized benchmarking for finite groups. *Journal of Physics A: Mathematical and Theoretical*, *51*(39), 395302.

Haah, J., Harrow, A. W., Ji, Z., Wu, X., & Yu, N. (2016). Sample-optimal tomography of quantum states. In *Proceedings of the forty-eighth annual acm symposium on theory of computing* (pp. 913–925).

Harper, R., Flammia, S. T., & Wallman, J. J. (2020). Efficient learning of quantum noise. *Nature Physics*, *16*(12), 1184–1188.

Helsen, J., Roth, I., Onorati, E., Werner, A. H., & Eisert, J. (2022). General framework for randomized benchmarking. *PRX Quantum*, *3*(2), 020357.

Helsen, J., Xue, X., Vandersypen, L. M., & Wehner, S. (2019). A new class of efficient randomized benchmarking protocols. *npj Quantum Information*, *5*(1), 1–9.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, *58*, 13–30. Retrieved from http://links.jstor.org/sici?sici=0162-1459(196303)58:301<13:PIFSOB>2.0.CO;2-D&origin=MSN

Isserlis, L. (1918). On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, *12*(1/2), 134–139.

Magesan, E., Gambetta, J. M., & Emerson, J. (2012). Characterizing quantum gates via randomized benchmarking. *Physical Review A*, *85*(4), 042311.

Mattila, P. (1999). *Geometry of sets and measures in euclidean spaces: fractals and rectifiability* (No. 44). Cambridge university press.

Montanaro, A., & de Wolf, R. (2013). A survey of quantum property testing. *arXiv preprint arXiv:1310.2035*.

Roth, I., Kueng, R., Kimmel, S., Liu, Y.-K., Gross, D., Eisert, J., & Kliesch, M. (2018, Oct). Recovering quantum gates from few average gate fidelities. *Phys. Rev. Lett.*, *121*, 170502. Retrieved from https://link.aps.org/doi/10.1103/PhysRevLett.121.170502 DOI: 10.1103/PhysRevLett.121.170502

Scholl, P., Schuler, M., Williams, H. J., Eberharter, A. A., Barredo, D., Schymik, K. N., ... Browaeys, A. (2021, jul). Quantum simulation of 2d antiferromagnets with hundreds of rydberg atoms. *Nature*, *595*(7866), 233–238. DOI: 10. 1038/s41586-021-03585-1

Van Handel, R. (2014). *Probability in high dimension* (Tech. Rep.). PRINCETON UNIV NJ.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* (Vol. 48). Cambridge University Press.

Wallman, J. J., & Emerson, J. (2016). Noise tailoring for scalable quantum computation via randomized compiling. *Physical Review A*, *94*(5), 052325.

Watrous, J. (2018). *The theory of quantum information.* Cambridge university press.

Wootters, W. K., & Fields, B. D. (1989). Optimal state-determination by mutually unbiased

measurements. *Annals of Physics*, *191*(2), 363–381.
Zhong, H. S., Wang, H., Deng, Y. H., Chen, M. C., Peng, L. C., Luo, Y. H., ... Pan, J. W. (2020, dec). Quantum computational advantage using photons. *Science*, *370*(6523), 1460–1463. DOI: 10.1126/science.abe8770

A Technical tools

A.1 Pauli group properties

In this section, we group some useful properties about the Pauli operators that we need for the proofs in this article.

Lemma A.1. *We have for all $Q \in \{\mathbb{I}, X, Y, Z\}^{\otimes n}$:*

$$\sum_{P \in \mathbb{P}_n} (-1)^{P.Q} = d^2 \cdot \mathbb{1}_{Q=\mathbb{I}}.$$

Proof. It is clear that for $Q = \mathbb{I}$, Q commutes with every $P \in \mathbb{P}_n$ and thus the equality holds. Now, let $Q \in \mathbb{P}_n \setminus \{\mathbb{I}\}$ and we write $Q = Q_1 \otimes \cdots \otimes Q_n$ where for all $i \in [n]$, $Q_i \in \{\mathbb{I}, X, Y, Z\}$ is a Pauli matrix. By the same decomposition for $P \in \mathbb{P}_n$, we can write:

$$\begin{aligned}
\sum_{P \in \mathbb{P}_n} (-1)^{P.Q} &= \sum_{P_1, \ldots, \mathbb{P}_n \in \{\mathbb{I}, X, Y, Z\}} (-1)^{P_1.Q_1 + 2 \cdots + 2 \mathbb{P}_n.Q_n} \\
&= \prod_{i=1}^{n} \sum_{P_i \in \{\mathbb{I}, X, Y, Z\}} (-1)^{P_i.Q_i} \\
&= \prod_{i=1}^{n} 4 \mathbb{1}_{Q_i = \mathbb{I}_2} \\
&= d^2 \mathbb{1}_{Q = \mathbb{I}_d}
\end{aligned}$$

where we have used in the third equality the fact that every non identity Pauli matrix Q_i commutes only with the identity and itself (so it anti-commutes with the two other Pauli matrices) thus the sum $\sum_{P_i \in \{\mathbb{I}, X, Y, Z\}} (-1)^{P_i.Q_i} = 0$. □

Lemma A.2. *We have for all matrices ρ:*

$$\sum_{P \in \{\mathbb{I}, X, Y, Z\}^{\otimes n}} P \rho P = d \mathrm{tr}(\rho) \mathbb{I}.$$

Proof. Let $d = 2^n$ and $\rho \in \mathbb{C}^{d \times d}$. It is known that $\frac{1}{\sqrt{d}} \{\mathbb{I}, X, Y, Z\}^{\otimes n}$ forms an ortho-normal basis of $\mathbb{C}^{d \times d}$ for the Hilbert-Schmidt scalar product. Thus, we can write ρ in this basis:

$$\rho = \sum_{P \in \{\mathbb{I}, X, Y, Z\}^{\otimes n}} \mathrm{tr}\left(\frac{P}{\sqrt{d}} \rho\right) \frac{P}{\sqrt{d}} = \frac{1}{d} \sum_{P \in \{\mathbb{I}, X, Y, Z\}^{\otimes n}} \mathrm{tr}\left(P \rho\right) P.$$

Therefore we can simplify the LHS by using the identity $PQ = (-1)^{P.Q} QP$ for all $P, Q \in \mathbb{P}_n$:

$$\begin{aligned}
\sum_{P \in \mathbb{P}_n} P \rho P &= \frac{1}{d} \sum_{P, Q \in \mathbb{P}_n} \mathrm{tr}(Q\rho) PQP = \frac{1}{d} \sum_{P, Q \in \mathbb{P}_n} \mathrm{tr}(Q\rho) (-1)^{P.Q} QPP \\
&= \sum_{Q \in \mathbb{P}_n} \mathrm{tr}(Q\rho) Q \frac{1}{d} \sum_{P \in \mathbb{P}_n} (-1)^{P.Q} = \sum_{Q \in \mathbb{P}_n} \mathrm{tr}(Q\rho) Q \cdot d \cdot \mathbb{1}_{Q=\mathbb{I}} \\
&= d \mathrm{tr}(\rho) \mathbb{I},
\end{aligned}$$

where we have used Lemma A.1 to obtain the fourth equality. □

A.2 Gaussian integration by parts

Gaussian integration by parts (see e.g. Van Handel (2014)) is a generalization of Isserlis' formula Isserlis (1918).

Theorem A.3. *Let $(X_1, \ldots, X_d)$ be a Gaussian vector and $f : \mathbb{R}^d \to \mathbb{R}$ be a smooth function. We have:*

$$\mathbb{E}\left(X_1 f(X_1, \ldots, X_d)\right) = \sum_{i=1}^{d} \mathrm{Cov}(X_1, X_i) \mathbb{E}\left(\partial_i f(X_1, \ldots, X_d)\right)$$

B Optimal non-adaptive algorithm for learning a Pauli channel with incoherent measurements

In this section, we simplify the algorithm of Flammia and Wallman (2020), and consider only the learning algorithm. In particular, we show that, if we don't have errors in SPAM, only one copy per step is needed. Hence, this algorithm can learn a Pauli channel to within ε in the diamond norm with only $\mathcal{O}\left(\frac{d^3\log(d)}{\varepsilon^2}\right)$ measurements/steps. Since we have proved the lower bound of $\Omega\left(\frac{d^3}{\varepsilon^2}\right)$ measurements in Theorem 4.1 with one use per step, this algorithm is thus optimal up to a logarithmic factor. Note that we only add this part for completeness and we don't claim any new contribution, all the proofs are similar to those in the mentioned references.

Recall the set of Pauli operators $\mathbb{P}_n = \{\mathbb{I}, X, Y, Z\}^{\otimes n}$. Let S be a subset of $\mathbb{P}_n$, we define the commutant of S as the set of Pauli operators that commute with every element in S: $C_S = \{P \in \mathbb{P}_n : \forall Q \in S, PQ = QP\}$.

Before stating the algorithm, we start by some important Lemmas we need:

Lemma B.1. *$\mathbb{P}_n$ can be covered by $d+1$ stabilizer (Abelian) groups $G_1, \ldots, G_{d+1}$ satisfying for all $i \neq j$:*

- $|G_i| = d$,
- $C_{G_i} = G_i$,
- $G_i \cap G_j = \{\mathbb{I}\}$.

The proof is taken from Bandyopadhyay, Boykin, Roychowdhury, and Vatan (2002) and Wootters and Fields (1989).

Proof. Here we use the correspondence between $\mathbb{P}_n$ and $\left(\mathbb{Z}_2^2\right)^n \simeq \mathbb{Z}_2^{2n}$. We can encode:

$$\begin{aligned}\mathbb{I} &\mapsto (0,0) \in \mathbb{Z}_2^2,\\ X &\mapsto (1,0) \in \mathbb{Z}_2^2,\\ Y &\mapsto (1,1) \in \mathbb{Z}_2^2,\\ Z &\mapsto (0,1) \in \mathbb{Z}_2^2\end{aligned}$$

and we generalize to $\mathbb{P}_n$ by concatenating the encoding of each tensor. We define the inner product of $a = (a_1, a_2), b = (b_1, b_2) \in \mathbb{Z}_2^2$ as follows:

$$a \times b = a_1 b_2 + a_2 b_1 \mod (2).$$

We generalize to $a, b \in \left(\mathbb{Z}_2^2\right)^n$:

$$a \times b = a_1 \times b_1 + \cdots + a_n \times b_n \mod (2).$$

It is not difficult to see that P and Q commute iff their corresponding images a and b have inner product 0. We group the first coordinates together then the second coordinates by using the isomorphism $\left(\mathbb{Z}_2^2\right)^n \simeq \left(\mathbb{Z}_2^n\right)^2 : a \mapsto (\alpha|\beta) = (((a_1)_1, \ldots (a_n)_1)|((a_1)_2, \ldots (a_n)_2)$. Moreover, by specifying an $n \times (2n)$ matrix $(A|B)$ we can construct the corresponding set of Pauli operators as the preimage of the linear combinations of $\{(A_i|B_i)\}_i$ where M_i denotes the i^{th} row of the matrix M. For instance, the corresponding set of $(0_n|\mathbb{I}_n)$ is the Z only Pauli operators. The partition we are looking for will be given by $\{G_i\}_{i=1}^{d+1} = \{\{\mathbb{1}\} \cup \mathcal{C}_i\}_{i=1}^{d+1}$ Bandyopadhyay et al. (2002) where $\{\mathcal{C}_i\}_{i=1}^{d+1}$ are the sets of non identity Pauli operators corresponding to matrices of the form:

$$(0_n|\mathbb{I}_n), (\mathbb{I}_n|A_1), \ldots, (\mathbb{I}_n|A_d).$$

It can be shown that in order to have G_i a stabilizer group, it is sufficient to have A_i a symmetric matrix and $|\mathcal{C}_i| = d - 1$. The former condition implies that for all $k, l \in [n]$, the preimages of $(\mathbb{I}_n|A_i)_k$ and $(\mathbb{I}_n|A_i)_l$ commute. Indeed, $(\mathbb{I}_n|A_i)_k \times (\mathbb{I}_n|A_i)_l = (A_i)_{k,l} + (A_i)_{l,k} = 0 \mod (2)$. The latter condition is satisfied since the matrix $(\mathbb{I}_n|A_i)$ has rank n and thus generates $2^n - 1$ non identity Pauli operators (the preimages of $\sum_{j=1}^n \alpha_j (\mathbb{I}_n|A_i)_j$ where $\{\alpha_j\}_{j\in[n]} \in \{0,1\}^n \setminus \{0\}$). Also, the condition of $|\mathcal{C}_i| = d - 1$ implies that $G_i = \{\mathbb{1}\} \cup \mathcal{C}_i$ is a group, since the maximal cardinal of a stabilizer set is d Bandyopadhyay et al. (2002). The same argument shows that the commutant of G_i is itself.

Now, we want to have $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$. If this is not the case, then we can find $\alpha \in \mathbb{Z}_2^n$ such that $\alpha(\mathbb{I}_n|A_i) = \alpha(\mathbb{I}_n|A_j)$ which is equivalent to $\alpha(A_i - A_j) = 0$. So in order to avoid this situation, we would like to have for all $i \neq j$, $\det(A_i - A_j) \neq 0$. Let $B_1, \dots, B_n \in \mathbb{Z}_2^{n \times n}$ be n symmetric matrices. If we have for all $\alpha \in \mathbb{Z}_2^n \setminus \{0\}$:

$$\det\left(\sum_{i=1}^n \alpha_i B_i\right) \neq 0$$

then we can choose $A_i = \sum_{k=1}^n i_k B_k$ where $i = (i_1, \dots, i_n)$ is the expansion of i in the binary basis. These $\{A_i\}_{i=1}^{2^n}$ have the wanted conditions. Such construction of $\{B_i\}_{i=1}^n$ can be found in Wootters and Fields (1989). Let $f_1, \dots, f_n$ be a basis of $\mathbb{F}_{2^n}$ as a vector space over $\mathbb{Z}_2$. Then we can write for all $i, j \in [n]$:

$$f_i f_j = \sum_{k=1}^n B_{i,j}^{(k)} f_k.$$

Then $B_k = \left(B_{i,j}^{(k)}\right)_{i,j\in[n]}$ satisfy the wanted condition Wootters and Fields (1989). We can verify this, let $\alpha \in \mathbb{Z}_2^n \setminus \{0\}$, we have:

$$\sum_{k=1}^n \alpha_k B_k = \left(\sum_k \alpha_k B_{i,j}^{(k)}\right)_{i,j\in[n]}.$$

Suppose that $\det\left(\sum_{k=1}^n \alpha_k B_k\right) = 0$, so there is $x \in \mathbb{Z}_2^n \setminus \{0\}$ such that $\sum_{k=1}^n \alpha_k B_k x^T = 0$. Let $y = \left(\sum_k \alpha_k f_k\right)\left(\sum_k x_k f_k\right)^{-1} = \sum_k y_k f_k \in \mathbb{F}_{2^n}$. We have for all $i \in [n]$, $\sum_{k,j} \alpha_k B_{i,j}^{(k)} x_j = 0$ so $\sum_i y_i \sum_{k,j} \alpha_k B_{i,j}^{(k)} x_j = 0$ therefore:

$$\sum_{k,i,j} y_i \alpha_k B_{i,j}^{(k)} x_j = 0.$$

On the other hand, the definition of y implies $\sum_{j,k,i} y_i x_j B_{i,j}^{(k)} f_k = \sum_j \left(\sum_i y_i f_i\right) x_j f_j = \sum_k \alpha_k f_k$ hence $\sum_{i,j} y_i x_j B_{i,j}^{(k)} = \alpha_k$ and therefore $\sum_k \alpha_k^2 = \sum_{k,i,j} y_i \alpha_k B_{i,j}^{(k)} x_j = 0$ finally $\alpha = 0$ which is a contradiction. □

Lemma B.2. *Let $G \in \{G_1, \dots, G_{d+1}\}$. Denote by $A_G = \mathbb{P}_n/C_G$, we have $\mathbb{P}_n = G \oplus \mathbb{P}_n/G$, $C_{A_G} = \mathbb{P}_n/G$, and*

$$\frac{1}{|G|} \sum_{P\in G} (-1)^{P.Q} = \mathbb{1}\{Q \in C_G\}.$$

Proof. It is clear that when $Q \in C_G$, the identity holds since for all $P \in G$, $P.Q = 0$. Let $Q \notin C_G$, so we can find $P \in G$ such that $Q.P = 1$ i.e. $QP = -PQ$. Let C (resp. A) be the set of elements of G that commutes (resp. anti commutes) with Q. By the action of P, we have the isomorphism $C \to A : R \mapsto PR$. Hence G can be partitioned into two sets C and A of the same size. Therefore:

$$\sum_{P\in G} (-1)^{P.Q} = \sum_{P\in C} (-1)^0 + \sum_{P\in A} (-1)^1 = |C| - |A| = 0.$$

We have $G \cap \mathbb{P}_n/G = \{\mathbb{I}\}$. Let $P \in \mathbb{P}_n$ and $Q \in \mathbb{P}_n/G$ the class of P. So, $P - Q \in G$ and P can be written as $P = (P - Q) + Q \in G + \mathbb{P}_n/G$. Therefore $\mathbb{P}_n = G \oplus \mathbb{P}_n/G$.

Finally since $C_G = G$, we have $C_{A_G} = C_{\mathbb{P}_n/C_G} = \mathbb{P}_n/C_{C_G} = \mathbb{P}_n/G$.

□

Lemma B.3. *Let $G \in \{G_1, \dots, G_{d+1}\}$. We have*

- *$\rho_G := \frac{1}{d}\sum_{P\in G} P$ is a rank one quantum state.*
- *$\mathcal{M}_G := \{M_G^Q := \frac{1}{d}\sum_{P\in G}(-1)^{P.Q}P\}_{Q\in A_G}$ is a POVM.*

Proof. Since G is an Abelian group we have for all $Q \in G$: $Q\rho_G Q = \frac{1}{d}\sum_{P\in G} QPQ = \frac{1}{d}\sum_{P\in G} P = \rho_G$ so ρ_G is stabilized by G. Moreover, since $|G| = d$:

$$\rho_G^2 = \left(\frac{1}{d}\sum_{P\in G} P\right)^2 = \frac{1}{d^2}\sum_{P,Q\in G} PQ = \frac{1}{d^2}\sum_{R\in G}\sum_{P\in G:PR\in G} R = \frac{1}{d}\sum_{R\in G} R = \rho_G$$

therefore ρ_G is a projector of trace $\mathrm{tr}(\rho_G) = \frac{1}{d}\cdot\mathrm{tr}(\mathbb{I}) = 1$ thus it is a rank 1 quantum state.

For the second point, we'll show that each term of $\mathcal{M}_G$ is a rank one projector and their sum is $\mathbb{I}$. Let $R \in A_G$, we have

$$\begin{aligned}(M_G^R)^2 &= \frac{1}{d^2}\sum_{P,Q\in G}(-1)^{(P+Q).R}PQ = \frac{1}{d^2}\sum_{P\in G}\sum_{Q\in G:QP\in G}(-1)^{P.R}P \\ &= \frac{1}{d}\sum_{P\in G}(-1)^{P.R}P = M_G^R\end{aligned}$$

and $\mathrm{tr}(M_G^R) = \frac{1}{d}(-1)^{R.I}\mathrm{tr}(\mathbb{I}) = 1$ so M_G^R is a rank 1 projector. Moreover, by using the fact that $C_{A_G} = \mathbb{P}_n/G$ and $G\cap\mathbb{P}_n/G = \{\mathbb{I}\}$, we obtain

$$\begin{aligned}\sum_{Q\in A_G} M_G^Q &= \sum_{Q\in A_G}\frac{1}{d}\sum_{P\in G}(-1)^{P.Q}P = \sum_{P\in G}\frac{1}{d}\sum_{Q\in A_G}(-1)^{P.Q}P \\ &= \sum_{P\in G}\mathbb{1}\{P\in C_{A_G}\}P = \sum_{P\in G\cap\mathbb{P}_n/G} P = \mathbb{I}.\end{aligned}$$

Finally, these two conditions imply that $\mathcal{M}_G$ is a POVM. □

Now, we can state a simplified version of the algorithm proposed by Flammia and Wallman (2020).

Algorithm 1 Learning a Pauli channel in diamond norm

Require: $N = \mathcal{O}(d^3\log(d)/\varepsilon^2)$ independent copies of the unknown Pauli channel $\mathcal{P}(\rho) = \sum_{P\in\mathbb{P}_n} p(P)P\rho P$.
Ensure: An approximated Pauli channel $\mathcal{R}$ such that $\|\mathcal{P}-\mathcal{R}\|_\diamond \le \varepsilon$.
  for $G \in \{G_1,\dots,G_{d+1}\}$ **do**
    Take the input $\rho_G = \frac{1}{d}\sum_{P\in G} P$, the output state is $\mathcal{P}(\rho_G)$.
    Perform $N_G = d^2\log(2d(d+1))/(4\varepsilon^2)$ measurements on $\mathcal{P}(\rho_G)$ using the POVM $\mathcal{M}_G := \{M_G^Q := \frac{1}{d}\sum_{P\in G}(-1)^{P.Q}P\}_{Q\in A_G}$ and observe $Q_1,\dots,Q_{N_G}\in A_G$.
    For $P\in G$, define $\hat{q}(P) = \frac{1}{N_G}\sum_{i=1}^{N_G}(-1)^{Q_i.P}$.
  end for
  Define for $P\in\mathbb{P}_n$, $q(P) = \frac{1}{d^2}\sum_{Q\in\mathbb{P}_n}(-1)^{Q.P}\hat{q}(Q)$.
  Let r be the orthogonal projection of q on the set of probability distributions on $\mathbb{P}_n$.
  return the Pauli channel $\mathcal{R}(\rho) = \sum_{P\in\mathbb{P}_n} r(P)P\rho P$.

Theorem B.4. *Algorithm 1 performs $\mathcal{O}\left(d^3\log(d)/\varepsilon^2\right)$ measurements to learn a Pauli channel to within ε in diamond norm with at least a probability $2/3$.*

The proof is taken from Flammia and Wallman (2020).

Proof. If we choose the input ρ_G for some stabilizer group G, apply the Pauli channel $\mathcal{P}$ and perform the measurement using the POVM $\mathcal{M}_G$, the induced probability distribution is given by:

$$p_G = \{\mathrm{tr}(M_G^Q\mathcal{P}(\rho_G))\}_{Q\in A_G} = \left\{\sum_{P\in G} p(P+Q)\right\}_{Q\in A_G},$$

because for $Q \in A_G$, we have:

$$\begin{aligned}
\operatorname{tr}(M_G^Q \mathcal{P}(\rho_G)) &= \frac{1}{d^2} \sum_{P_1,P_3 \in G, P_2} (-1)^{P_1.Q} p(P_2) \operatorname{tr}(P_1 P_2 P_3 P_2) \\
&= \frac{1}{d^2} \sum_{P_1,P_3 \in G, P_2} p(P_2)(-1)^{P_1.Q+P_2.P_3} \operatorname{tr}(P_1 P_3) \\
&= \frac{1}{d} \sum_{P_1 \in G, P_2} p(P_2)(-1)^{P_1.(Q+P_2)} \\
&= \sum_{P_2} p(P_2) \mathbb{1}(Q + P_2 \in C_G) \\
&= \sum_{P \in C_G} p(P+Q).
\end{aligned}$$

Therefore if $Q \sim p_G$ and $P \in G$, then:

$$\begin{aligned}
\mathbb{E}\left((-1)^{Q.P}\right) &= \sum_{Q \in A_G} p_G(Q)(-1)^{Q.P} = \sum_{Q \in A_G} \sum_{R \in C_G} p(R+Q)(-1)^{(Q+R).P} \\
&= \sum_{S \in \mathbb{P}_n} p(S)(-1)^{S.P} = \hat{p}(P)
\end{aligned}$$

because $P \in G$ thus P commutes with $R \in C_G$ and $\mathbb{P}_n = C_G \oplus A_G$. Therefore, by Hoeffding's inequality Hoeffding (1963), we can estimates $\{\hat{p}(P) = \mathbb{E}_{Q \sim p_G}(-1)^{Q.P}\}_{P \in G}$ to within ε/d with $N_G = \log(2d(d+1))/(2(\varepsilon/d)^2) = \mathcal{O}\left(d^2 \log(d)/\varepsilon^2\right)$ samples to have a probability of error at most $\delta/(d(d+1))$ for each $P \in G$. We repeat this procedure for each $G \in \{G_1, \ldots, G_{d+1}\}$ to estimate all $\{\hat{p}(P)\}_{P \in \mathbb{P}_n}$ to within ε/d. The total complexity is thus $N = \sum_{i=1}^{d+1} N_{G_i} = \mathcal{O}\left(d^3 \log(d)/\varepsilon^2\right)$. Let $\{\hat{q}(P)\}_{P \in \mathbb{P}_n}$ the approximations of $\{\hat{p}(P)\}_{P \in \mathbb{P}_n}$ given by the empirical means. We can define $\{q(P)\}_{P \in \mathbb{P}_n}$ as follows:

$$q(P) = \frac{1}{d^2} \sum_{Q \in \mathbb{P}_n} (-1)^{Q.P} \hat{q}(Q)$$

so that we have by the Parseval–Plancherel identity:

$$d\|q - p\|_2 = \|\hat{q} - \hat{p}\|_2 = \sqrt{\sum_{P \in \mathbb{P}_n} (\hat{q} - \hat{p})^2} \leq \sqrt{\sum_{P \in \mathbb{P}_n} (\varepsilon/d)^2} = \varepsilon.$$

However, q is not necessarily a probability distribution. The set of probability distributions on $\mathbb{P}_n$ is convex, hence if we define r as the orthogonal projection on this set, we have:

$$\|r - p\|_2 \leq \|q - p\|_2 \leq \varepsilon/d.$$

We conclude by the Cauchy-Schwartz inequality that r is a good approximation of p:

$$\|r - p\|_1 \leq d\|r - p\|_2 \leq \varepsilon.$$

□

C Proof of Ineq. (6)

In this Section, we give the proof of an approximation used in the proof of Theorem 5.1, more precisely in Ineq. (6), where we showed that the empirical average over the ensemble of a certain function is well-approximated by its mean.

Proposition C.1. *There is a universal constant $C > 0$ such that with probability at least $9/10$ we have:*

$$\begin{aligned}
&\sum_{k=1}^{N} \frac{1}{M} \sum_{x=1}^{M} \sum_{i_1,\ldots,i_{k-1}} \left(\prod_{t=1}^{k-1} \lambda_{i_t}^t \left(\frac{1 + u_{i_t}^{t,x}}{d} \right) \right) \sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,x})^2 \\
&\leq \sum_{k=1}^{N} \mathbb{E}_\alpha \left[\sum_{i_1,\ldots,i_{k-1}} \left(\prod_{t=1}^{k-1} \lambda_{i_t}^t \left(\frac{1 + u_{i_t}^{t,\alpha}}{d} \right) \right) \sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,\alpha})^2 \right] + N\varepsilon^2 \exp(-Cd^2).
\end{aligned}$$

Proof. Let $k \in [N]$. For $x \in [M]$, let $f_k(x)$ be the function:

$$f_k(x) = \sum_{i_1,\dots,i_{k-1}} \left(\prod_{t=1}^{k-1} \lambda_{i_t}^t \left(\frac{1 + u_{i_t}^{t,x}}{d} \right) \right) \sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,x})^2.$$

Similarly we define:

$$f_k(\alpha) = \sum_{i_1,\dots,i_{k-1}} \left(\prod_{t=1}^{k-1} \lambda_{i_t}^t \left(\frac{1 + u_{i_t}^{t,\alpha}}{d} \right) \right) \sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,\alpha})^2.$$

Since for all $k \in [N], x \in [M]$ and $i_k \in \mathcal{I}_k$, we have $(u_{i_k}^{k,x})^2 \leq 1$, $\sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,x})^2 \leq 16\varepsilon^2$ (Lemma 5.5) and $\sum_{i_k} \frac{\lambda_{i_k}^k u_{i_k}^{k,x}}{d} = 0$ thus $f_k(x) \in [0, 16\varepsilon^2]$. Therefore the function $\sum_{k=1}^N \frac{1}{M} \sum_{x=1}^M f_k(x)$ is $\left(\frac{32N\varepsilon^2}{M}, \dots, \frac{32N\varepsilon^2}{M}\right)$-bounded. Hence Hoeffding's inequality Hoeffding (1963) writes:

$$\mathbb{P}\left(\left| \sum_{k=1}^N \frac{1}{M} \sum_{x=1}^M f_k(x) - \mathbb{E}\left(\sum_{k=1}^N \frac{1}{M} \sum_{x=1}^M f_k(x) \right) \right| > s \right) \leq 2 \exp\left(- \frac{2s^2}{\sum_{j=1}^M \left(\frac{32N\varepsilon^2}{M}\right)^2} \right).$$

Since for all $x \in [M]$ we have $\mathbb{E}\left(f_k(x)\right) = \mathbb{E}_\alpha(f_k(\alpha))$, we deduce:

$$\mathbb{P}\left(\left| \sum_{k=1}^N \frac{1}{M} \sum_{x=1}^M f_k(x) - \sum_{k=1}^N \mathbb{E}_\alpha(f_k(\alpha)) \right| > s \right) \leq 2 \exp\left(- \frac{s^2 M}{512 N^2 \varepsilon^4} \right).$$

Finally, by taking $s = 25N\varepsilon^2 \sqrt{\frac{\log(20)}{M}}$, with probability at least $9/10$, we have:

$$\begin{aligned}
&\sum_{k=1}^N \frac{1}{M} \sum_{x=1}^M \sum_{i_1,\dots,i_{k-1}} \left(\prod_{t=1}^{k-1} \lambda_{i_t}^t \left(\frac{1 + u_{i_t}^{t,x}}{d} \right) \right) \sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,x})^2 \\
&= \sum_{k=1}^N \frac{1}{M} \sum_{x=1}^M f_k(x) \leq \sum_{k=1}^N \mathbb{E}_\alpha(f_k(\alpha)) + 25N\varepsilon^2 \sqrt{\frac{\log(20)}{M}} \\
&\leq \sum_{k=1}^N \mathbb{E}_\alpha \left[\sum_{i_1,\dots,i_{k-1}} \left(\prod_{t=1}^{k-1} \lambda_{i_t}^t \left(\frac{1 + u_{i_t}^{t,\alpha}}{d} \right) \right) \sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,\alpha})^2 \right] + N\varepsilon^2 \exp(-Cd^2)
\end{aligned}$$

where $C > 0$ is a universal constant and we used the fact that $M = \exp(\Omega(d^2))$. □