On Adaptivity in Classical and Quantum Learning

Aadil Oufkir


Numéro National de Thèse : 2023ENSL0037

THÈSE

en vue de l'obtention du grade de Docteur, délivré par
l'ÉCOLE NORMALE SUPÉRIEURE DE LYON

École Doctorale N°512
École Doctorale en Informatique et Mathématiques de Lyon

Discipline : Informatique

Soutenue publiquement le 14/09/2023, par :
Aadil OUFKIR

On Adaptivity in Classical and Quantum Learning

L'Adaptativité dans l'Apprentissage Classique et Quantique

Devant le jury composé de :

Richard Kueng, Professeur, Johannes Kepler University Linz	Rapporteur
Marco Tomamichel, Professeur, National University of Singapore	Rapporteur
Sébastien Bubeck, Professeur, Microsoft	Examinateur
Alice Guionnet, Directrice de recherche CNRS, ENS de Lyon	Examinatrice
Cécilia Lancien, Chargée de recherche CNRS, UGA	Examinatrice
Omar Fawzi, Directeur de recherche Inria, ENS de Lyon	Directeur
Aurélien Garivier, Professeur des universités, ENS de Lyon	Co-directeur

Contents

Abstract

Learning properties of data is a fundamental problem in statistics that has applications in almost every aspect of our lives. There are two natural classes of strategies for learning properties of data. Non-adaptive strategies collect all the required data and then processes all this data in one batch, while adaptive strategies can use the accumulated information to decide whether more samples should be collected. Furthermore, with adaptive methods the way that new information is gathered can also be adapted according to what was observed in previous steps. Since our goal is to minimize the required amount of data, in this thesis, we study the optimal number of resources for both approaches in various classical and quantum learning problems.

For testing classical distributions, we show that adaptive strategies outperform their non-adaptive counterparts by a factor of four when the alphabet size is small. In addition, adaptive strategies can stop earlier when the tested distributions are far apart from each other. This advantage holds even with larger alphabets.

Concerning testing quantum states, we exhibit situations where adaptive strategies have provable advantage against the non-adaptive ones. These include the (binary) hypothesis testing, testing identity and closeness of quantum states.

In certain situations however, we have multiple ways of querying the data. These situations can be modeled by quantum channels that output data after receiving an input chosen by the learning strategy. These inputs can also be adapted to the previous observations by adaptive strategies while non-adaptive ones should determine all inputs in advance. We determine the optimal complexity of testing whether a channel is perfect or not. Moreover, we characterize the optimal number of resources for learning a general channel or a Pauli noise in the non-adaptive setting. Furthermore, we reduce the gap between adaptive and non-adaptive strategies for learning Pauli noise.

Résumé

Apprendre à partir de données est un problème fondamental en statistique qui a des applications dans presque tous les aspects de notre vie. Il existe deux classes naturelles de stratégies pour apprendre à partir de données. Les stratégies non adaptatives collectent toutes les données nécessaires, puis traitent toutes ces données en une seule fois, tandis que les stratégies adaptatives peuvent utiliser les informations accumulées pour décider si davantage d'échantillons doivent être collectés. De plus, avec les méthodes adaptatives, la façon dont les nouvelles informations sont obtenues peut également être adaptée en fonction de ce qui a été observé lors des étapes précédentes. Étant donné que notre objectif est de minimiser la quantité de données requises, dans cette thèse, nous étudions le nombre optimal de ressources pour les deux approches pour différents problèmes d'apprentissage classiques et quantiques.

Pour tester les distributions classiques, nous montrons que les stratégies adaptatives surpassent leurs analogues non adaptatives d'un facteur de quatre lorsque la taille de l'alphabet est petite. De plus, les stratégies adaptatives peuvent s'arrêter plus tôt lorsque les distributions testées sont très différentes les unes des autres. Cet avantage est également présent même pour les grands alphabets.

En ce qui concerne le test des états quantiques, nous présentons des situations où les stratégies adaptatives ont un avantage significatif par rapport aux stratégies non adaptatives. Celles-ci comprennent le test d'hypothèses (binaires), le test d'identité et de proximité des états quantiques.

Cependant, dans certaines situations, nous avons plusieurs façons de générer les données. Ces situations peuvent être modélisées par des canaux quantiques qui produisent des données après avoir reçu une entrée choisie par la stratégie d'apprentissage. Ces entrées peuvent également être adaptées aux observations précédentes par les stratégies adaptatives, tandis que les stratégies non adaptatives doivent déterminer toutes les entrées à l'avance. Nous déterminons la complexité optimale pour tester si un canal est parfait ou non. De plus, nous caractérisons le nombre optimal de ressources pour apprendre un canal général ou un bruit de type Pauli dans le cadre non adaptatif. En outre, nous réduisons l'écart entre les stratégies adaptatives et non adaptatives pour l'apprentissage du bruit de type Pauli.

Notation

Common	
$[n]$	The set of integers between 1 and n: $[n] = \{1, \ldots, n\}$
log	Natural logarithm (base e)
$a \wedge b$	$\min\{a, b\}$
$a \vee b$	$\max\{a, b\}$
$\mathfrak{S}_n$	The group of permutations of $[n]$
$\mathcal{N}(\mu, \sigma^2)$	The Gaussian distribution of mean μ and variance σ^2 of probability density function $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
$\mathcal{N}(0, 1)$	The standard Gaussian distribution
$\|x\|_p$	The p-norm of the vector x: $\|x\|_p = \left(\sum_{i=1}^d x_i^p\right)^{1/p}$
$\mathbb{C}^{d \times d'}$	The set of complex $d \times d'$ matrices
$\mathbb{I}_d$	The identity matrix of $\mathbb{C}^{d \times d}$
$\mathrm{diag}(X)$	The diagonal matrix whose diagonal entries are the elements of X.
$\mathbb{U}(d)$	The group of unitary matrices $\{A \in \mathbb{C}^{d \times d} : AA^\dagger = \mathbb{I}\}$
$\mathbf{S}^d$	The set of unit vectors of $\mathbb{C}^d$
$A^\top$	The transpose of the matrix A
$A^\dagger$	The conjugate transpose (adjoint) of the matrix A
$\|A\|$	The absolute value of A: $\|A\| = \sqrt{A^\dagger A}$
$\|A\|_p$	The Schatten p-norm of the matrix A: $\|A\|_p = \left[\mathrm{Tr}\left(\|A\|^p\right)\right]^{1/p}$
$\|A\|_{\mathrm{Tr}}$	The trace norm of the matrix A: $\|A\|_{\mathrm{Tr}} = \frac{1}{2}\mathrm{Tr}(\|A\|)$
$A \succcurlyeq 0$	The matrix A is positive semi-definite
$A \succcurlyeq B$	The matrix $A - B$ is positive semi-definite
Bra-ket	
$\|\phi\rangle$	A column vector
$\langle\phi\|$	The row vector adjoint to $\|\phi\rangle$
$\{\|i\rangle\}_{i \in [d]}$	The canonical basis $\{\|i\rangle\}_{i \in [d]} := \{\|e_i\rangle\}_{i \in [d]}$ where e_i has the i-th entry equal to 1 and the other entries are equal to 0
$\|ij\rangle$	The tensor product of $\|i\rangle$ and $\|j\rangle$: $\|ij\rangle = \|i\rangle \otimes \|j\rangle$
$\langle\phi\|\psi\rangle$	The scalar product of the vectors $\|\psi\rangle$ and $\|\phi\rangle$
$\|\phi\rangle\langle\phi\|$	The projector on the space spanned by the unit vector $\|\phi\rangle$

Chapter 1

Introduction

1.1 Adaptivity in classical and quantum Learning

Learning properties of data is an essential part of statistics. It involves understanding the underlying patterns which can then be used to make predictions about future outcomes. This task is essential and ubiquitous in every aspect of human life [WW70; RRSRR73] including clinical trials [FFDRG15], economics [ASWCC16], finance [Rup04], machine learning [JDM00], fraud detection [BH02], etc. In this thesis, we will explore two different strategies for learning properties of data : non-adaptive/non-sequential and adaptive/sequential strategies. Our goal is to minimize the required amount of data required to perform the learning task for both types of strategies.

Non-adaptive or non-sequential strategies are those where all parameters of the learning process are chosen before it begins; these choices remain fixed throughout the whole procedure. This type of strategy relies heavily on prior knowledge and assumptions about how best to approach the learning task, as well as having access to the total batch of observations before starting the learning procedure.

In contrast, adaptive or sequential strategies [Wal45; BV94; Ünl04; LHSS06; GK19] are capable of adapting the choices of the learning procedure during its execution. This means that these models learn actively rather than passively like their non-adaptive counterparts do. Importantly, the choices determining how the newly acquired data will be generated can depend on the previously observed data.

By "learning properties of data" we mean every task of extracting classical information from random observations. In this thesis, we focus on two types of problems that fit in this category. The first type concerns learning fully the object we are dealing with. Of course, because of the inherent randomness, we cannot exactly recover this object without allowing any error. Still, we can require to construct a classical description that approximates this object in most cases. We refer to this type of problem as *tomography* or simply *learning*. The second type of problem we consider in this thesis is about *testing*. In this case, we do not ask to approximate completely the object we are dealing with, but only to test whether it satisfies some property or is far from satisfying it. This kind of problems is interesting because it requires less resources than its learning counterpart and in many cases knowing whether a data satisfies some property is all that is needed. Furthermore, from a theoretical point of view, the techniques we need for testing problems are in general different than those used for learning problems.

The main question of the thesis can be formulated as follows: **Can adaptive/sequential strategies outperform non-adaptive/non-sequential strategies for some testing**

or learning problem? We investigate the difference between adaptive/sequential and non-adaptive/non-sequential strategies for testing discrete classical distributions, quantum states and quantum channels.

1.2 Testing classical distributions

We start with classical discrete distributions, the simplest model for unknown data. Learning classical distributions is at the heart of classical machine learning [AB09; JM15]. We here focus on learning properties of unstructured discrete distributions, that is situations where we have no prior knowledge or information other than the number of different outcomes.

Definition 1.2.1. *Let $n \in \mathbb{N}^*$ be a positive integer. A discrete probability distribution $\mathcal{D}$ on $[n]$ is a set of n non negative reals $\{\mathcal{D}(i)\}_{i\in[n]}$ that sum to 1:*

- $\forall i \in [n] : \mathcal{D}(i) \geq 0$,
- $\sum_{i=1}^{n} \mathcal{D}(i) = 1$.

An important example of discrete distributions is the uniform distribution whose parts are all equal: $\forall i \in [n] : \mathcal{D}(i) = \frac{1}{n}$. We denote this distribution by Uniform($[n]$) or simply U_n whenever there is no confusion. A Bernoulli distribution is any probability distribution on a set of size 2. It is denoted by $\text{Bern}(p) := \{1-p, p\}$.
An event in the discrete case is any subset $S \subset [n]$. Under the distribution $\mathcal{D}$, the probability that the event S occurs is $\mathcal{D}(S) = \sum_{i\in S} \mathcal{D}(i)$.
A sample $x \sim \mathcal{D}$ from the distribution $\mathcal{D}$ is a random variable that takes the value i with probability $\mathcal{D}_i$. For independent and identically distributed samples, we use the shorthand "i.i.d.". In the modeling phase, we may have a candidate of a probability distribution $\mathcal{D}_0$ which represent our guess about this unknown distribution the data is sampled from. It is then natural to ask whether this source of randomness is exactly what we think, or not at all. This problem known as *testing identity* is formally defined as follows.

Definition 1.2.2. *Let $\varepsilon > 0$ be a threshold parameter, δ be a confidence parameter and let $\mathcal{D}_0$ be a fixed known probability distribution on the set $[n]$. Given τ i.i.d. samples from an unknown distribution $\mathcal{D}$, testing identity problem is about distinguishing between $\mathcal{D} = \mathcal{D}_0$ and $\text{dist}(\mathcal{D}, \mathcal{D}_0) \geq \varepsilon$ with at least a probability $1 - \delta$.*

Note that in this kind of tests, we have two requirements. If we define the null hypothesis as $\mathcal{H}_0 = \{\mathcal{D} = \mathcal{D}_0\}$ and the alternate hypothesis as $\mathcal{H}_1 = \{\mathcal{D} : \text{dist}(\mathcal{D}, \mathcal{D}_0) \geq \varepsilon\}$, two types of error are to be distinguished. The type I error occurs if the algorithm rejects the null hypothesis while it is the true one. On the other hand, the type II error happens when the algorithm rejects the alternate hypothesis while it is the true one. The type I and II errors represent the bad situations a tester would like to avoid. Then, the two requirements in the above definition can be reformulated as follows: both the type I and II errors should happen with a probability at most δ.
Thus far, we have not specified the distance "dist" we choose for this problem. To do so, let us suppose, only for now, that both distributions $\mathcal{D}_0$ and $\mathcal{D}$ are known. We are given a sample x from one of these distributions and would like to decide which distribution generates this sample. Since we have only one sample, any deterministic algorithm would choose a subset A_0 of $[n]$ so that it answer $\mathcal{D}_0$ if the sample $x \in A_0$ and $\mathcal{D}$ otherwise.

Suppose furthermore that $\mathcal{D}_0$ and $\mathcal{D}$ are equally likely to be the true distribution. In this case the success probability can be computed exactly:

$$\mathbb{P}\left(\mathcal{D}_0\right)\mathbb{P}_{x\sim\mathcal{D}_0}(x\in A_0)+\mathbb{P}\left(\mathcal{D}\right)\mathbb{P}_{x\sim\mathcal{D}}(x\notin A_0)=\frac{1}{2}\left(\mathcal{D}_0(A_0)+\mathcal{D}(A_0^c)\right). \tag{1.1}$$

Note that this probability only depends on the set A_0 and the two distributions $\mathcal{D}_0$ and $\mathcal{D}$. So, in order to minimize the error probability, we can take the set A_0 that maximizes the latter expression. It turns out that one such set is given by $A_0=\{i\in[n]:\mathcal{D}_0(i)>\mathcal{D}(i)\}$ which is natural: it is unlikely that the distribution generating x is $\mathcal{D}$ while we observe $x=i$ such that $\mathcal{D}_0(i)>\mathcal{D}(i)$. In this case the success probability can be written as follows

$$\frac{1}{2}\left(\mathcal{D}_0(A_0)+\mathcal{D}(A_0^c)\right)=\frac{1}{2}+\frac{1}{2}\sum_{i:\mathcal{D}_0(i)>\mathcal{D}_1(i)}\left(\mathcal{D}_0(i)-\mathcal{D}(i)\right)=\frac{1}{2}+\frac{1}{4}\sum_{i=1}^{n}\left|\mathcal{D}_0(i)-\mathcal{D}(i)\right| \tag{1.2}$$

where we use the fact that $\mathcal{D}_0$ and $\mathcal{D}$ are probability distributions in the last equality. This expression of the best success probability to discriminate between two distributions $\mathcal{D}_0$ and $\mathcal{D}$ motivates the total variation distance which we denote by TV and define formally:

Definition 1.2.3. *Let $\mathcal{D}_0$ and $\mathcal{D}$ be two probability distributions on $[n]$. The total variation distance between $\mathcal{D}_0$ and $\mathcal{D}$ is defined as*

$$\mathrm{TV}(\mathcal{D},\mathcal{D}_0)=\frac{1}{2}\sum_{i=1}^{n}\left|\mathcal{D}(i)-\mathcal{D}_0(i)\right|. \tag{1.3}$$

The total variation distance characterizes the minimal error probability to distinguish between the distributions $\mathcal{D}_0$ and $\mathcal{D}$ when considering only one sample. Moreover, it satisfies the Data-Processing inequality which can be seen in the following equivalent formulation.

Proposition 1.2.1. *Let $\mathcal{D}_0$ and $\mathcal{D}$ be two probability distributions on $[n]$. The total variation distance between $\mathcal{D}_0$ and $\mathcal{D}$ satisfies*

$$\mathrm{TV}(\mathcal{D},\mathcal{D}_0)=\max_{S\subset[n]}\left(\mathcal{D}(S)-\mathcal{D}_0(S)\right). \tag{1.4}$$

In particular, this can be used to prove lower bounds for testing two hypotheses problems (see Section 4.4). In the literature, this argument is called LeCam's method [LeC73]. Now, it is clear that the natural distance for testing identity problem is the total variation (TV). It remains to clarify what would be the difference between a sequential and non-sequential strategy for this task. Since a testing strategy has no role in the distribution that generates the i.i.d. samples, it can only choose whether it needs a new sample or not. As such, we can distinguish two settings of testing identity: *non-sequential* or *batch* setting in which the number of samples is fixed beforehand and *sequential* setting where the strategy has the ability to choose at each step whether a) to stop and answer the null or alternate hypothesis or b) to continue and ask for a new sample if it does not have a sufficient level of confidence to answer the correct hypothesis. In the latter setting, the stopping time can be random so we need to choose how to compare it with the deterministic number of samples in the batch setting. For this, it is natural to choose the expected stopping time.

If we want to see the exact difference between the optimal sample complexity for the testing identity problem in the batch and sequential settings, we could focus on the case of small alphabet where n is small, δ and ε goes to 0. For the simplicity of the presentation, we choose $n = 2$ and test Bernoulli distributions. In this case, we can characterize the exact sample complexity. Before stating this result, we need to introduce an important divergence that appears naturally in the complexity of both batch and sequential strategies. It is known as the Kullback Leibler (KL) divergence and will play a central role in this thesis.

Definition 1.2.4. *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two probability distributions on $[n]$. The Kullback Leibler divergence between $\mathcal{D}_1$ and $\mathcal{D}_2$ is defined as*

$$\mathrm{KL}(\mathcal{D}_1\|\mathcal{D}_2) = \sum_{i=1}^{n} \mathcal{D}_1(i) \log\left(\frac{\mathcal{D}_1(i)}{\mathcal{D}_2(i)}\right). \tag{1.5}$$

For two real numbers $p, q \in [0,1]$, we denote by $\mathrm{KL}(p\|q) = \mathrm{KL}(\mathrm{Bern}(p)\|\,\mathrm{Bern}(q)) = p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right)$.

This is only defined when $\mathcal{D}_1$ is absolutely continuous with respect to $\mathcal{D}_2$ where we use the convention "$0\log\left(\frac{0}{0}\right) = 0$". Otherwise, we can take $\mathrm{KL}(\mathcal{D}_1\|\mathcal{D}_2) = +\infty$. The KL divergence can be related to the TV distance by Pinsker's inequality.

Proposition 1.2.2 ([FHT03])**.** *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two distributions on $[n]$. We have*

$$\mathrm{KL}(\mathcal{D}_1\|\mathcal{D}_2) \geq 2\,\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2)^2. \tag{1.6}$$

One can wonder why the KL divergence appears in the sample complexity of testing identity problem. This is simply because of the Chernoff Hoeffding's inequality and its reverse. Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli random variables of parameter p. The law of large numbers says that the empirical mean $\frac{1}{n}\sum_{t=1}^{n} X_t$ converges to the theoretical mean $\mathbb{E}(X_1) = p$ almost surely when $n \to \infty$. To have a precise estimation of the number of samples sufficient to achieve a good precision we need to use a concentration inequality. In this case, we can apply the Chernoff-Hoeffding inequality.

Theorem 1.2.1 ([Hoe63])**.** *Let $n \in \mathbb{N}^*$ and $X_1, \ldots X_n \sim \mathrm{Bern}(p)$. We have for any $t > 0$:*

$$\mathbb{P}\left(\frac{\sum_{i=1}^{n}(X_i - \mathbb{E}(X_i))}{n} > t\right) \leq \exp(-n\,\mathrm{KL}(p+t\|p)). \tag{1.7}$$

Now we can state a simplified version of the main result of this part of the thesis. We start with the batch setting.

Theorem 1.2.2 (Batch setting)**.**

- *There is a non-sequential algorithm for testing identity to $\mathcal{D}_0 = \mathrm{Bern}(p)$ that uses a number of samples τ satisfying*

$$\tau = \max\left\{\frac{\log(2/\delta)}{\mathrm{KL}(p \pm \epsilon/2\|p)}, \frac{\log(2/\delta)}{\mathrm{KL}(p \pm \epsilon/2\|p \pm \epsilon)}\right\} \underset{\epsilon,\delta\to 0}{\sim} \frac{8p(1-p)}{\epsilon^2}\log(2/\delta)\,. \tag{1.8}$$

- *Any non-sequential algorithm for testing identity to $\mathcal{D}_0 = \mathrm{Bern}(p)$ needs to use a number of samples τ satisfying*

$$\liminf_{\delta\to 0} \frac{\tau}{\log(1/\delta)} \geq \max\left\{\frac{1}{\mathrm{KL}(p+\alpha\epsilon\|p)}, \frac{1}{\mathrm{KL}(p-\beta\epsilon\|p)}\right\} \underset{\epsilon\to 0}{\sim} \frac{8p(1-p)}{\epsilon^2}\,, \tag{1.9}$$

where $\alpha \in (0,1)$ and $\beta \in (0,1)$ are defined such that $\mathrm{KL}(p+\alpha\epsilon\|p) = \mathrm{KL}(p+\alpha\epsilon\|p+\epsilon)$ and $\mathrm{KL}(p-\beta\epsilon\|p) = \mathrm{KL}(p-\beta\epsilon\|p-\epsilon)$.

The detailed proof of this theorem can be found in Section 2.3.1. The upper bound is given by an application of the Chernoff-Hoeffding inequality. For the lower bound, we use instead Stirling's approximation. This result can be related to the fact that Chernoff-Hoeffding inequality is almost optimal, indeed we have the reverse Chernoff-Hoeffding inequality which also can be proved using Stirling's approximation.

Theorem 1.2.3. *Let $n \in \mathbb{N}^*$, $X_1, \dots X_n \sim \mathrm{Bern}(p)$. We have for any $t > 0$:*

$$\mathbb{P}\left(\frac{\sum_{i=1}^n (X_i - \mathbb{E}(X_i))}{n} > t\right) \geq \frac{1}{e\sqrt{2\pi n}} \exp(-n \,\mathrm{KL}(p+t\|p)). \tag{1.10}$$

The idea of the proof is simple, let $l = n(t+p)$ which we suppose is an integer for simplicity, we can express and lower bound the previous probability as follows:

$$\mathbb{P}\left(\frac{\sum_{i=1}^n (X_i - \mathbb{E}(X_i))}{n} > t\right) = \sum_{n \geq k > n(t+p)} \binom{n}{k} p^k (1-p)^{n-k} \geq \binom{n}{l} p^l (1-p)^{n-l}. \tag{1.11}$$

Then Stirling's approximation [Leu85] implies

$$\binom{n}{l} p^l (1-p)^{n-l} \geq \frac{1}{e\sqrt{2\pi n}} \exp(-n \,\mathrm{KL}(p+t\|p)). \tag{1.12}$$

After finding the optimal sample complexity of testing identity in the batch setting, we move to the sequential setting. Here, the number of samples τ is a (random) stopping time that can behave differently under the null and alternate hypotheses. Let q be the unknown parameter of the Bernoulli distribution $\mathcal{D}$. We state the main result in the sequential setting whose proof can be found in Section 2.3.2.

Theorem 1.2.4 (Sequential setting)**.** *There is a sequential algorithm for testing identity to $\mathcal{D}_0 = \mathrm{Bern}(p)$ whose stopping time τ satisfies*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}(\tau)}{\log(1/\delta)} \leq \frac{1}{\min\{\mathrm{KL}(p\|p \pm \epsilon)\}} \underset{\epsilon \to 0}{\sim} \frac{2p(1-p)}{\epsilon^2} \text{ if } q = p, \text{ and} \tag{1.13}$$

$$\limsup_{\delta \to 0} \frac{\mathbb{E}(\tau)}{\log(1/\delta)} \leq \frac{1}{\min\{\mathrm{KL}(p \pm |q-p|\|p)\}} \underset{|q-p| \to 0}{\sim} \frac{2p(1-p)}{|q-p|^2} \text{ if } |q-p| > \epsilon\,. \tag{1.14}$$

Moreover, any sequential algorithm for testing identity to $\mathcal{D}_0 = \mathrm{Bern}(p)$ has a stopping time τ satisfying

$$\mathbb{E}(\tau) \geq \frac{\log(1/3\delta)}{\min\{\mathrm{KL}(p\|p \pm \epsilon)\}} \underset{\epsilon \to 0}{\sim} \frac{2p(1-p)}{\epsilon^2} \log(1/3\delta) \text{ if } q = p, \text{ and} \tag{1.15}$$

$$\mathbb{E}(\tau) \geq \frac{\log(1/3\delta)}{\mathrm{KL}(q\|p)} \underset{|q-p| \to 0}{\sim} \frac{2p(1-p)}{|q-p|^2} \log(1/3\delta) \text{ if } |q-p| > \epsilon. \tag{1.16}$$

The sequential upper bound uses time uniform concentration inequalities. In our case, it is not difficult to deduce this type of inequalities using the Chernoff Hoeffding inequality and the union bound. Indeed, if $X_1, \cdots \sim \mathrm{Bern}(p)$ are i.i.d. random variables then we have by applying the union bound then the Chernoff Hoeffding inequality and Pinsker's inequality:

$$\mathbb{P}\left(\exists n \in \mathbb{N}^* : \frac{\sum_{i=1}^n (X_i - \mathbb{E}(X_i))}{n} > \sqrt{\frac{\log(2n(n+1)/\delta)}{2n}}\right) \tag{1.17}$$

$$\leq \sum_{n \geq 1} \exp\left(-2n\left(\sqrt{\frac{\log(2n(n+1)/\delta)}{2n}}\right)^2\right) = \sum_{n \geq 1} \frac{\delta}{2n(n+1)} = \frac{\delta}{2}. \tag{1.18}$$

This is only to illustrate the method but not used in the actual proof of this theorem. To obtain an expression involving the KL divergence as in the previous Theorem, we need to choose the time dependent thresholds carefully. On the other hand, this way of using the union bound is in general not optimal and leads to sub-optimal second terms in the complexity. A better technique is based on Ville's maximal inequality for non-negative super-martingales [HRMS20] which we use for general alphabets (see Lemma 2.6.2).
For the lower bound, we use again the Kullback Leibler divergence which is not a distance, but it is non negative and satisfies the tensorization property.

Proposition 1.2.3. *Let $\mathcal{D}_1, \mathcal{D}'_1, \mathcal{D}_2$ and $\mathcal{D}'_2$ distributions on $[n]$, we have*

- ***Non negativity*** $\mathrm{KL}(\mathcal{D}_1 \| \mathcal{D}_2) \geq 0$.
- ***Tensorization*** $\mathrm{KL}(\mathcal{D}_1 \times \mathcal{D}'_1 \| \mathcal{D}_2 \times \mathcal{D}'_2) = \mathrm{KL}(\mathcal{D}_1 \| \mathcal{D}_2) + \mathrm{KL}(\mathcal{D}'_1 \| \mathcal{D}'_2)$.

Furthermore, the Kullback Leibler divergence satisfies the Data-Processing property. For a distribution $\mathcal{D}$ on $[n]$ and a random variable $X : [n] \to \mathcal{X}$, we define the distribution of X under $\mathcal{D}$ as $\mathcal{D}^X = \{\mathcal{D}(X = x)\}_{x \in \mathcal{X}}$.

Proposition 1.2.4. *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two distributions on $[n]$. Let X be a random variable and g a function. Define the random variable $Y = g(X)$, we have*

$$\mathrm{KL}\left(\mathcal{D}_1^X \| \mathcal{D}_2^X\right) \geq \mathrm{KL}\left(\mathcal{D}_1^Y \| \mathcal{D}_2^Y\right). \tag{1.19}$$

The last ingredient of the sequential lower bound is Wald's lemma:

Lemma 1.2.1 ([Wal44])**.** *Let $X_1, \ldots$ be i.i.d random variables and $\tau \in \mathbb{N}$ be a stopping time for the sequence $(X_n)_n$. Suppose that τ and X_1 have finite expectations. We have*

$$\mathbb{E}(X_1 + \cdots + X_\tau) = \mathbb{E}(\tau)\mathbb{E}(X_1)\,. \tag{1.20}$$

Combining the tensorization property (Proposition 1.2.3), the data processing property (Proposition 1.2.4) of the KL divergence along with Wald's lemma in various scenarios permit to prove the sequential lower bound of Theorem 1.2.4. Let us illustrate this method for a stopping time under the null hypothesis. Let $q = p \pm \varepsilon$, $X_1, \ldots$ be random variables i.i.d. as $\mathrm{Bern}(p)$ and $Y_1, \ldots$ be random variables i.i.d. as $\mathrm{Bern}(q)$. Let τ be the stopping time of a testing strategy that answers with an error probability at most δ. On the one hand the tensorization property of the KL and Wald's Lemma imply:

$$\mathrm{KL}(\mathbb{P}_{X_1,\ldots,X_\tau} \| \mathbb{P}_{Y_1,\ldots,Y_\tau}) = \mathbb{E}\left(\tau\right) \mathrm{KL}(p \| q) \tag{1.21}$$

where $\mathbb{P}_X$ is the probability distribution of the random variable X. On the other hand, if $\mathcal{E}$ denotes the event that the testing algorithm $\mathcal{A}$ answers the null hypothesis, we have under the null hypothesis, $\mathbb{P}\left(\mathcal{E}\right) \geq 1 - \delta$ while under the alternate hypothesis we have $\mathbb{P}\left(\mathcal{E}\right) \leq \delta$. Hence, we can apply the data processing inequality on the KL divergence to obtain:

$$\mathrm{KL}(\mathbb{P}_{X_1,\ldots,X_\tau} \| \mathbb{P}_{Y_1,\ldots,Y_\tau}) \geq \mathrm{KL}(\mathcal{A}(X) \in \mathcal{E} \| \mathcal{A}(Y) \in \mathcal{E}) \geq \mathrm{KL}(1 - \delta \| \delta) \geq \log(1/3\delta). \tag{1.22}$$

Combining these (in)equalities, we obtain the lower bound on the expected stopping time under the null hypothesis:

$$\mathbb{E}\left(\tau\right) \geq \frac{\log(1/3\delta)}{\mathrm{KL}(p \| q)} = \frac{\log(1/3\delta)}{\mathrm{KL}(p \| p \pm \varepsilon)}. \tag{1.23}$$

So far, we have characterized the optimal sample complexity for testing identity to the Bernoulli distribution $\mathrm{Bern}(p)$ in both batch (Theorem 1.2.2) and sequential (Theorem 1.2.4) settings. In particular, when $\varepsilon, \delta \to 0$, the batch complexity is equivalent to $8p(1-p)\varepsilon^{-2}$ whereas the expected sequential complexity is equivalent to $2p(1-p)\varepsilon^{-2}$ under the null hypothesis and $2p(1-p)|q-p|^{-2}$ under the alternate hypothesis (recall $\mathcal{D} = \mathrm{Bern}(q)$). We remark that sequential strategies outperform batch ones by at least a factor 4. Furthermore, under the alternate hypothesis, sequential strategies adapt to the actual difficulty of the problem: the complexity depends on the distance $|q-p| = \mathrm{TV}(\mathcal{D}, \mathcal{D}_0)$ rather than the threshold parameter ε. This is the main message of Chapter 2 where we exhibit the same advantage for testing identity and closeness of distributions on small alphabets. For general alphabets, it is hard to find the optimal constant either in the batch or the sequential settings. Still, we show that sequential strategies can adapt to the actual distance between the tested distributions. Moreover, we prove a sequential lower bound that shows that (up to a constant) we cannot hope for more than this advantage in the worst case.
Since the advantage of sequential strategies over batch ones is understood for discrete distributions, one can wonder whether there is a learning problem for which this advantage is more significant. For instance, can we obtain a separation between sequential and non-sequential strategies at least polynomial in the parameters of some learning task? In the sequel, we will explore the advantage of sequentiality/adaptivity for more general models of learning properties of data.

1.3 Testing quantum states

One of the generalizations of discrete distributions is given by quantum states. We need to introduce some basics about quantum information theory before formulating learning or testing problems of quantum states. For this, we follow the excellent textbook [NC02].

1.3.1 Postulates of quantum mechanics

Quantum mechanics has four main postulates which connect the mathematical formalism with the real physical world.

1. **State space** The first postulate says that every quantum system is associated to a Hilbert space $\mathcal{H}$. A (pure) state $|\psi\rangle$ is any unit element of the Hilbert space $\mathcal{H}$. The quantum system is completely described by its state vector $|\psi\rangle \in \mathcal{H}$. In finite dimension, we can take the Hilbert space $\mathcal{H} = \mathbb{C}^d$. In particular, when $d = 2$, the quantum mechanical system is the qubit.

2. **Evolution** The second postulate states that the evolution of a closed quantum system is described by a unitary transformation. Concretely, the time evolution of the state $|\psi(t)\rangle \in \mathcal{H}$ is given by:

$$|\psi(t)\rangle = U(t)\,|\psi(0)\rangle \tag{1.24}$$

where $U(t)$ is a unitary operator. This is equivalent to saying that the state $|\psi(t)\rangle$ satisfies the Schrödinger equation:

$$\frac{d}{dt}\,|\psi(t)\rangle = -\frac{i}{\hbar} H\,|\psi(t)\rangle \tag{1.25}$$

where H represents the Hamiltonian, which is a Hermitian operator. The solution to the Schrödinger equation takes the form $|\psi(t)\rangle = U(t)\,|\psi(0)\rangle$ with $U(t) = \exp\left(-\frac{itH}{\hbar}\right)$ being a unitary operator.

3. **Quantum measurement** The third postulate is about describing the effects of measurements on quantum systems. A quantum measurement is a collection of $\{A_x\}_{x\in\mathcal{X}}$ of measurement operators satisfying $\sum_{x\in\mathcal{X}} A_x^\dagger A_x = \mathbb{I}$. The measurement outcome is one of the indices $x \in \mathcal{X}$. If the state $|\psi\rangle \in \mathcal{H}$ is measured with this quantum measurement then we observe x with probability $\langle\psi|\, A_x^\dagger A_x\,|\psi\rangle$ and the state of the system becomes

$$|\psi\rangle_{|x} = \frac{A_x\,|\psi\rangle}{\sqrt{\langle\psi|\, A_x^\dagger A_x\,|\psi\rangle}}. \tag{1.26}$$

4. **Composite systems** The fourth postulate says that state space of a composite physical system is the tensor product of the state spaces of the component physical systems. Concretely, if for $x \in \mathcal{X}$, $\mathcal{H}_x$ is the Hilbert space of the x-th quantum system then the Hilbert space of the combined quantum systems is $\otimes_{x\in\mathcal{X}}\mathcal{H}_x$. For instance a pure bipartite state $|\phi\rangle \in \mathcal{H}_1\otimes\mathcal{H}_2$ can be written as $|\phi\rangle = \sum_x \sqrt{\lambda_x}\,|r_x\rangle\otimes|s_x\rangle$ (Schmidt decomposition). It is called *separable* or *product* if it can be decomposed as $|\phi\rangle = |r\rangle\otimes|s\rangle$, otherwise it is called *entangled*. An example of an entangled state is given by $|\Psi\rangle = \frac{1}{\sqrt{2}}(|0\rangle\otimes|0\rangle + |1\rangle\otimes|1\rangle)$ where $\{|0\rangle, |1\rangle\}$ is the canonical basis of $\mathbb{C}^2$.

The first and third postulate are necessary to formulate every quantum testing or learning problem (Section 1.3.4, Section 1.4). The second postulate is only necessary when the object of the testing or learning procedure is the evolution of the system and not its state (Section 1.4). The fourth postulate permits to differentiate between two types of strategies of learning: coherent (or entangled) and incoherent strategies. Besides, it also allows us to consider situations when an algorithm can use auxiliary systems. We refer to Section 1.3.4 (resp. Section 1.4.3) for different models of learning properties of states (resp. evolutions).

1.3.2 Quantum states

Let $\mathcal{H} = \mathbb{C}^d$ be the Hilbert space associated to a quantum system. So far, we have mentioned only *pure* states, that is unit vectors $|\psi\rangle \in \mathcal{H}$. Now, we could be in a situation where we only have a probabilistic description of the state: with probability p_y, the state is $|\psi_y\rangle \in \mathcal{H}$. In this case the state is given by a pure state ensemble $\{p_y, |\psi_y\rangle\}_{y\in\mathcal{Y}}$. By the linearity of the measurement, this pure state ensemble would produce the same observations as $\rho = \sum_{y\in\mathcal{Y}} p_y\,|\psi_y\rangle\langle\psi_y|$. Indeed, for a POVM measurement $\{M_x\}_{x\in\mathcal{X}}$, the probability of observing $x \in \mathcal{X}$ can be written as

$$\sum_{y\in\mathcal{Y}} p_y\,\langle\psi_y|\,M_x\,|\psi_y\rangle = \mathrm{Tr}\left(M_x \sum_{y\in\mathcal{Y}} p_y\,|\psi_y\rangle\langle\psi_y|\right) = \mathrm{Tr}\left(M_x\rho\right). \tag{1.27}$$

Note that $\mathrm{Tr}(\rho) = \sum_{y\in\mathcal{Y}} p_y \mathrm{Tr}(|\psi_y\rangle\langle\psi_y|) = \sum_{y\in\mathcal{Y}} p_y = 1$ since p is a probability and for all $|\phi\rangle$ we have $\langle\phi|\,\rho\,|\phi\rangle = \sum_{y\in\mathcal{Y}} p_y |\langle\phi|\psi_y\rangle|^2 \geq 0$ so ρ is positive semi-definite. Any matrix satisfying these conditions is called a density matrix or simply a quantum state.

Definition 1.3.1. *A quantum state (density matrix) is a positive semi-definite matrix of trace* 1*.*

Conversely, by the spectral theorem, a quantum state can be written as $\rho = \sum_{y\in\mathcal{Y}} p_y |\psi_y\rangle\langle\psi_y|$ such that $p = \{p_y\}_{y\in\mathcal{Y}}$ is a probability. Now, with this definition, a pure state is any quantum state of rank 1 since for two unit vectors $|\phi\rangle$ and $|\psi\rangle$ we have $|\phi\rangle\langle\phi| = |\psi\rangle\langle\psi| \iff |\langle\phi|\psi\rangle|^2 = 1 \iff \exists\theta \in [0, 2\pi) : |\phi\rangle = e^{i\theta}|\psi\rangle$. Three particularly important examples of quantum states are:

1. **A classical state** is given by a diagonal density matrix $\rho = \mathrm{diag}(p)$ where $p = \{p_x\}_{x\in\mathcal{X}}$ is a probability distribution. We will see in Section 1.3.4 how testing classical states reduces to testing classical distributions.

2. **The maximally mixed state** is given by $\rho = \frac{\mathbb{I}}{d}$.

3. **The maximally entangled state** is given by $\rho = |\Psi\rangle\langle\Psi|$ where $|\Psi\rangle = \frac{1}{\sqrt{d}}\sum_{i=1}^{d} |i\rangle_A \otimes |i\rangle_B$ and $\{|i\rangle_A\}_{i=1}^d$ (resp. $\{|i\rangle_B\}_{i=1}^d$) is the canonical basis of $\mathcal{H}_A \cong \mathbb{C}^d$ (resp. $\mathcal{H}_B \cong \mathbb{C}^d$). For example when $d = 2$, the maximally entangled state is

$$\rho = |\Psi\rangle\langle\Psi| = \frac{1}{2}\begin{pmatrix}1\\0\\0\\1\end{pmatrix}\begin{pmatrix}1 & 0 & 0 & 1\end{pmatrix} = \begin{pmatrix}\frac{1}{2} & 0 & 0 & \frac{1}{2}\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ \frac{1}{2} & 0 & 0 & \frac{1}{2}\end{pmatrix}.$$

We can generalize the definition of an entangled state for non-pure states.

Definition 1.3.2. *Let ρ be a bipartite quantum state on $\mathcal{H}_1 \otimes \mathcal{H}_2$. It is called separable if it can be written as*

$$\rho = \sum_x p_x(\sigma_x \otimes \zeta_x) \tag{1.28}$$

for a probability distribution $p = \{p_x\}_x$ and quantum states $\{\sigma_x\}_x$ and $\{\zeta_x\}_x$.
A quantum state is called entangled if it is not separable.

For instance, the maximally mixed state on $\mathbb{C}^{d_1} \otimes \mathbb{C}^{d_2}$ is separable. On the other hand, the maximally entangled state $\rho = |\Psi\rangle\langle\Psi| = \frac{1}{d}(|i\rangle \otimes |i\rangle)(\langle j| \otimes \langle j|)$ on $\mathbb{C}^d \otimes \mathbb{C}^d$ is, as the name suggests, entangled (not separable).

1.3.3 POVM measurements

POVM (positive operator-valued measure) measurements provide an elegant mathematical tool to analyze quantum measurement in situations the post measurement states are not important and we are only interested in the measurement statistics.

Definition 1.3.3. *A POVM measurement is a set of positive semi-definite matrices $\mathcal{M} = \{M_x\}_{x\in\mathcal{X}}$ acting on the Hilbert space $\mathbb{C}^d$ and satisfying $\sum_{x\in\mathcal{X}} M_x = \mathbb{I}$. Each element M_x in the POVM $\mathcal{M}$ is associated with the outcome $x \in \mathcal{X}$.*

Starting from a quantum measurement $\{A_x\}_{x\in\mathcal{X}}$, we can construct a POVM measurement $\mathcal{M} = \{M_x = A_x^\dagger A_x\}_{x\in\mathcal{X}}$ since for all $x \in \mathcal{X}$, $M_x = A_x^\dagger A_x \succcurlyeq 0$ and

$\sum_{x\in\mathcal{X}} M_x = \sum_{x\in\mathcal{X}} A_x^\dagger A_x = \mathbb{I}$. As such, we can see how to perform a POVM measurement on a state $|\psi\rangle$: the probability that the measurement on a quantum state $|\psi\rangle$ using the POVM measurement $\mathcal{M}$ will output x is exactly $\langle\psi| A_x^\dagger A_x |\psi\rangle = \langle\psi| M_x |\psi\rangle$. An important POVM is given by a random basis. The rule is simple, when we do not have enough information about the eigenbasis of an unknown given state, we measure with a random basis. But, how to generate such a basis? A simple way to generate a random basis would be to choose a Haar(d)-distributed unitary matrix U (Haar(d) is the Haar probability measure over the compact group $\mathbb{U}(d)$ of unitary $d \times d$ matrices [Haa33]). To sample such a matrix $U \sim \text{Haar}(d)$, one could start with a random matrix M whose entries are i.i.d. standard complex Gaussian random variables and apply the Gram-Schmidt orthonormalization process to its columns [Mec19]. Then it is not difficult to check that for any unitary matrix $U \in \mathbb{U}(d)$, the set $\mathcal{M}_U = \{U |i\rangle\langle i| U^\dagger\}_{i\in[d]}$ is a POVM measurement. In this case, after measuring the state $|\psi\rangle$, the post measurement states have the form

$$|\psi\rangle_{|i} = \left(\frac{\langle i| U^\dagger |\phi\rangle}{|\langle i| U^\dagger |\phi\rangle|}\right) U |i\rangle \quad \text{for} \quad i = 1, \ldots, d \tag{1.29}$$

thus they are in general useless.

1.3.4 Testing quantum states

After this brief introduction, we can now move to test quantum states. Since a state generalizes a probability distribution, our first question would be to find the right way to measure the distance between two quantum states. We have seen previously that the TV distance determines the minimal error probability to discriminate two distributions. It turns out that the trace norm plays a similar role.

Theorem 1.3.1 (Holevo-Helstrom, [Hol73; Hel69])**.** *The minimal error probability to discriminate between two known states σ_1 and σ_2 is given by*

$$\frac{1}{2} + \frac{1}{2}\|\sigma_1 - \sigma_2\|_{\text{Tr}} = \frac{1}{2} + \frac{1}{2}\max_{0\preccurlyeq O\preccurlyeq\mathbb{I}} \text{Tr}((\sigma_1 - \sigma_2)O). \tag{1.30}$$

A maximizing observable $0 \preccurlyeq O \preccurlyeq \mathbb{I}$ in the last equality provides an optimal strategy to discriminate two states. Indeed, the condition $0 \preccurlyeq O \preccurlyeq \mathbb{I}$ permits to construct the measurement device $\mathcal{M}_O = \{\mathbb{I} - O, O\}$. Measuring $\rho \in \{\sigma_1, \sigma_2\}$ with the POVM $\mathcal{M}_O$ produces a sample $X \sim \text{Bern}(\text{Tr}(O\rho))$. Thus, in the setting where $\rho = \sigma_1$ and $\rho = \sigma_2$ occur with probability $1/2$ each, the error probability is:

$$\frac{1}{2}\text{Tr}(O\sigma_1) + \frac{1}{2}\text{Tr}((\mathbb{I} - O)\sigma_2) = \frac{1}{2} + \frac{1}{2}\text{Tr}((\sigma_1 - \sigma_2)O). \tag{1.31}$$

Such observable can be constructed easily once the spectral decomposition of $\sigma_1 - \sigma_2$ is known. Indeed, since the matrix $\sigma_1 - \sigma_2$ is Hermitian, by the spectral theorem, it can be written as:

$$\sigma_1 - \sigma_2 = \Pi^+ - \Pi^- \tag{1.32}$$

where Π^+ and Π^- are two orthogonal projectors onto the positive and negative eigenspace respectively. In this case we can choose $O = \Pi^+$ and the minimal error probability of Theorem 1.3.1 is exactly:

$$\frac{1}{2} + \frac{1}{2}\text{Tr}((\sigma_1 - \sigma_2)\Pi^+) = \frac{1}{2} + \frac{1}{2}\sum_i \lambda_i \tag{1.33}$$

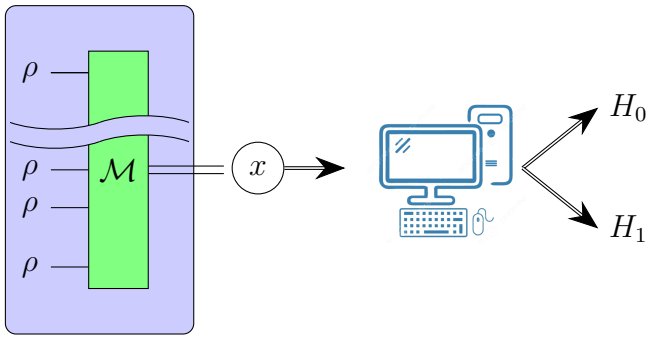

Figure 1.1: Illustration of an entangled strategy for testing quantum states. Here, x is the classical outcome of a measurement of the state $\rho^{\otimes N}$ with the POVM $\mathcal{M}$.

where $\{\lambda_i\}_i$ is the set of positive eigenvalues of $\sigma_1 - \sigma_2$. Note that $\sum_i \lambda_i = \frac{1}{2}\|\sigma_1 - \sigma_2\|_1 = \|\sigma_1 - \sigma_2\|_{\mathrm{Tr}}$ as $\mathrm{Tr}(\sigma_1 - \sigma_2) = 0$. Observe that the previous minimal error probability is in general strictly less than 1 unless the states σ_1 and σ_2 have disjoint supports. In particular, there are examples of states (σ_1, σ_2) for which the best error probability to discriminate them is very close to $1/2$. In other words, they are almost indistinguishable with only *one* measurement. In this case, repeating the test proves to be essential. With that we can give the formal definition of the binary hypothesis selection problem.

Definition 1.3.4. *Let $\delta \in (0, 1/2)$ be a confidence parameter and let σ_1 and σ_2 be two fixed and known d dimensional quantum states. Given N i.i.d. copies of an unknown quantum state $\rho \in \{\sigma_1, \sigma_2\}$, the binary hypothesis selection problem is to distinguish between $\rho = \sigma_1$ and $\rho = \sigma_2$ with at least a probability $1 - \delta$.*

This problem cannot be solved without specifying how we can use the N i.i.d. copies of ρ. For instance, one can think of a strategy where we put these copies all together in parallel and measure them at once. This type of strategies is called *entangled* and illustrated in Figure 1.1. Here $\mathcal{M}$ is a large d^N dimensional POVM. The entangled strategies are powerful and subsumes every possible strategy for testing states. However, in order to be able to use these strategies, we need to be able to keep the entanglement between the copies of ρ. This requires to have a quantum memory of large size: observe that the dimension of N systems grows exponentially with N. To circumvent these restrictions, we can think of simpler strategies without entanglement. In this case, one could think of a scenario where each copy of the state ρ is measured at each step. This enables, for instance, testing in situations we have not all the copies at once and we receive the copies at different times. In other words, we do not need to have a quantum memory to store all the copies before performing a global measurement. Here also we can distinguish between different strategies depending on the adaptiveness of the choice of the measurement devices on the previous observations. Concretely, an **incoherent strategy** is given by a sequence of POVMs $\{\mathcal{M}_t\}_{t\in[N]}$, each of them acts on the Hilbert space $\mathcal{H} = \mathbb{C}^d$. In this case, we measure at step t the quantum state ρ using the POVM $\mathcal{M}_t$. Depending on whether the measurement devices and the number of copies can adapt on the previous observations or not, we distinguish four types of incoherent strategies:

- **Non-adaptive non-sequential** is the setting where the number of copies N and the POVMs $\{\mathcal{M}_t\}_t$ are fixed in advance (i.e., do not depend on the outcomes of the previous measurements) (see Figure 1.2 for an illustration).

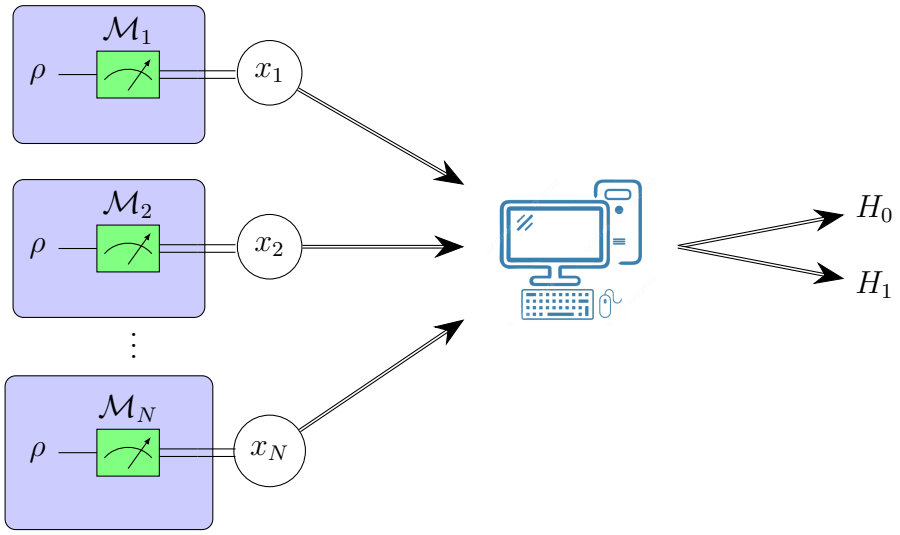

Figure 1.2: Illustration of a non-adaptive non-sequential incoherent strategy for testing quantum states. The classical computer processes the observations $(x_1, \ldots, x_N)$ to distinguish between two hypotheses H_0/H_1.

- **Non-adaptive sequential** is the setting where the POVMs $\mathcal{M}_t$ are fixed before hand but the the number of copies N can be chosen depending on the results of the previous measurements with the POVMs $\{\mathcal{M}_s\}_{s<t}$ (see Figure 1.3 for an illustration).
- **Adaptive non-sequential** is the setting where the number of copies N is fixed beforehand but the POVMs $\mathcal{M}_t$ can be chosen depending on the results of the previous measurements with the POVMs $\{\mathcal{M}_s\}_{s<t}$ (see Figure 1.4 for an illustration).
- **Adaptive sequential** is the setting where both the number of copies N and the POVMs $\mathcal{M}_t$ can be chosen depending on the results of the previous measurements with the POVMs $\{\mathcal{M}_s\}_{s<t}$ (see Figure 1.5 for an illustration).

For sequential strategies, the complexity is given by the *expected copy complexity* of the procedure $\mathbb{E}(N)$. Non-sequential strategies have a fixed number of measurements N.

It turns out that for incoherent strategies, the difference between sequential and non-sequential complexities is essentially a factor 4 as in the classical case (Section 1.2).

Theorem 1.3.2 (Informal)**.** *We can characterize the optimal copy complexity of binary hypothesis problem:*

- $\frac{2\log(1/\delta)}{\|\sigma_1-\sigma_2\|_{\mathrm{Tr}}^2}$ *copies of* ρ *are necessary and sufficient to distinguish between* $\rho=\sigma_1$ *and* $\rho=\sigma_2$ *with incoherent non-sequential strategies.*
- *In expectation,* $\frac{\log(1/\delta)}{2\|\sigma_1-\sigma_2\|_{\mathrm{Tr}}^2}$ *copies of* ρ *are necessary and sufficient to distinguish between* $\rho=\sigma_1$ *and* $\rho=\sigma_2$ *with incoherent sequential strategies.*

Actually, this result only holds asymptotically when $\delta \to 0$ and $\|\sigma_1-\sigma_2\|_{\mathrm{Tr}} \to 0$. Non-asymptotic complexities can be found with the same methods. Since the upper bounds use similar techniques of classical testing identity in Section 1.2, we prefer to focus on the lower bounds. Now, the testing algorithm can choose the measurement device at each

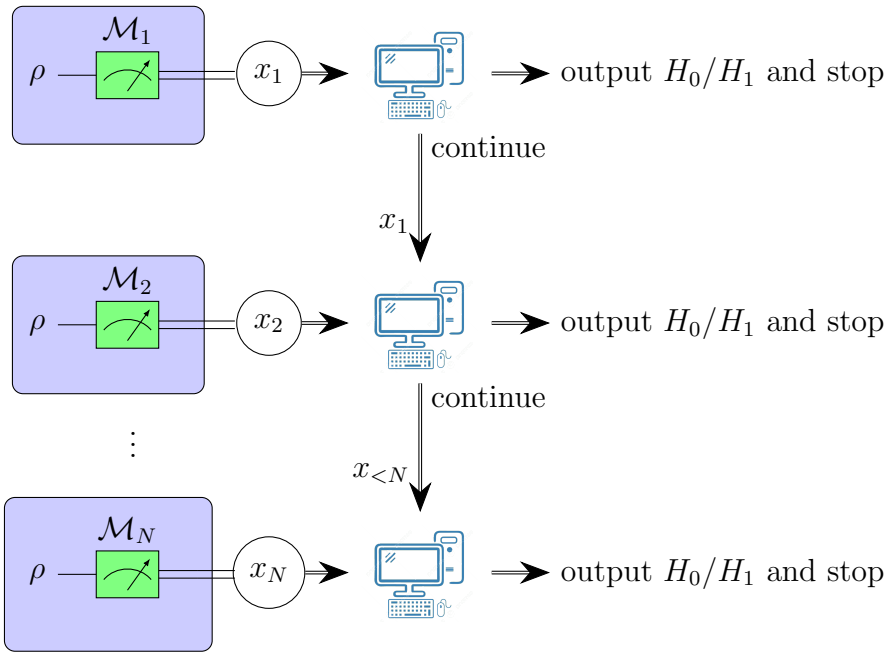

Figure 1.3: Illustration of a non-adaptive sequential incoherent strategy for testing quantum states. The classical computer processes the observations $(x_1, \ldots, x_N)$ to distinguish between two hypotheses H_0/H_1.

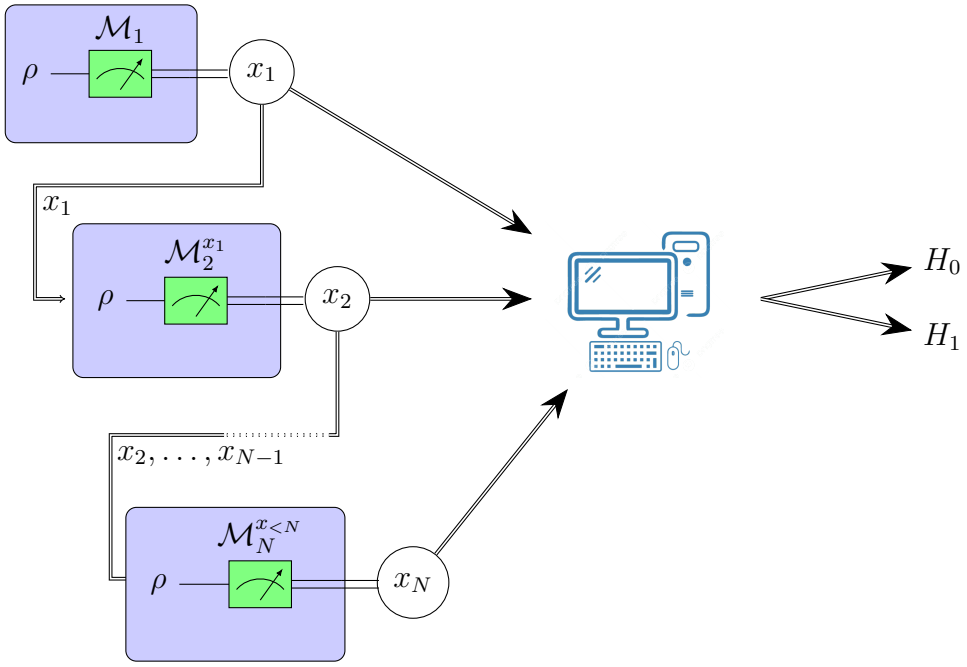

Figure 1.4: Illustration of an adaptive non-sequential incoherent strategy for testing quantum states. The classical computer processes the observations $(x_1, \ldots, x_N)$ to distinguish between two hypotheses H_0/H_1.

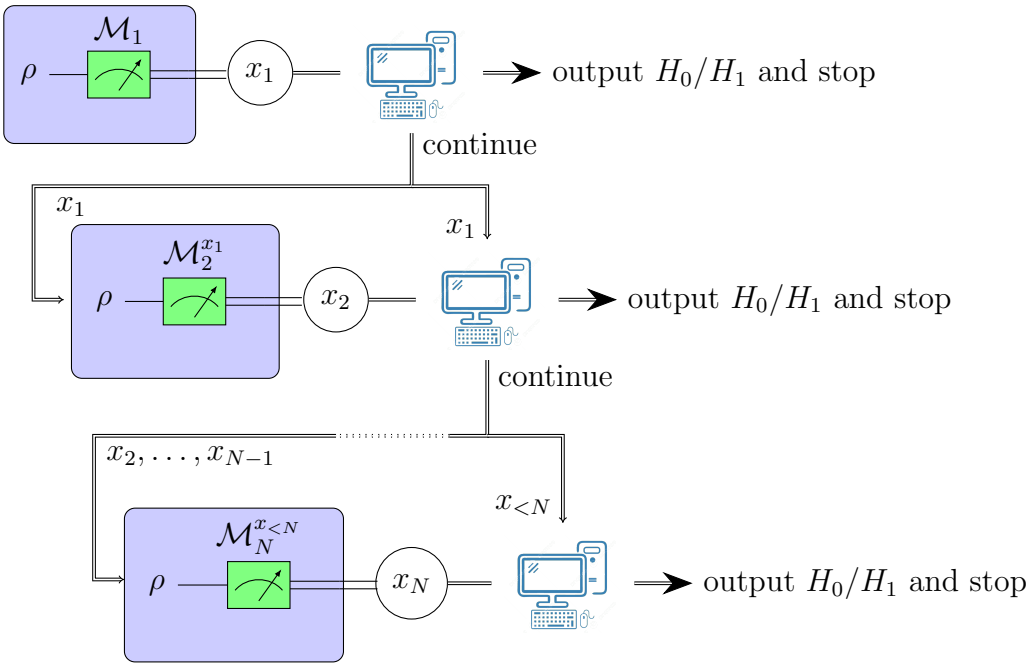

Figure 1.5: Illustration of an adaptive sequential incoherent strategy for testing quantum states. The classical computer processes the observations $(x_1, \ldots, x_N)$ to distinguish between two hypotheses H_0/H_1.

step so proving a lower bound against it is harder than the classical setting where the testing algorithm does not interact with the source of randomness. Still, we can choose a type of states for which the quantum testing algorithm cannot extract more information than a classical testing algorithm. The states we choose to prove the lower bounds are:

$$\sigma_1 = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \quad \text{and} \quad \sigma_2 = \begin{pmatrix} \frac{1}{2} - \varepsilon & 0 \\ 0 & \frac{1}{2} + \varepsilon \end{pmatrix}$$

where $\varepsilon = \|\sigma_1 - \sigma_2\|_{\mathrm{Tr}}$. With this choice of states, we show that every measurement the testing algorithm performs on σ_1, σ_2 can be seen as a post processing of a sample from $\mathrm{Bern}(1/2), \mathrm{Bern}(1/2 + \varepsilon)$ respectively. We can generalize this idea to any diagonal quantum state(s).

Lemma 1.3.1. *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two discrete distributions and σ_1 and σ_2 their corresponding diagonal quantum states. Let $\mathcal{M}$ be a POVM. Measuring the quantum state σ_1 (resp. σ_2) with the POVM $\mathcal{M}$ can be seen as post-processing (independent of the quantum states) of samples from the distribution $\mathcal{D}_1$ (resp. $\mathcal{D}_2$).*

This lemma permits to translate every lower bound for a classical testing problem to a lower bound for its quantum analogue. However, in some situations, this lower bound is not optimal. We refer to Section 3.3.1 for the proofs of these results.
So for the problem of discrimination two quantum states, the difference between sequential and non-sequential algorithms is the same as in the classical case. Moreover, adaptive algorithms cannot outperform non-adaptive ones since one of the optimal POVM is already known to the testing algorithm. It turns out that adaptive algorithms have the same performance as the non-adaptive ones for other problems as well, for instance testing identity [CHLL22] and state tomography [CHLLS22]. One can conjecture that this is

always the case. To show the contrary, we construct a (contrived) learning problem for which there is a polynomial separation between adaptive and non-adaptive strategies. We refer the reader to Section 3.4 for more details. We would like to finish with one of the main open problems about testing quantum states known as the composite hypothesis testing.

Definition 1.3.5 (Composite hypothesis testing)**.** *Given a known family* $\{\sigma_1, \ldots, \sigma_{2m}\}$ *of* $2m$ *quantum states* ε*-separated in the trace norm. The goal is to distinguish between the hypotheses*

$$H_0 : \rho \in \{\sigma_1, \ldots, \sigma_m\} \quad vs. \quad H_1 : \rho \in \{\sigma_{m+1}, \ldots, \sigma_{2m}\} \tag{1.34}$$

with high probability.

In this problem we can have up to $2m^2 - m$ optimal POVMs to choose. So a non-adaptive algorithm can be designed with a sample complexity $\mathcal{O}(m^2/\varepsilon^2)$. Moreover a lower bound of $\Omega(m/\varepsilon^2)$ can be proved for adaptive algorithms when $m \leq d$. We conjecture a separation between adaptive and non-adaptive algorithms for this problem.

1.4 Learning properties of quantum channels

After discussing how to learn properties of quantum states, we move to analyse how to learn properties of quantum channels. Let us first consider the classical analogue problem: learning properties of the classical channels.

1.4.1 Learning properties of classical channels

A classical channel is a stochastic matrix $W(y|x)_{x\in[m],y\in[d]}$. For an input $x \in [m]$, the output follows the probability distribution $y \sim \{W(y|x)\}_{y\in[d]}$. In other words, if we send an input x through the channel W, we receive the output y with a probability $W(y|x)$. For instance we have these examples of classical channels.

- The identity channel for which $m = d$ and $W(\cdot|x) = \delta_x$ for all $x \in [d]$.
- The noisy channel satisfying for all $x \in [m]$, $W(\cdot|x) \sim \text{Uniform}(d)$.

The identity channel is a perfect channel without any loss of information while the noisy channel is completely useless since the output is independent of the input. We can distinguish between these two types of channels using a few number of inputs (for $d \geq 2$). Still, we can imagine more interesting testing and learning tasks for classical channels. For instance, we can consider learning completely the channel or testing whether it is equal to a fixed channel or far from it. The problem of learning a channel can be formally defined as follows.

Definition 1.4.1 (Learning a classical channel)**.** *Given an unknown classical channel* W *as a black box. The goal is to construct a classical channel* $\tilde{W}(y|x)_{x\in[m],y\in[d]}$ *satisfying for all* $x \in [m]$*:*

$$\text{TV}(\tilde{W}(\cdot|x), W(\cdot|x)) \leq \varepsilon. \tag{1.35}$$

with an error probability at most $1/3$ *while minimizing the total number of channel uses* N*.*

It is not difficult to see that a number $\tilde{\Theta}(md/\varepsilon^2)$ of channel uses is necessary and sufficient for this task.
Moreover, testing whether an unknown channel W is exactly the identity channel or far from it can be formally defined as follows.

Definition 1.4.2 (Testing identity to the identity channel)**.** *Given an unknown channel W as a black box. The goal is to distinguish between the hypotheses*

$$H_0 : \forall x \in [d], W(\cdot|x) = \delta_x \quad \text{vs.} \quad H_1 : \exists x \in [d], \mathrm{TV}(W(\cdot|x), \delta_x) \geq \varepsilon \tag{1.36}$$

with an error probability at most $1/3$ while minimizing the total number of channel uses N.

The condition of the alternate hypothesis is equivalent to the existence of $x \in [d]$ such that $W(x|x) \leq 1 - \varepsilon$. In this case, a simple strategy would be to choose each input $x \in [d]$ a sufficient number of times then answer H_0 if always receiving the same output as input, otherwise answer H_1. If the channel is identity, this test makes no error. On the other hand, if the channel is not identity, then there is at least one input x such that the output is not always x. Since $\mathrm{TV}(W(\cdot|x), \delta_x) \geq \varepsilon$, the probability of seeing a different output than x is at least ε so the error probability after n repetitions is at most $(1-\varepsilon)^n$ thus we can choose $n = \log(1/3)/\log(1-\varepsilon) = \mathcal{O}(1/\varepsilon)$ to reduce the error probability to $1/3$. Hence, the total number of uses of the channel is $N = dn = \mathcal{O}(d/\varepsilon)$ because we do not know x and we need to test all the inputs to be sure we pass on x. Moreover, it is not difficult to see that a number of copies of the channel satisfying $N = \Omega(d/\varepsilon)$ is also necessary for any correct test. As we shall see in the sequel, generalizing this learning and testing results is not straightforward in the quantum setting.
Before we consider the quantum analogue of these learning/testing tasks, we introduce the definition of quantum channels and illustrate them with some examples. Then we compare different access models for learning properties of quantum channels.

1.4.2 Quantum channels

We have seen in the second postulate (Section 1.3.1) that the time evolution of a *closed* quantum system is described by a unitary transformation. However, the operator is no longer unitary when the quantum system is *open*, that is when it interacts with an environment. The overall evolution of the system and the environment is driven by Schrödinger's equation. If we want to focus only on the open quantum system (without the environment), we may ignore (trace out) the environment and analyse the new evolution. It can be proved that any such transformation can be described by a quantum channel (see e.g., [Lid19]).

Definition 1.4.3 (Kraus decomposition of quantum channels)**.** *A $(d_{\text{in}}, d_{\text{out}})$-dimensional quantum channel (process) is a map $\mathcal{N} : \mathbb{C}^{d_{\text{in}} \times d_{\text{in}}} \to \mathbb{C}^{d_{\text{out}} \times d_{\text{out}}}$ of the form*

$$\mathcal{N}(\rho) = \sum_k A_k \rho A_k^\dagger$$

where the Kraus operators $\{A_k\}_k$ satisfy $\sum_k A_k^\dagger A_k = \mathbb{I}$.

There are many important examples of quantum channels, including:

1. **The identity channel** $\mathrm{id}_d(\rho) = \rho$ admits the Kraus operator $\{\mathbb{I}_d\}$.
2. **Unitary channels:** given a unitary matrix U, the corresponding unitary channel $\mathcal{N}_U(\rho) = U\rho U^\dagger$ admits the Kraus operator $\{U\}$.
3. **The completely depolarizing channel** $\mathcal{D}(\rho) = \mathrm{Tr}(\rho)\frac{\mathbb{I}}{d_{\mathrm{out}}}$ admits the Kraus operators $\left\{\frac{1}{\sqrt{d_{\mathrm{out}}}}|i\rangle\langle j|\right\}_{j\in[d_{\mathrm{in}}],i\in[d_{\mathrm{out}}]}$.
4. **Measurement channels (instruments):** given a quantum measurement $\mathcal{M} = \{A_x\}_{x\in\mathcal{X}}$, after measuring a quantum state ρ with $\mathcal{M}$, we see the outcome $x \in \mathcal{X}$ with a probability $p_x = \mathrm{Tr}(A_x\rho A_x^\dagger)$ and the post-measurement state is $\rho_{|x} = \frac{A_x\rho A_x^\dagger}{\mathrm{Tr}(A_x\rho A_x^\dagger)}$. The corresponding quantum channel, denoted by $\mathcal{N}_\mathcal{M}$, can be defined as $\mathcal{N}_\mathcal{M}(\rho) = \sum_{x\in\mathcal{X}} p_x\rho_{|x} = \sum_{x\in\mathcal{X}} A_x\rho A_x^\dagger$, which gives the state of the system after the measurement has been performed. Therefore, the Kraus operators for the channel $\mathcal{N}_\mathcal{M}$ are simply given by the measurement operators $\{A_x\}_{x\in\mathcal{X}}$.

Since a convex combination of two quantum channels remains a quantum channel, we can construct more examples using the previous ones.
Note that the condition that the map $\mathcal{N}$ can be written as $\mathcal{N}(\rho) = \sum_k A_k\rho A_k^\dagger$ is equivalent to be completely positive (CP): for all d and $\rho \succcurlyeq 0$, $\mathrm{id}_d \otimes \mathcal{N}(\rho) \succcurlyeq 0$. The second condition $\sum_k A_k^\dagger A_k = \mathbb{I}$ is equivalent for the map $\mathcal{N}$ to be trace preserving (TP): for all ρ, $\mathrm{Tr}(\mathcal{N}(\rho)) = \mathrm{Tr}(\rho)$. Hence a map $\mathcal{N}$ is a quantum channel if, and only if, it is completely positive and trace preserving (CPTP).
It turns out that we can see a quantum channel as a quantum state with an additional property. Before stating this connection, we define the partial trace.

Definition 1.4.4 (Partial trace)**.** *Let $M = \sum_k A_k \otimes B_k \in \mathbb{C}^{d\times d} \otimes \mathbb{C}^{d'\times d'}$. We define the partial trace of the matrix M with respect to the first system as*

$$\mathrm{Tr}_1(M) = \sum_k \mathrm{Tr}(A_k)B_k \in \mathbb{C}^{d'\times d'}.$$

Similarly, we define the partial trace of the matrix M with respect to the second system as

$$\mathrm{Tr}_2(M) = \sum_k \mathrm{Tr}(B_k)A_k \in \mathbb{C}^{d\times d}.$$

The Choi state of a quantum channel $\mathcal{N}$ is a useful tool in quantum information theory that captures all the information about the channel. By linearity, once we know the set of images $\{\mathcal{N}(|i\rangle\langle j|)\}_{i,j=1}^{d_{\mathrm{in}}}$, we can compute exactly the output state $\mathcal{N}(\rho)$ for any given state ρ. Therefore, we can define the Choi state as the bipartite state $\mathcal{J}_\mathcal{N}$, which is obtained by applying the channel $\mathcal{N}$ to one part of a maximally entangled state, and leaving the other part untouched.

Definition 1.4.5 (Choi state)**.** *Let $\mathcal{N}$ be a $(d_{\mathrm{in}}, d_{\mathrm{out}})$ dimensional quantum channel. We define the Choi state $\mathcal{J}_\mathcal{N}$ of the channel $\mathcal{N}$ as*

$$\mathcal{J}_\mathcal{N} = \mathrm{id}\otimes\mathcal{N}(|\Psi\rangle\langle\Psi|) = \frac{1}{d_{\mathrm{in}}}\sum_{i,j=1}^{d_{\mathrm{in}}} |i\rangle\langle j| \otimes \mathcal{N}(|i\rangle\langle j|)$$

where $|\Psi\rangle = \frac{1}{\sqrt{d_{\mathrm{in}}}}\sum_{i=1}^{d_{\mathrm{in}}} |i\rangle\otimes|i\rangle$ is the maximally entangled state.

The map $\mathcal{J} : \mathcal{N} \mapsto \mathrm{id} \otimes \mathcal{N}(|\Psi\rangle\langle\Psi|)$ is an isomorphism called the Choi–Jamiołkowski isomorphism [Cho75; Jam72]. Note that for any quantum channel $\mathcal{N}$, $\mathcal{J}_{\mathcal{N}}$ is positive semi-definite and satisfies $\mathrm{Tr}_2(\mathcal{J}_{\mathcal{N}}) = \frac{\mathbb{I}}{d_{\mathrm{in}}}$. Moreover, any $\mathcal{K}$ satisfying these conditions is a Choi state: we can construct a quantum channel $\mathcal{N}$ such that $\mathcal{K} = \mathcal{J}_{\mathcal{N}}$.
We can compute the Choi states of the previous examples:

1. **The identity channel** $\mathrm{id}_d(\rho) = \rho$ has a Choi state $\mathcal{J}_{\mathrm{id}} = |\Psi\rangle\langle\Psi|$.
2. **A unitary channel** $\mathcal{N}_U(\rho) = U\rho U^\dagger$ has a Choi state $\mathcal{J}_U = (\mathbb{I} \otimes U)\,|\Psi\rangle\langle\Psi|\,(\mathbb{I} \otimes U^\dagger)$.
3. **The depolarizing channel** $\mathcal{D}(\rho) = \mathrm{Tr}(\rho)\frac{\mathbb{I}}{d_{\mathrm{out}}}$ has a Choi state $\mathcal{J}_{\mathcal{D}} = \frac{\mathbb{I}_{d_{\mathrm{in}}} \otimes \mathbb{I}_{d_{\mathrm{out}}}}{d_{\mathrm{in}} d_{\mathrm{out}}}$.

We have not specified so far how a learning algorithm would interact with an unknown channel. In the following, we describe various settings we can consider for learning properties of quantum channels.

1.4.3 Different models of learning properties of channels

In any natural scenario we could imagine, the learning algorithm would have the ability to choose the input state, send it through the channel and finally measure it. The classical observations the algorithm obtains after the measurement(s) are then processed by a classical computer to either a) distinguish between two hypotheses H_0/H_1 (testing problem) or b) return a classical approximation $\tilde{\mathcal{N}}$ of $\mathcal{N}$ (learning problem). The complexity is always measured by the total number of the channel uses and the number of measurements.

Coherent strategies.

The most general model of learning properties of a channel is given by quantum circuits. In this setting, we allow using auxiliary systems of arbitrary dimensions, the unknown channel can be applied sequentially before performing a measurement and finally multiple copies of the channel can be used in parallel so that the input state and the output state can be entangled.

Definition 1.4.6 (Coherent strategy)**.** *Let $n, N \in \mathbb{N}^*$ be positive integers, n (resp. N) represents the width (resp. depth) of the circuit. The input state is 2^n-dimensional and equal to $|0\rangle\langle 0|^{\otimes n}$. The general coherent strategy would choose*

1. *a $d_{\mathrm{in}}^n \times 2^n$-dimensional isometry matrix V_0,*
2. *$(N-1)$ $d_{\mathrm{out}}^n \times d_{\mathrm{in}}^n$-dimensional isometry matrices $V_1, \ldots, V_{N-1}$ and*
3. *a $2^n \times d_{\mathrm{out}}^n$-dimensional isometry matrix V.*

For each layer $1 \leq l \leq N-1$, the learner chooses a vector $x_l \in \{0,1\}^n$ whose $i-$th entry determines whether the channel $\mathcal{N}$ is applied on the $i-$th system or not. The state before measurement is given by:

$$\rho_{output} = \mathcal{N}_V \circ \mathcal{N}^{\otimes x_l} \circ \cdots \circ \mathcal{N}_{V_1} \circ \mathcal{N}^{\otimes x_1} \circ \mathcal{N}_{V_0}(|0\rangle\langle 0|^{\otimes n}) \tag{1.37}$$

*where $\mathcal{N}_V(\rho) = V\rho V^\dagger$. The output state is then measured with the canonical basis $\{|i\rangle\langle i|\}_{i\in[2^n]}$. The learner thus sees the outcome $i \in [2^n]$ with a probability $\langle i|\,\rho_{output}\,|i\rangle$. See Figure 1.6 for an illustration of this model. Shallow circuits where the depth is $N = 1$ are called **coherent batch** and are illustrated in Figure 1.7.*

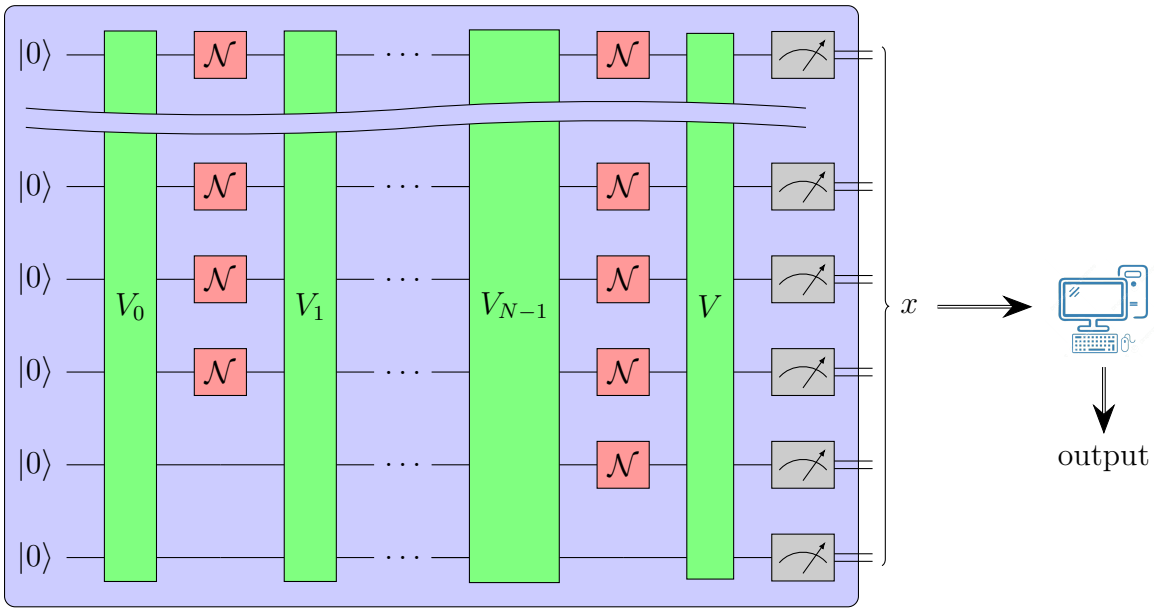

Figure 1.6: Illustration of a general strategy for learning properties of an unknown channel $\mathcal{N}$. The classical computer processes the observation x to distinguish between two hypotheses H_0/H_1 or produce an approximate channel $\tilde{\mathcal{N}}$.

Observe that a coherent strategy can also be represented in a sequential circuit where at each layer only one channel is applied (see Figure 1.8 for an illustration). This model permits to apply and combine several powerful quantum subroutines such as Grover's search algorithm [Gro96], quantum phase estimation [DDDSLWBW09] and quantum singular value transformation [GSLW19], among others. However, it presents practical challenges because, in this case, the learning algorithm must maintain entanglement throughout the circuit. Additionally, the number of required copies of quantum channels is generally exponential in the number of qubits. In this thesis, we will not consider such strategies and we focus only on incoherent strategies. We can imagine a variety of settings depending on whether an auxiliary system is allowed or not and whether the input/measurements can be chosen adaptively or not.

Incoherent strategies.

In the case we do not have a quantum memory, we have to measure the output of the unknown quantum channel immediately. So an incoherent strategy cannot use entangled input states or entangled measurement devices. Moreover, incoherent strategies are not allowed to apply the channel successively before performing a measurement. Depending on whether the ancilla is allowed or not and whether the input and measurement devices can depend on the previous observations or not, we distinguish four types of incoherent strategies. When auxiliary systems are not allowed, the strategy is called **ancilla-free**.

Definition 1.4.7 (Ancilla-free independent/incoherent strategy)**.** *At each step $1 \leq t \leq N$, the strategy would choose an input d_{in}-dimensional state ρ_t and a d_{out}-dimensional measurement device $\mathcal{M}_t = \{M_x^t\}_{x \in \mathcal{X}_t}$. It thus sees the outcome $x_t \in \mathcal{X}_t$ with a probability $\text{Tr}(\mathcal{N}(\rho_t)M_{x_t}^t)$. If the choice of the state ρ_t and measurement device $\mathcal{M}_t$ can depend on the previous observations $(x_1, \ldots, x_{t-1})$ the strategy is called **adaptive** (see Figure 1.10 for an illustration), otherwise it is called **non-adaptive** (see Figure 1.9 for an illustration).*

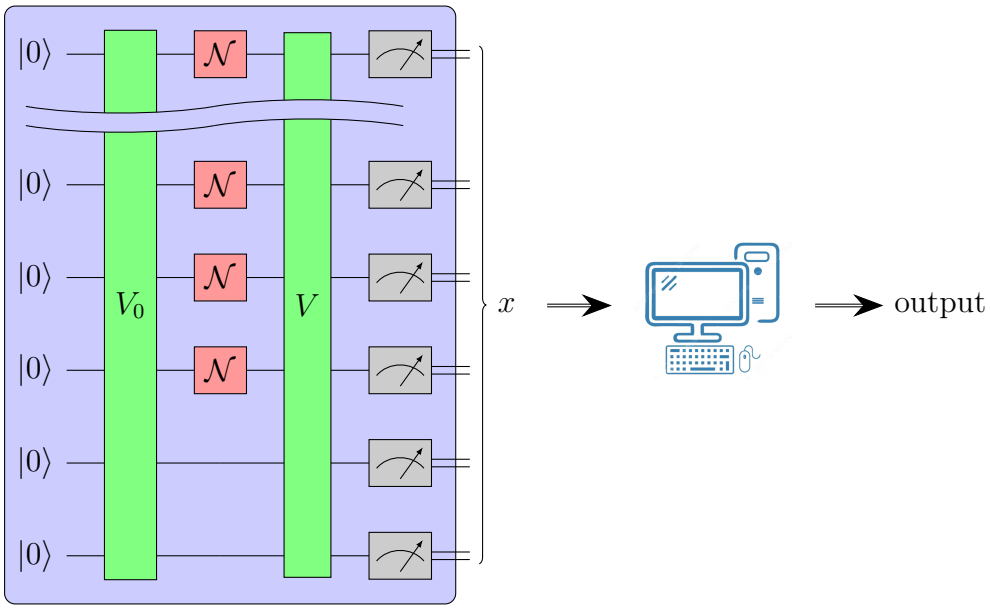

Figure 1.7: Illustration of a batch strategy for learning properties of an unknown channel $\mathcal{N}$. The classical computer processes the observation x to distinguish between two hypotheses H_0/H_1 or produce an approximate channel $\tilde{\mathcal{N}}$.

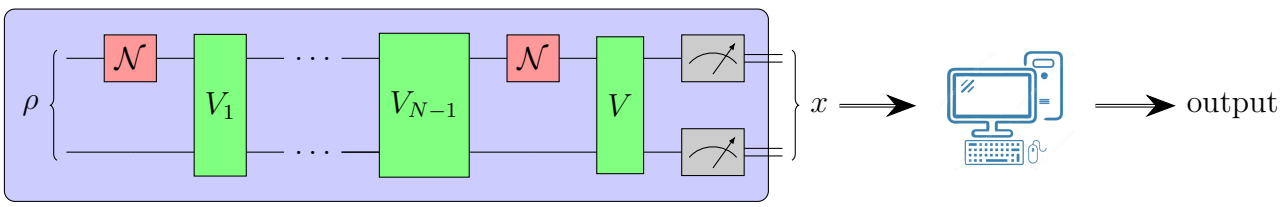

Figure 1.8: Illustration of a general coherent strategy for learning properties of an unknown channel $\mathcal{N}$. The classical computer processes the observation x and distinguish between two hypotheses H_0/H_1 or produce an approximate channel $\tilde{\mathcal{N}}$.

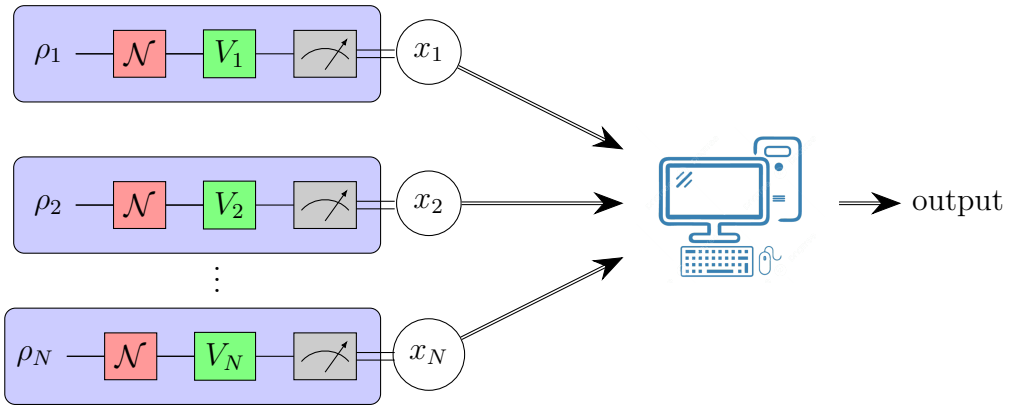

Figure 1.9: Illustration of an ancilla-free incoherent non-adaptive strategy for learning properties of quantum channels. The classical computer processes the observations $(x_1, \ldots, x_N)$ to distinguish between two hypotheses H_0/H_1 or produce an approximate channel $\tilde{\mathcal{N}}$.

Note that adaptive strategies are at least as powerful as their non-adaptive counterparts. On the other hand, if an auxiliary system is allowed to be used, these strategies are called **ancilla-assisted**.

Definition 1.4.8 (Ancilla-assisted independent/incoherent strategy)**.** *At each step t, the learner would choose an input $d_{\text{anc}}d_{\text{in}}$-dimensional state ρ_t and a $d_{\text{anc}}d_{\text{out}}$-dimensional measurement device $\mathcal{M}_t = \{M_x^t\}_{x\in\mathcal{X}_t}$. It thus sees the outcome $x_t \in \mathcal{X}_t$ with a probability* $\text{Tr}(\text{id}_{d_{\text{anc}}} \otimes \mathcal{N}(\rho_t)M_{x_t}^t)$. *If the choice of the state ρ_t and measurement device $\mathcal{M}_t$ can depend on the previous observations $(x_1, \ldots, x_{t-1})$ the strategy is called **adaptive** (see Figure 1.12 for an illustration), otherwise it is called **non-adaptive** (see Figure 1.11 for an illustration).*

Relation between different models of learning properties of channels

Here, we give the obvious reductions between the models we have seen previously and discuss about the interesting separations we would like to know in general.
First of all, every strategy can be cast in the general coherent model (see Definition 1.4.6). Moreover, each model where the use of auxiliary systems is allowed contains its ancilla-free counterpart as a special case. Also, non-adaptive incoherent strategies can be turned to coherent batch ones. But it is not clear how adaptive incoherent strategies perform compared to coherent batch ones. We known that there are examples of problems for which batch algorithms outperform adaptive ones. This includes the state tomography [HHJWY16; CHLLS22] and shadow tomography [Aar19; CCHL22] among others. However, finding an example for which adaptive strategies have a provable advantage over coherent batch ones remains an open question. We refer to Figure 1.13 for an illustration of these relations.

Now we see that different settings can be chosen for learning properties of quantum channels. We shall focus on incoherent strategies. We start with the problem of learning quantum channels then we analyse the same task for a restricted class of channels and finally we move to study the problem of testing whether a channel is equal to a fixed channel or far from it.

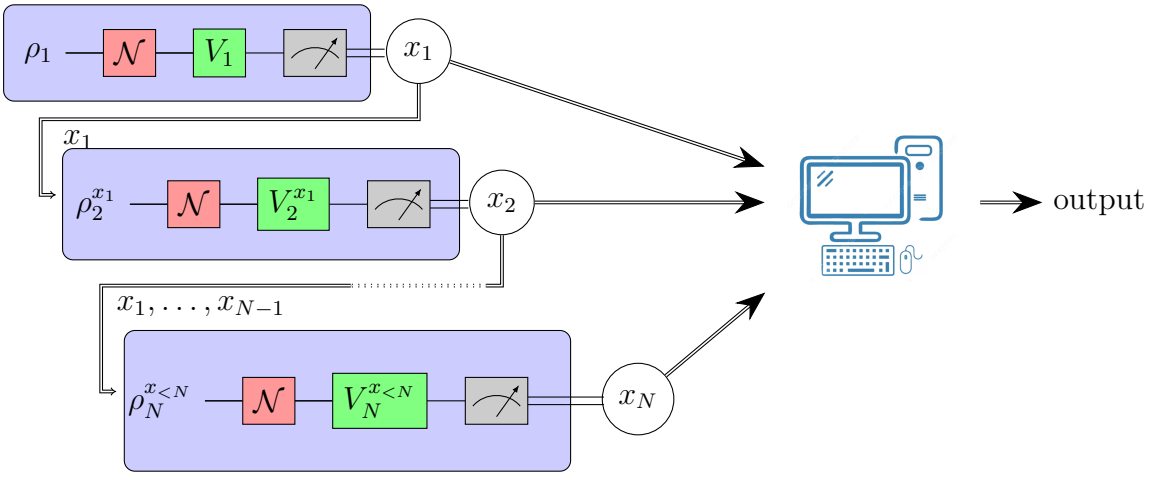

Figure 1.10: Illustration of an ancilla-free incoherent adaptive strategy for learning properties of quantum channels. The classical computer processes the observations $(x_1, \ldots, x_N)$ to distinguish between two hypotheses H_0/H_1 or produce an approximate channel $\tilde{\mathcal{N}}$.

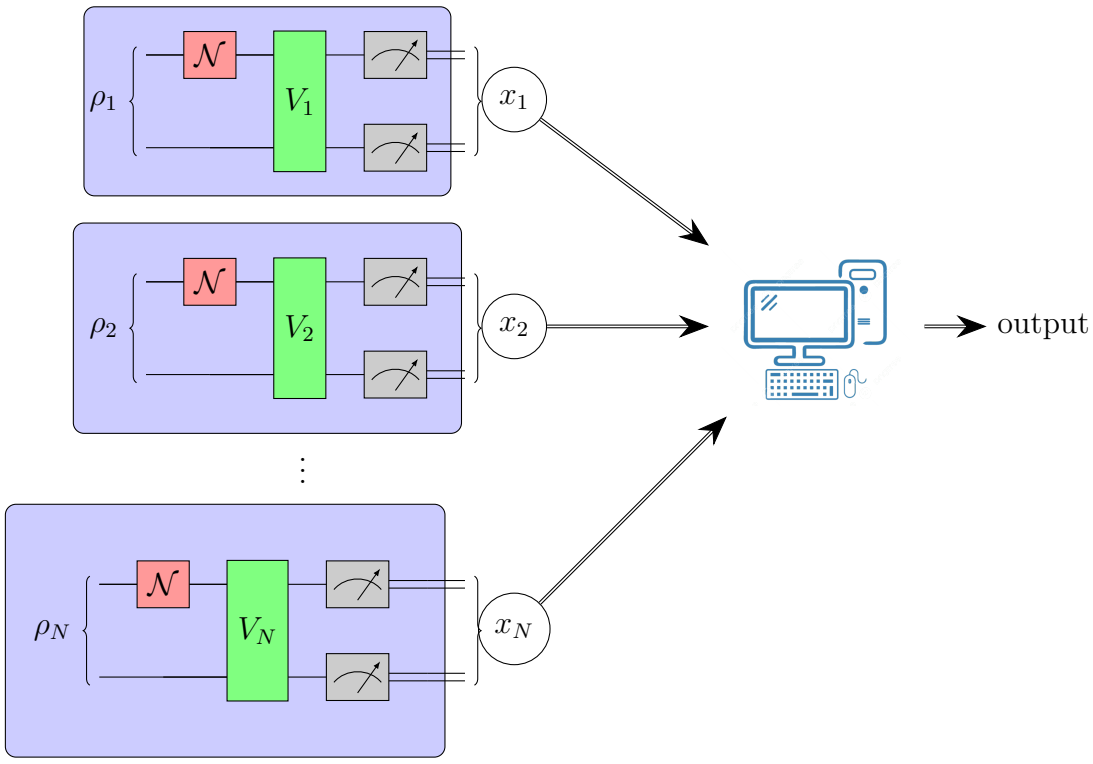

Figure 1.11: Illustration of an ancilla-assisted incoherent non-adaptive strategy for learning properties of quantum channels. The classical computer processes the observations $(x_1, \ldots, x_N)$ to distinguish between two hypotheses H_0/H_1 or produce an approximate channel $\tilde{\mathcal{N}}$.

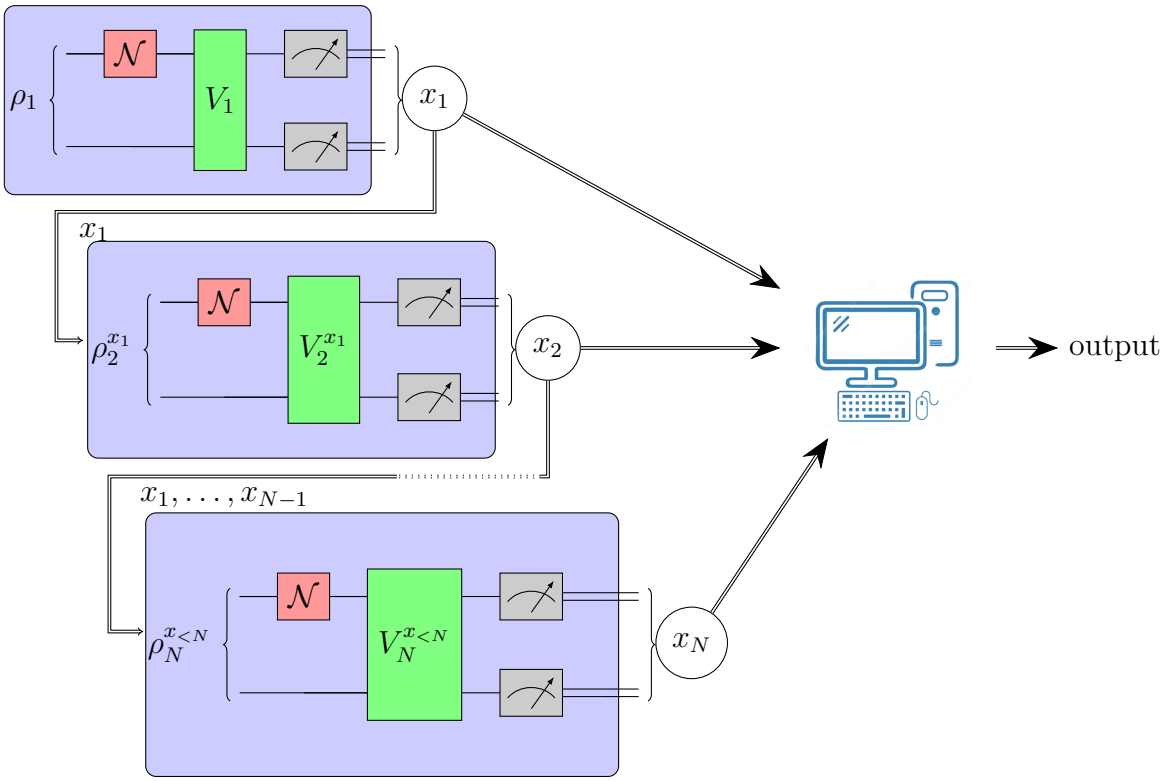

Figure 1.12: Illustration of an ancilla-assisted incoherent adaptive strategy for learning properties of quantum channels. The classical computer processes the observation $(x_1, \ldots, x_N)$ and distinguish between two hypotheses H_0/H_1 or produce an approximate channel $\tilde{\mathcal{N}}$.

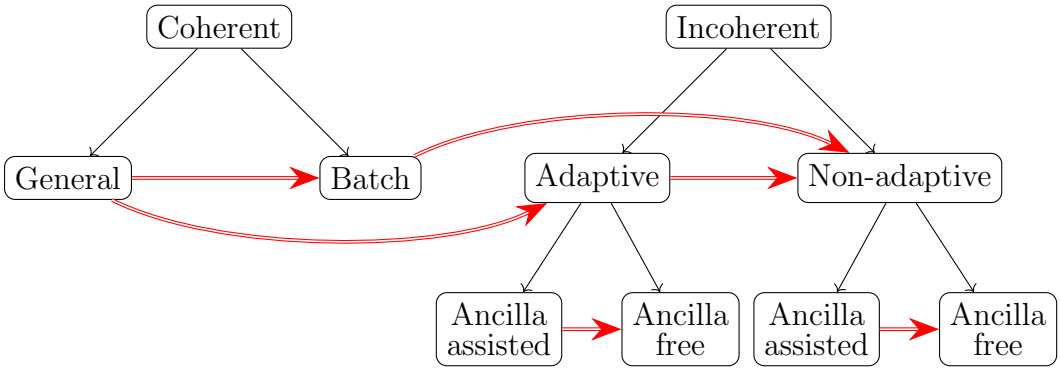

Figure 1.13: Different models of learning for quantum channels. A red arrow from strategy A to strategy B means that the former is more general than the latter.

1.4.4 Learning quantum channels

Learning completely a quantum channel is the most important question we can ask when we want to extract classical information from a quantum evolution. The reason is simple, once we have a classical approximation of the unknown process, we can compute (classically) an approximation of the output state of any input state of our choice. Thus we can answer any statistical test about the channel using its approximation. This problem is known in the literature as *"quantum process tomography"* which we define formally in the following.

Definition 1.4.9 (Quantum Process Tomography)**.** *Let $\varepsilon > 0$ be a precision parameter and $\mathcal{N} : \mathbb{C}^{d_{\text{in}} \times d_{\text{in}}} \to \mathbb{C}^{d_{\text{out}} \times d_{\text{out}}}$ be an unknown quantum channel. Given a collection of N copies of $\mathcal{N}$ and the ability to choose the input states and measure the corresponding output states, the task of quantum process tomography is to produce a quantum channel $\tilde{\mathcal{N}}$ such that with high probability:*

$$d_\diamond(\tilde{\mathcal{N}}, \mathcal{N}) \leq \varepsilon. \tag{1.38}$$

The natural figure of merit for this learning task is the diamond distance $d_\diamond$ because it characterizes the optimal error of discriminating two channels when auxiliary systems are allowed [Wat18]. The diamond distance $d_\diamond$ between two channels $\mathcal{N}$ and $\mathcal{M}$ is defined as:

$$d_\diamond(\mathcal{N}, \mathcal{M}) = \sup_{|\phi\rangle \in \mathbf{S}^{d_{\text{in}} \times d_{\text{in}}}} \|\text{id} \otimes (\mathcal{N} - \mathcal{M})(|\phi\rangle\langle\phi|)\|_1. \tag{1.39}$$

For the problem of quantum process tomography, we can characterize the optimal number of channel uses for non-adaptive incoherent strategies.

Theorem 1.4.1. *A number $N = \tilde{\Theta}\left(\frac{d_{\text{in}}^3 d_{\text{out}}^3}{\varepsilon^2}\right)$ of copies is sufficient and necessary to solve the quantum process tomography problem with non-adaptive strategies.*

The upper bound is achieved by an ancilla-free algorithm while the lower bound is proven for any ancilla-assisted strategy. We refer to Chapter 5 for the proof of this theorem. The algorithm and its analysis are similar to the ones proposed by [SSKKG22]. Then we use a standard method to prove lower bounds for learning problems known as Fano's inequality. To state this result, we need first to define the mutual information.

Mutual information In order to quantify the correlations between two discrete random variables X and Y, we use the mutual information which is given by the Kullback Leibler divergence between the joint distribution $\mathbb{P}_{(X,Y)}$ and the product distribution $\mathbb{P}_X \times \mathbb{P}_Y$:

Definition 1.4.10. *Given two random variables X and Y taking values in the sets $[n]$ and $[m]$ respectively, the mutual information between X and Y is defined as*

$$\mathcal{I}(X : Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{P}(X = i, Y = j) \log\left(\frac{\mathbb{P}(X = i, Y = j)}{\mathbb{P}(X = i)\,\mathbb{P}(Y = j)}\right). \tag{1.40}$$

The mutual information is non negative, symmetric and satisfies the Data-Processing inequality.

Proposition 1.4.1. *Let X and Y be two discrete random variables and f be a function. We have*

$$\mathcal{I}(X : f(Y)) \leq \mathcal{I}(X : Y). \tag{1.41}$$

When $X \sim \text{Uniform}([n])$, we have a useful lower bound on the mutual information known as Fano's inequality.

Proposition 1.4.2 (Fano, [Fan61])**.** *Let $X \sim \text{Uniform}([n])$ and Y be a random variable taking values in $[n]$. Let $\delta = \mathbb{P}(X \neq Y)$, the mutual information between X and Y is at least*

$$\mathcal{I}(X : Y) \geq (1 - \delta)\log(n) - \log(2). \tag{1.42}$$

This means that if X and Y are equal with a positive probability, then they are correlated and should share at least $\Omega(\log(n))$ nats of information. To use this inequality, we proceed by constructing an ε-separated family of quantum channels $\{\mathcal{N}_x\}_{x\in[n]}$. Then we use this family to encode the uniform random variable $X \sim \text{Uniform}([n])$ by the map $X \mapsto \mathcal{N}_X$. The next step is to use the learning algorithm to construct an approximation $\mathcal{M}$ of $\mathcal{N}_X$ to within $\varepsilon/2$. Since the family $\{\mathcal{N}_x\}_{x\in[n]}$ is ε-separated, the channel $\mathcal{M}$ is $\varepsilon/2$ close to at most one channel in the family, we denote it by $\mathcal{N}_Y$. This defines a random variable Y verifying $\delta = \mathbb{P}(Y \neq X) \leq 1/3$ since the learning algorithm approximates a channel with success probability at least $2/3$. By Fano's inequality, the random variables X and Y satisfy:

$$\mathcal{I}(X : Y) \geq 2/3\log(n) - \log(2). \tag{1.43}$$

Next, we need to pack a large number n of channels while keeping the mutual information $\mathcal{I}(X : Y)$ small. We construct the family of channels randomly by choosing the Choi states having the following expression:

$$\mathcal{J}_U = \frac{\mathbb{I}}{d_{\text{in}}d_{\text{out}}} + \frac{\varepsilon}{d_{\text{in}}d_{\text{out}}}(U + U^\dagger) - \frac{\varepsilon}{d_{\text{in}}d_{\text{out}}}\text{Tr}_2(U + U^\dagger) \otimes \frac{\mathbb{I}}{d_{\text{out}}} \tag{1.44}$$

where $U \sim \text{Haar}(d_{\text{in}}d_{\text{out}})$. This construction is different than the usual constructions for states because the Choi state $\mathcal{J}$ has the additional property $\text{Tr}_2(\mathcal{J}) = \frac{\mathbb{I}}{d_{\text{in}}}$. The existence of the family of cardinal $\exp(\Omega(d_{\text{in}}^2 d_{\text{out}}^2))$ is proven using the concentration inequality of Lipschitz functions of Haar distributed unitary matrices. Before stating this theorem, we recall the definition of a Lipschitz function.

Definition 1.4.11. *Let $f : \mathbb{U}(d)^n \to \mathbb{R}$. we say that f is an L-Lipschitz function if for all $U = (U_1, \ldots, U_n) \in \mathbb{U}(d)^n$ and $V = (V_1, \ldots, V_n) \in \mathbb{U}(d)^n$ we have*

$$|f(U) - f(V)| \leq L\|U - V\|_{2,HS} \tag{1.45}$$

where the 2-norm of Hilbert-Schmidt metric is defined as $\|U - V\|_{2,HS} = (\sum_{i=1}^n \|U_i - V_i\|_2^2)^{1/2}$.

Then we have the following concentration inequality.

Theorem 1.4.2 ([MM13])**.** *Let $M = U(d)^n$ endowed by the 2-norm of Hilbert-Schmidt metric. If $f : M \to \mathbb{R}$ is L-Lipschitz, then for any $t > 0$*

$$\mathbb{P}(|f(U_1, \ldots, U_n) - \mathbb{E}(f(U_1, \ldots, U_n))| \geq t) \leq \exp\left(-\frac{dt^2}{12L^2}\right), \tag{1.46}$$

where $U_1, \ldots, U_n$ are independent $\text{Haar}(d)$-distributed unitary matrices.

The last ingredient of the lower bound's proof is to upper bound the mutual information with an expression of $N, d_{\text{in}}, d_{\text{out}}$ and ε. For this we use Weingarten calculus.

Lemma 1.4.1 ([Gu13])**.** *Let U be a* Haar(d)*-distributed unitary matrix and $\{A_i, B_i\}_i$ be a sequence of $d \times d$ complex matrices. We have the following formula*

$$\mathbb{E}\left(\text{Tr}(UB_1U^\dagger A_1 U \dots UB_nU^\dagger A_n)\right) \tag{1.47}$$

$$= \sum_{\alpha,\beta\in\mathfrak{S}_n} \text{Wg}(\beta\alpha^{-1}, d)\text{Tr}_{\beta^{-1}}(B_1, \dots, B_n)\text{Tr}_{\alpha\gamma_n}(A_1, \dots, A_n), \tag{1.48}$$

where $\gamma_n = (12\cdots n)$ and $\text{Tr}_\sigma(M_1, \dots, M_n) = \Pi_j \text{Tr}(\Pi_{i\in C_j} M_i)$ for $\sigma = \Pi_j C_j$ and C_j are cycles.

The following values of Weingarten function suffice to compute the expectation of any polynomial in the entries of $U \sim$ Haar(d) of degree at most 6.

Lemma 1.4.2. *The Weingarten function* Wg(π, d) *depends only on the cycle type of the permutation $\pi \in \mathfrak{S}_n$. We have:*

- $\text{Wg}([1], d) = \frac{1}{d}$,
- $\text{Wg}(([2], d) = \frac{-1}{d(d^2-1)}$,
- $\text{Wg}([1, 1], d) = \frac{1}{d^2-1}$,
- $\text{Wg}([3], d) = \frac{2}{d(d^2-1)(d^2-4)}$,
- $\text{Wg}([2, 1], d) = \frac{-1}{(d^2-1)(d^2-4)}$ *and*
- $\text{Wg}([1, 1, 1], d) = \frac{d^2-2}{d(d^2-1)(d^2-4)}$.

The complexity of quantum process tomography is thus $\tilde{\Theta}\left(\frac{d_{\text{in}}^3 d_{\text{out}}^3}{\varepsilon^2}\right)$. Since this complexity is optimal, we should add more restrictions on the channel in order to reduce the query complexity of approximating it. One particularly important class of processes is given by Pauli channels. We will investigate the query complexity of learning a Pauli channel in the next section.

1.4.5 Learning pauli channels

In this section, we consider the problem of learning Pauli channels. This task is motivated by the fact that quantum computation are in reality imperfect. So we need to consider a model of computation where the noise is taken into account. Then, in order to correct the noise affecting this computation, we need to be able to learn its type. An important and simple model of n-qubit noise ($d_{\text{in}} = d_{\text{out}} = 2^n$) is given by Pauli channels since every noisy channel can be transformed to this form while keeping the same fidelity [WE16]. Formally, Pauli channels are quantum channels whose Kraus operators are weighted Pauli operators. For an n-qubit system, Pauli operators have the form:

$$P = P_1 \otimes P_2 \otimes \cdots \otimes P_n \tag{1.49}$$

where for each $i \in [n]$, P_i is a 1-qubit Pauli operator:

$$P_i \in \left\{ \mathbb{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \right\}. \tag{1.50}$$

These Pauli operators generalize the notion of an error that occurs on a qubit. For instance X can be seen as a bit flip error because it satisfies:

$$X|0\rangle = |1\rangle \qquad \text{and} \qquad X|1\rangle = |0\rangle. \tag{1.51}$$

On the other hand, Z can be seen as a phase (sign) flip error since it verifies:

$$Z|0\rangle = |0\rangle \qquad \text{and} \qquad Z|1\rangle = (-1)^1|1\rangle. \tag{1.52}$$

Finally, Y represents both flip and phase errors since $Y = (-i)ZX$.
We denote the set of n-qubit Pauli operators by $\mathbb{P}_n = \{\mathbb{I}, X, Y, Z\}^{\otimes n}$. Then a Pauli channel is any quantum channel of the form:

$$\mathcal{P}(\rho) = \sum_{P \in \{\mathbb{I},X,Y,Z\}^{\otimes n}} p(P) P\rho P \tag{1.53}$$

where $\{p(P)\}_{P\in\mathbb{P}_n}$ is a probability distribution. Note that the Kraus operators of this channel can be taken as $\left\{\sqrt{p(P)}P\right\}_{P\in\mathbb{P}_n}$ which satisfy:

$$\sum_{P\in\mathbb{P}_n} \left(\sqrt{p(P)}P\right)^\dagger \left(\sqrt{p(P)}P\right) = \sum_{P\in\mathbb{P}_n} p(P)P^\dagger P = \sum_{P\in\mathbb{P}_n} p(P)\mathbb{I} = \mathbb{I} \tag{1.54}$$

because Pauli operators are unitary. Given such channel as a black box, our goal is to learn it in the diamond norm.

Definition 1.4.12 (Learning Pauli Channels)**.** *Let $\varepsilon > 0$ be a precision parameter and $\mathcal{P} : (\mathbb{C}^2)^{\otimes n} \to (\mathbb{C}^2)^{\otimes n}$ be an unknown n-qubit Pauli channel. Given a collection of N copies of $\mathcal{P}$ and the ability to choose the input states and measure the corresponding output states, the task of learning a Pauli channel is to produce a Pauli channel $\tilde{\mathcal{P}}$ such that with high probability:*

$$d_\diamond(\tilde{\mathcal{P}}, \mathcal{P}) \leq \varepsilon. \tag{1.55}$$

Since Pauli channels are quantum channels, we can apply directly a quantum process tomography algorithm as in Theorem 1.4.1 which requires in this case a total number of channels uses $\mathcal{O}(n2^{6n}/\varepsilon^2)$ (because $d_{\text{in}} = d_{\text{out}} = 2^n$). However, Pauli channels are structured and their Kraus operators are not arbitrary: they should be among the Pauli operators. So we expect that learning a Pauli channel requires fewer resources than the general process tomography problem. Indeed, [FW20] propose an algorithm for learning p in the 2-norm using only $\tilde{\mathcal{O}}(2^n/\varepsilon^2)$ copies of the channel. In order to translate this result to a learning algorithm in the diamond norm it is sufficient to apply a Cauchy Schwarz inequality. Indeed, the diamond norm between two Pauli channels is equivalent to the 1-norm between their corresponding probability distributions [MGE12] hence:

$$\|\mathcal{P} - \tilde{\mathcal{P}}\|_\diamond = \|p - \tilde{p}\|_1 \leq 2^n \|p - \tilde{p}\|_2 \tag{1.56}$$

as $p - \tilde{p}$ is a vector of 4^n entries. Therefore we deduce the following upper bound for learning Pauli channels.

Theorem 1.4.3 ([FW20]). *There is a non-adaptive ancilla-free incoherent algorithm using $\mathcal{O}\left(n2^{3n}/\varepsilon^2\right)$ copies to learn a Pauli channel to within ε in diamond norm with at least a probability $2/3$.*

Note that this complexity is smaller than the one we obtain by applying quantum process tomography. We reproduce this algorithm and its analysis in Section 6.7. Moreover, we show a matching lower bound in the non-adaptive case.

Theorem 1.4.4. *Non-adaptive ancilla-free incoherent strategies for the problem of Pauli channel tomography require a number of channel uses satisfying:*

$$N = \Omega\left(\frac{2^{3n}}{\varepsilon^2}\right). \tag{1.57}$$

Furthermore, we prove a lower bound in the adaptive setting and high precision regime that matches the optimal non-adaptive complexity if the memory of the algorithm is limited.

Theorem 1.4.5. *Let $\varepsilon \leq 2^{-n-5}$. Adaptive ancilla-free incoherent strategies for the problem of Pauli channel tomography require a number of channel uses satisfying:*

$$N = \Omega\left(\frac{2^{5n/2}}{\varepsilon^2}\right). \tag{1.58}$$

Furthermore, any adaptive strategy that uses $\mathcal{O}(2^{2n}/\varepsilon^2)$ memory requires a number of channel uses verifying:

$$N = \Omega\left(\frac{2^{3n}}{\varepsilon^2}\right). \tag{1.59}$$

In both non-adaptive and adaptive cases, the proof of these lower bounds relies on Fano's inequality (Proposition 1.4.2). In the non-adaptive case we construct our family of Pauli channels hard to learn by choosing the probability p randomly. Concretely, the elements of this family have the form:

$$\mathcal{P}(\rho) = \sum_{P\in\{\mathbb{I},X,Y,Z\}^{\otimes n}} p(P)P\rho P = \sum_{P\in\{\mathbb{I},X,Y,Z\}^{\otimes n}} \left(\frac{1+4\alpha(P)\varepsilon}{4^n}\right) P\rho P \tag{1.60}$$

where $\alpha(P) = \pm 1$ to be chosen randomly so that $\alpha(P) = -\alpha(\sigma(P))$ for some matching σ of $\{\mathbb{I}, X, Y, Z\}^{\otimes n}$. However, it is challenging to obtain a non trivial adaptive lower bound with such construction. For this reason, we change the construction and choose the coefficients normally distributed. Precisely, the new construction have elements of the form:

$$\mathcal{P}(\rho) = \sum_{P\in\{\mathbb{I},X,Y,Z\}^{\otimes n}} p(P)P\rho P = \mathrm{Tr}(\rho)\frac{\mathbb{I}}{2^n} + \frac{2\varepsilon}{2^n} \sum_{P\in\{\mathbb{I},X,Y,Z\}^{\otimes n}} \frac{\tilde{\alpha}(P)}{\|\alpha\|_2} P\rho P \tag{1.61}$$

where $\tilde{\alpha}(P) = \alpha(P) - \frac{1}{4^n}\sum_{Q\in\mathbb{P}_n}\alpha(Q)$, $\{\alpha(P)\}_P$ are 4^n random variables i.i.d. as $\mathcal{N}(0,1)$ and $p(P) = \frac{1}{4^n} + \frac{2\varepsilon}{2^n}\cdot\frac{\tilde{\alpha}(P)}{\|\alpha\|_2}$. One inconvenience of this construction is that we add the condition $2^{n+5}\varepsilon \leq 1$ in order to be sure that $\mathcal{P}$ is a valid Pauli quantum channel. On the other hand, we can pack the same order of channels as in the previous construction

using a remarkable concentration inequality of Lipschitz functions of Gaussian random variables. Precisely, we say that $f : \mathbb{R}^k \to \mathbb{R}$ is an L-Lipschitz function if for all $x, y \in \mathbb{R}^k$ we have

$$|f(x) - f(y)| \leq L\|x - y\|_2. \tag{1.62}$$

Any such function concentrates around its mean if its entries are standard Gaussians.

Theorem 1.4.6 ([MS86])**.** *Let $(X_1, \ldots, X_k)$ be a vector of i.i.d. standard Gaussian variables and $f : \mathbb{R}^k \to \mathbb{R}$ be an L-Lipschitz function. We have for any $t > 0$:*

$$\mathbb{P}\left(|f(X_1, \ldots, X_k) - \mathbb{E}\left(f(X_1, \ldots, X_k)\right)| > t\right) \leq 2\exp\left(\frac{-t^2}{2L^2}\right). \tag{1.63}$$

Finally, with Gaussian variables instead of Bernoulli variables, we can control large products in a better way. To this end, we use Gaussian integration by parts (see e.g., [VH14]) which is a generalization of Isserlis' formula [Iss18].

Theorem 1.4.7 (Gaussian integration by parts)**.** *Let $(X_1, \ldots, X_k)$ be a Gaussian vector and $f : \mathbb{R}^k \to \mathbb{R}$ be a smooth function. We have:*

$$\mathbb{E}\left(X_1 f(X_1, \ldots, X_k)\right) = \sum_{i=1}^{k} \mathrm{Cov}(X_1, X_i)\mathbb{E}\left(\partial_i f(X_1, \ldots, X_k)\right) \tag{1.64}$$

We refer to Chapter 6 for the proofs of these lower bounds.
Another approach to reduce the copy complexity of quantum process tomography is to only consider testing properties of the quantum channel instead of completely approximating it. This will be the subject of the next section.

1.4.6 Testing quantum channels

Here our goal would be to check whether an unknown process behaves as intended. We consider a setting where we have two quantum channels $\mathcal{N}_0$ and $\mathcal{N}$. The first process $\mathcal{N}_0$ is completely known to the tester. For instance, the tester has a complete classical description of the Kraus operators or possibly the unitary operator that defines exactly the process. In particular, the tester knows what would be the output state after applying the process $\mathcal{N}_0$ on any chosen input state. We can think of the process $\mathcal{N}_0$ as an ideal channel. In reality, the tester interacts with an unknown process $\mathcal{N}$ that could be very different from the ideal process $\mathcal{N}_0$. The goal of this test is to certify whether the unknown process $\mathcal{N}$ is exactly what we think it should be (i.e., $\mathcal{N}_0$) or far from it. This testing question is important for various reasons. For instance, quantum computation can be modelled by quantum circuits. These are the analogue of classical circuits in the quantum world where bits become qubits and gates become quantum gates. These gates are nothing but unitary quantum channels. So, in order to succeed in a quantum computation, the quantum gates should act as designed. It is thus important to have a way to check these gates and characterize the exact amount of resources needed for this task. Let us give the formal definition of this testing problem.

Definition 1.4.13 (Quantum channel certification)**.** *Let $\varepsilon > 0$ be a threshold parameter and let $\mathcal{N}_0$ be a fixed known quantum channel. Given a collection of N copies of an unknown quantum channel $\mathcal{N}$ and the ability to choose the input states and measure the corresponding output states, the task of quantum channel certification is to distinguish between $\mathcal{N} = \mathcal{N}_0$ and $\mathrm{dist}(\mathcal{N}, \mathcal{N}_0) \geq \varepsilon$ with high probability.*

The distance in this type of tests is crucial and the optimal number of copies necessary varies substantially from a distance to another. Moreover, we would not consider a distance without an operational interpretation. Two particularly important examples of distances are related to the trace and diamond norms. Recall that the diamond distance between two channels $\mathcal{N}$ and $\mathcal{M}$ is

$$d_\diamond(\mathcal{N},\mathcal{M}) = \sup_{|\phi\rangle\in\mathbf{S}^{d_{\text{in}}\times d_{\text{in}}}} \|\text{id}\otimes(\mathcal{N}-\mathcal{M})(|\phi\rangle\langle\phi|)\|_1. \tag{1.65}$$

When auxiliary systems are not allowed, we obtain instead the trace distance

$$d_{\text{Tr}}(\mathcal{N},\mathcal{M}) = \sup_{|\phi\rangle\in\mathbf{S}^{d_{\text{in}}}} \|(\mathcal{N}-\mathcal{M})(|\phi\rangle\langle\phi|)\|_1. \tag{1.66}$$

So two channels are close in trace or diamond distance means that for *any* input state, the channels output close states in the 1-norm. Understanding the complexity of testing with these distances is important because we want to avoid situations where two channels appear close but can yield significantly different outputs for the same input. For instance, such a scenario could disrupt the overall functionality of a quantum computation.
Then, we need to fix a known process $\mathcal{N}_0$: this problem differs from one channel $\mathcal{N}_0$ to another. In this thesis, we focus on two extreme cases:

- **A fixed unitary channel.** Here $\mathcal{N}_0(\rho) = \mathcal{N}_U(\rho) = U\rho U^\dagger$ where U is a unitary matrix. This corresponds to the example of the evolution of a closed quantum system or a gate in a quantum circuit. It turns out that we can reduce this problem to testing to the identity channel where $U = \mathbb{I}$ and $\mathcal{N}_0(\rho) = \text{id}(\rho) = \rho$. We refer to this problem as *"testing identity to identity"*. This problem can be thought of the generalization of testing the classical identity channel, testing a rank 1 state or a Dirac distribution. In the setting of ancilla-free incoherent strategies, this problem has a different complexity than any of the mentioned problems.

- **The completely depolarizing channel.** Here $\mathcal{N}_0(\rho) = \mathcal{D}(\rho) = \text{Tr}(\rho)\frac{\mathbb{I}}{d_{\text{out}}}$ which corresponds to a noisy channel. Testing to such channel is the natural generalization to the well studied testing uniform of distribution problem [DGPP17] or testing mixedness of states [BCL20]. Furthermore, if the channel has no input $d_{\text{in}} = 1$ then testing such channels becomes exactly testing the corresponding states. Besides, testing to $\mathcal{D}$ is believed to be more difficult than any testing to another channel $\mathcal{N}_0$. We also conjecture that this test can be used as a subroutine for the general quantum certification problem as in [CLO22].

Testing identity to identity

We consider the quantum analogue of the test on classical channels in Section 1.4.1. Our proposed algorithm uses a random pure state as input and measure with the corresponding measurement device (the POVM constituted with this pure state and its complement). When the ideal channel we would like to test to is the identity channel, we know that the output state is the same as the input state. In this case we always see "0" under the null hypothesis and thus we can achieve a zero type I error. Under the alternate hypothesis, the channel $\mathcal{N}$ is far from the identity channel so we can see "1" with a positive probability. This latter can be boosted to 2/3 if the number of repetition is sufficiently large. Concretely we can prove:

Theorem 1.4.8. *There is an ancilla-free algorithm for testing identity to identity in the trace distance using only* $N = \mathcal{O}\left(\frac{d}{\varepsilon^2}\right)$ *incoherent measurements. Moreover, this algorithm can also solve the testing identity to identity problem in the diamond distance using only* $N = \mathcal{O}\left(\frac{d}{\varepsilon^4}\right)$ *incoherent measurements.*

The number of copies required by this test is smaller for the trace distance. However, we remark that the dependency in the threshold parameter ε is quadratic compared to the classical testing identity to identity. A natural question arises, is this complexity necessary or this algorithm is suboptimal? We answer this question in the following theorem:

Theorem 1.4.9. *Let* $\text{dist} \in \{d_{\text{Tr}}, d_\diamond\}$ *be the trace or diamond distance. Any adaptive ancilla-free strategy using incoherent measurements requires a number of steps satisfying:*

$$N = \Omega\left(\frac{d}{\varepsilon^2}\right) \tag{1.67}$$

to distinguish between $\mathcal{N} = \text{id}$ *and* $\text{dist}(\mathcal{N}, \text{id}) > \varepsilon$ *with a probability at least* $2/3$.

This theorem shows that a number $N = \Theta\left(\frac{d}{\varepsilon^2}\right)$ of channels is necessary and sufficient for testing identity to identity in the trace distance which is slightly surprising. Indeed, we usually obtain the same dependency on the threshold parameter ε when we consider the quantum analogue of a classical problem and a different dependency in the dimension parameter d. Here the reverse occurs. To prove this lower bound, we use the well known method of LeCam [LeC73]. Precisely, we consider two situations:

- **Null hypothesis** H_0**.** Here the unknown channel $\mathcal{N}$ is exactly the identity channel $\mathcal{N}(\rho) = \rho$.
- **Alternate hypothesis** H_1**.** Here we consider a random $\mathcal{N}$ channel ε far from the identity channel in the trace distance. Given a random unitary matrix $V \in \text{Haar}(d)$, we construct the channel $\mathcal{N}_V(\rho) = \frac{1}{2}\rho + \frac{1}{2}U_V \rho U_V^\dagger$ where U_V satisfies:

$$U_V V |l\rangle = \begin{cases} \sqrt{1-\varepsilon^2} V |0\rangle + \varepsilon V |1\rangle & \text{if } l = 0 \\ \sqrt{1-\varepsilon^2} V |1\rangle - \varepsilon V |0\rangle & \text{if } l = 1 \\ V |l\rangle & \text{otherwise.} \end{cases}$$

The intuition behind this choice is the following. We would like to make it difficult for the tester to find a state ρ such that the output states ρ (under H_0) and $\mathcal{N}_V(\rho)$ (under H_1) are distinguishable. To do so, we *"hide"* this maximizing state by choosing a random basis given by the Haar distributed unitary V. The other details of the construction including adding the off diagonal elements and taking the convex combination with the identity channel are important to achieve the optimal lower bound. Finally, the construction is inspired from the skew divergence [Aud14].
After a sufficient number of measurements, the observations under the two hypotheses should be distinguishable and thus, by the data processing inequality, they should be $\Omega(1)$ nats separated. However, we can show that for this particular choice of tested channels, any ancilla-free adaptive tester can only extract $\mathcal{O}(\varepsilon^2/d)$ nats of information after a measurement no matter the dependence on the previous observations. Our techniques are based on the Chain rule and Weingarten calculus.
We refer to Section 4.3 for the proof of these theorems.

Testing identity to the depolarizing channel

The second example we consider in this type of problems is when $\mathcal{N}_0 = \mathcal{D}$ is the completely depolarizing channel. If the unknown channel $\mathcal{N} = \mathcal{N}_0$ then the outputs are always equal to the maximally mixed state $\frac{\mathbb{I}}{d_{\text{out}}}$. In this case, we observe samples from the uniform distribution if we measure with any orthonormal basis. On the other hand, if $\mathcal{N}$ is different than $\mathcal{D}$ then there is at least one state ρ^* such that the output state $\mathcal{N}(\rho^*)$ is far from $\frac{\mathbb{I}}{d_{\text{out}}}$. If we know this state, then we can apply the testing mixedness algorithm of [BCL20] to distinguish between the two situations. Since we have no information about ρ^* (or any state playing the same role), we take inspiration from the testing identity to identity problem we discussed earlier and we choose a random rank 1 input state. It turns out that with such an input state we can detect whether the channel is completely depolarizing (noisy) or far from it. The reason behind this is that even though we have no information about ρ^*, a random rank 1 state $|\phi\rangle\langle\phi|$ has a non vanishing overlap with ρ^* with high probability. This overlap helps us to reduce the testing to the depolarizing channel problem to the testing mixedness problem [BCL20] with a new threshold parameter $\varepsilon' = \frac{\varepsilon}{2d_{\text{in}}\sqrt{d_{\text{out}}}}$. Precisely we show that:

Theorem 1.4.10. *There is an ancilla-free algorithm requiring a number of incoherent measurements*

$$N = \mathcal{O}\left(\frac{d_{\text{in}}^2 d_{\text{out}}^{1.5}}{\varepsilon^2}\right) \tag{1.68}$$

to distinguish between $\mathcal{N} = \mathcal{D}$ and $d_\diamond(\mathcal{N}, \mathcal{D}) \geq \varepsilon$ with a success probability $2/3$.

If the channels have no input $d_{\text{in}} = 1$, the channels are constant states and the certification to $\mathcal{D}$ problem becomes a certification to the state $\frac{\mathbb{I}}{d_{\text{out}}}$ problem. In this case, our algorithm uses $N = \mathcal{O}\left(\frac{d_{\text{out}}^{1.5}}{\varepsilon^2}\right)$ to solve the testing mixedness problem. Note that this complexity is optimal even for adaptive strategies [CHLL22]. However, we cannot rely merely on this particular case to deduce that our algorithm is optimal. For this reason, we prove a general lower bound in the non-adaptive and adaptive settings.

Theorem 1.4.11. *Let $\varepsilon \leq 1/32$ and $d_{\text{out}} \geq 10$. Any ancilla-free non-adaptive algorithm for testing identity to the depolarizing channel requires, in the worst case, a number of measurements satisfying:*

$$N = \Omega\left(\frac{d_{\text{in}}^2 d_{\text{out}}^{1.5}}{\log(d_{\text{in}} d_{\text{out}}/\varepsilon)^2 \varepsilon^2}\right). \tag{1.69}$$

Moreover, any ancilla-free adaptive algorithm for testing identity to the depolarizing channel requires, in the worst case,

$$N = \Omega\left(\frac{d_{\text{in}}^2 d_{\text{out}} + d_{\text{out}}^{1.5}}{\varepsilon^2}\right)$$

incoherent measurements.

The first lower bound shows that our algorithm is almost optimal and thus the optimal complexity in the non-adaptive setting is given by $N = \tilde{\Theta}\left(\frac{d_{\text{in}}^2 d_{\text{out}}^{1.5}}{\varepsilon^2}\right)$. The second lower bound shows that if adaptive strategies could outperform non-adaptive ones, then the

improvement is at most $d_{\text{out}}^{0.5}$.
The lower bound is proved using the same LeCam's method but with a different construction. In this case, we mix two hardness in the alternate hypothesis. First we hide the best input state, then we hide the eigenbasis of the best output state. Precisely we consider the two situations:

- **Null hypothesis H_0.** Here the unknown channel $\mathcal{N}$ is exactly the depolarizing channel $\mathcal{N}(\rho) = \mathcal{D}(\rho) = \text{Tr}(\rho)\frac{\mathbb{I}}{d_{\text{out}}}$.
- **Alternate hypothesis H_1.** Here we construct a random $\mathcal{N}$ channel ε far from the depolarizing channel having the expression $\mathcal{N}(\rho) = \mathcal{D}(\rho) + \frac{\varepsilon}{d_{\text{out}}} \langle w| \rho |w\rangle U$ where $|w\rangle$ is a Haar distributed vector and U has Gaussian entries: for all $i, j \in [d_{\text{out}}]$, $U_{j,i} = U_{i,j} \sim \mathbf{1}\{i \neq j\}\mathcal{N}(0, 16/d_{\text{out}})$ conditioned on the event $\mathcal{G} = \{\|U\|_1 \geq d_{\text{out}}, \|U\|_\infty \leq 32\}$.

We carry out the analysis using the Hypercontractivity of Gaussian polynomials (see e.g., [Jan97; AS17]).
We refer to Section 4.4 for the proof of these theorems.

Contents of the thesis

The remaining chapters of this thesis are based on my publications, in the following order:

- Chapter 2 is an extended version [FFGO22] of the article [FFGO21], published in NeurIPS 2021 and co-authored with Omar Fawzi, Nicolas Flammarion and Aurélien Garivier. It studies "Sequential Algorithms for Testing Identity and Closeness of Distributions".
- Chapter 3 corresponds to the article [FFGO23a], published in TMLR and co-authored with Omar Fawzi, Nicolas Flammarion and Aurélien Garivier. It studies "On Adaptivity in Quantum Testing".
- Chapter 4 corresponds to the article [FFGO23b], accepted to COLT 2023 and co-authored with Omar Fawzi, Nicolas Flammarion and Aurélien Garivier. It studies "Quantum Channel Certification with Incoherent Strategies".
- Chapter 5 corresponds to the article [Ouf23] published in ISIT 2023 and selected for a talk at TQC 2023. It studies "Sample-Optimal Quantum Process Tomography with Non-Adaptive Incoherent Measurements".
- Chapter 6 corresponds to the article [FOF23], submitted to IEEE Transactions on Information Theory and co-authored with Omar Fawzi and Daniel Stilck França. It studies "Lower Bounds on Learning Pauli Channels".

Chapter 2

Sequential Algorithms for Testing Identity and Closeness

2.1 Introduction

How to test if two discrete sources of randomness are similar or distinct? This basic and ubiquitous question is surprisingly not closed if frugality matters, that is if one wants to take the right decision using as few samples as possible.

To state the problem more precisely, one first needs to define what "distinct" means. In this chapter, we endow the set of probability distributions on $\{1, \dots, n\}$ with the *total variation distance* TV, and we fix a tolerance parameter $\varepsilon \in [0, 1]$. We consider two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$, and we assume that either $\mathcal{D}_1 = \mathcal{D}_2$ or $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$. Whenever $0 < \mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) \leq \varepsilon$, we do not expect any determined behaviour from our test. Two cases occur:

- when the first distribution $\mathcal{D}_1$ is fixed and known to the algorithm (but not $\mathcal{D}_2$), we say that we are *testing identity* using independent samples of $\mathcal{D}_2$;
- when both distributions are unknown we are *testing closeness*, based on an equal number of independent samples of both distributions.

We also need to specify what kind of "test" is considered. Here we treat the two hypotheses symmetrically (there is no "null hypothesis") : given a fixed risk $\delta \in (0, 1)$, we expect our procedure to find the true one with probability $1 - \delta$, whichever it is. We call such a procedure δ-*correct*.

Finally, we consider and compare two notions of "frugality": in the *batch* setting, the agent specifies in advance the number of samples needed for the test: it makes its decision just after observing the data all at once, and the sample complexity of the test is the smallest sample size of a $\delta-$correct procedure. In the *sequential* setting, the agent observes the samples one by one, and decides accordingly whether to make a decision or request more samples before deciding . Then, the sample complexity of the test is the smallest *expected number of samples* needed before a δ-correct procedure takes a decision. Note that this expected number depends on the unknown distributions $\mathcal{D}_1$ and $\mathcal{D}_2$, which turns out to be an important advantage of sequential procedures.

Contributions When $n \geq 2$ is a small constant integer, we show that the optimal sample complexities can be precisely characterized (up to lower order terms in ε and δ) in

Model	Lower bound	Upper bound
Batch	$8\frac{\lfloor n^2/4 \rfloor}{n^2}\log(1/\delta)\varepsilon^{-2}$ $-\mathcal{O}\left(n\log\log(1/\delta)\varepsilon^{-2}\right)$	$8\frac{\lfloor n^2/4 \rfloor}{n^2}\log(1/\delta)\varepsilon^{-2}$ $+8\frac{\lfloor n^2/4 \rfloor}{n^2}(n+1)\varepsilon^{-2}$
Sequential (τ_1)	$2\frac{\lfloor n^2/4 \rfloor}{n^2}\log(1/\delta)\varepsilon^{-2}$ $-\mathcal{O}(\varepsilon^{-2})$	$2\frac{\lfloor n^2/4 \rfloor}{n^2}\log(1/\delta)\varepsilon^{-2}$ $+\mathcal{O}\left((n+\log(1/\delta)^{2/3})\varepsilon^{-2}\right)$
Sequential (τ_2)	$2\frac{\lfloor n^2/4 \rfloor}{n^2}\log(1/\delta)d^{-2}$ $-\mathcal{O}(d^{-2})$	$2\frac{\lfloor n^2/4 \rfloor}{n^2}\log(1/\delta)d^{-2}$ $+\mathcal{O}\left((n+\log(1/\delta)^{2/3})d^{-2}\right)$

Table 2.1: Lower and upper bounds on sample complexity for uniformity testing in batch and sequential setting with $d = \mathrm{TV}(\mathcal{D}, U_n)$. τ_1 (resp. τ_2) represents the stopping time of the sequential algorithm when $\mathcal{D} = U_n$ (resp. $\mathrm{TV}(\mathcal{D}, U_n) > \varepsilon$). The $\mathcal{O}$ hides universal constants.

Model	Lower bound	Upper bound
Batch	$4\log(1/\delta)\varepsilon^{-2} - \mathcal{O}(\log\log(1/\delta)\varepsilon^{-2})$	$4\log(1/\delta)\varepsilon^{-2} + \mathcal{O}(n\varepsilon^{-2})$
Sequential (τ_1)	$\log(1/\delta)\varepsilon^{-2} - \mathcal{O}(\varepsilon^{-2})$	$\log(1/\delta)\varepsilon^{-2}$ $+\mathcal{O}((n+\log(1/\delta)^{2/3})\varepsilon^{-2})$
Sequential (τ_2)	$\log(1/\delta)d^{-2} - \mathcal{O}(d^{-2})$	$\log(1/\delta)d^{-2}$ $+\mathcal{O}\left((n+\log(1/\delta)^{2/3})d^{-2}\right)$

Table 2.2: Lower and upper bounds on the sample complexities for testing closeness in the batch and sequential settings with $d = \mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2)$. τ_1 (resp. τ_2) represents the stopping time of the sequential algorithm when $\mathcal{D}_1 = \mathcal{D}_2$ (resp. $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$). The $\mathcal{O}$ hides universal constants.

both the batch and sequential setting as shown in Table 2.1 and Table 2.2. This establishes a provable advantage for sequential strategies over batch strategies when $n \ll \log(1/\delta)$: sequential algorithms reduce the sample complexity by a factor of at least 4, and can stop rapidly if the tested distributions are far (i.e., $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$). The improvements of the sequential algorithm are illustrated in Figure 2.1. The sequential algorithms use stopping rules inspired by time uniform concentration inequalities. The problems of testing identity and closeness for small n are studied in Section 2.3 and Section 2.4 respectively.

For general $n \geq 2$, we improve the dependence on ε to $\varepsilon \vee \mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2)$ in the best batch algorithm due to [DGKPP20], which is known to be optimal up to multiplicative constants. Namely we obtain a sequential closeness testing algorithm using a number of samples given by

$$\mathcal{O}\left(\max\left(\frac{n^{2/3}\log^{1/3}(1/\delta)}{(\varepsilon \vee \mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2))^{4/3}}, \frac{n^{1/2}\log^{1/2}(1/\delta)}{(\varepsilon \vee \mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2))^{2}}, \frac{\log(1/\delta)}{(\varepsilon \vee \mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2))^{2}}\right)\right). \tag{2.1}$$

A doubling search technique could also lead to the same order of sample complexity, we explain this method and compare it with our proposed algorithm in Remark 2.6.1.

As a special case, when $\varepsilon = 0$ (the algorithm should not stop when $\mathcal{D}_1 = \mathcal{D}_2$ in this case) we show that there is an algorithm that stops after

$$\mathcal{O}\left(\max\left(\frac{\log\log(1/d)}{d^2}, \frac{n^{2/3}\log\log(1/d)^{1/3}}{d^{4/3}}, \frac{n^{1/2}\log\log(1/d)^{1/2}}{d^2}\right)\right) \tag{2.2}$$

samples where $d = \mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > 0$. This is an improvement over the sequential algorithm of [DK17] which uses $\Theta(\frac{n/\log n}{d^2}\log\log(1/d))$ samples. We design the stopping

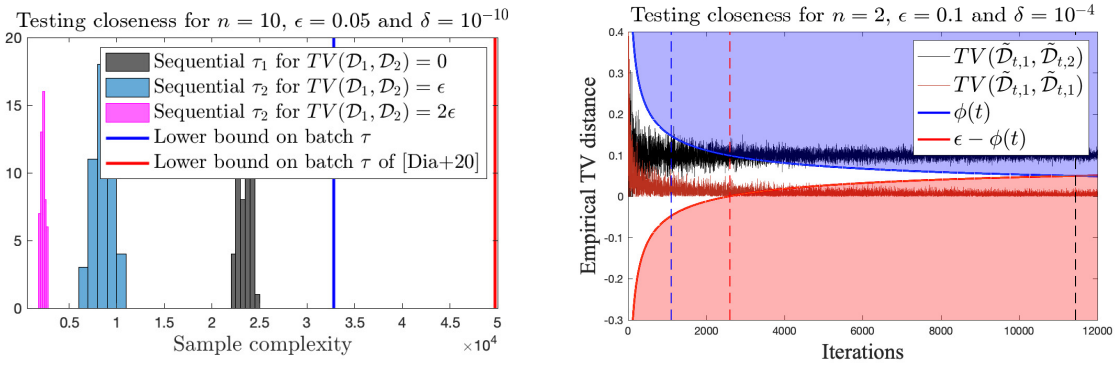

Figure 2.1: Left: histogram of the stopping times for 100 Monte-Carlo experiments. Black: $\mathcal{D}_1 = \mathcal{D}_2 = U_n$, blue (resp. magenta): $\mathcal{D}_1 = U_n$ and $\mathcal{D}_2 = \{(1 \pm 2\varepsilon)/n\}$ (resp. $\{(1 \pm 4\varepsilon)/n\}$). Right: $\mathcal{D}_1 = U_2$ and $\mathcal{D}_2 = \{(1 \pm 2\varepsilon)/2\}$. The sequential tester stops as soon as the statistic enters the red region (for H_1) or blue region (for H_2) whereas the batch tester waits for the red and blue regions to cover the whole segment $[0, 1]$. The blue/red and black dashed lines represent respectively the stopping times of the sequential and batch algorithms. The blue lower bound for batch algorithms is taken from Proposition 2.4.1. We note that, in both cases, the sequential tester stops long before the batch algorithm.

rules according to a time uniform concentration inequality deduced from McDiarmid's inequality, where we use the ideas of [HRMS18; HRMS20] in order to obtain powers of $\log\log(1/d)$ instead of $\log(1/d)$.

We show that the sample complexity for the testing closeness problem given by Equation (2.1) is optimal up to multiplicative constants in the worst case setting (i.e., when looking for a bound independent of the distributions $\mathcal{D}_1$ and $\mathcal{D}_2$). To do so, we construct two families of distributions whose cross TV distance is exactly $d \geq \varepsilon$ and hard to distinguish unless we have a number of samples given by Equation (2.1). This latter lower bound is based on properties of KL divergence along with Wald's Lemma. Using similar techniques, we also establish upper and lower bounds for testing identity that match up to multiplicative constants.

In addition, we establish a lower bound on the number of queries that matches Equation (2.2) up to multiplicative constants. The proof is inspired by [KK07] who proved lower bounds for testing whether the mean of a sequence of i.i.d. Bernoulli variables is smaller or larger than $1/2$. We construct well-chosen distributions $\mathcal{D}_k$ (for k integer) that are at distance ε_k (ε_k decreasing to 0) from uniform and then use properties of the Kullback-Leibler's divergence to show that no algorithm can distinguish between $\mathcal{D}_k$ and uniform using fewer samples than in Equation (2.2). Note that we could have used the testing closeness lower bound described in the previous paragraph and let $\varepsilon = 0$, however this gives sub-optimal lower bounds.

Discussion of the setting and related work It is clearly impossible to test $\mathcal{D}_1 = \mathcal{D}_2$ versus $\mathcal{D}_1 \neq \mathcal{D}_2$ in finite time: this is why the slack parameter ε is introduced in this setting. Other authors like [DK17] make a different choice: they fix no ε, but only require that the test decides for $\mathcal{D}_1 \neq \mathcal{D}_2$ as soon as it can, and never stops with high probability when $\mathcal{D}_1 = \mathcal{D}_2$.

We focus on the TV distance in testing closeness problems because it characterises the probability of error for the problem of distributions discrimination ; as noted by [DKW18],

using other distances such as KL and χ^2 is in general impossible.

For an overview of testing discrete distributions we recommend the survey of [Can20]. Testing identity for the uniform distribution was solved by [Pan08], then for general distribution by [VV17] and finally the high probability version by [DGPP17]. Likewise testing closeness was solved by [CDVV14], and a distribution dependent complexity was found by [DK16] and finally the high probability version by [DGKPP20]. Besides, the problem of testing $\mathcal{D}_1 = \mathcal{D}_2$ vs $\mathcal{D}_1 \neq \mathcal{D}_2$ was solved by [DK17] for $n = 2$, however the constants are not optimal. They also propose algorithms for the general case using black-box reduction from non-sequential hypothesis testers. Sequential and adaptive procedures have also been explored in active hypothesis setting [NJ13], channels' discrimination [Hay09] and quantum hypothesis testing [LTT22; LHT22a]. Sequential strategies have been also considered for testing continuous distributions by [ZZSE16] and [BR15]. In the latter, the authors design sequential algorithms whose stopping time adapts to the unknown difficulty of the problem. The techniques used are time uniform concentration inequalities which are surveyed by [HRMS20]. In contrast to the present work, however, they test properties of the *means* of the distributions.

2.2 Preliminaries

We mostly follow [DK17] for the notation.

2.2.1 Testing identity

Given two distributions $\mathcal{D}$ (known) and $\mathcal{D}'$ (unknown) on $[n] := \{1, \dots, n\}$, we want to distinguish between two hypothesis $H_1 : \mathcal{D}' = \mathcal{D}$ and $H_2 : \mathrm{TV}(\mathcal{D}', \mathcal{D}) > \varepsilon$. We call a stopping rule a function $T : [n]^* \to \{0, 1, 2\}$ such that if $T(x) \neq 0$ then $T(xy) = T(x)$ for all strings x and y. $T(x) = 1$ (resp. $T(x) = 2$) means that the rule accepts H_1 (resp. H_2) after seeing x while $T(x) = 0$ means the rule does not make a choice and continues sampling. We define two different stopping times, the first $\tau_1(T, \mathcal{D}') = \inf\{t, T(x_1 \cdots x_t) = 1\}$ and the second $\tau_2(T, \mathcal{D}') = \inf\{t, T(x_1 \cdots x_t) = 2\}$ where $x_1, \dots$ are i.i.d. samples from $\mathcal{D}'$. We want to find stopping rules satisfying

1. $\mathbb{P}\left(\tau_2(T, \mathcal{D}) \leq \tau_1(T, \mathcal{D})\right) \leq \delta$ and
2. $\mathbb{P}\left(\tau_1(T, \mathcal{D}') \leq \tau_2(T, \mathcal{D}')\right) \leq \delta$ whenever $\mathrm{TV}(\mathcal{D}', \mathcal{D}) > \varepsilon$.

We call such a stopping rule δ-correct. Our goal is to minimize the expected sample complexity $\mathbb{E}(\tau_1(T, \mathcal{D}))$ in case of the input is from $\mathcal{D}$ and $\mathbb{E}(\tau_2(T, \mathcal{D}'))$ in case of the input is from $\mathcal{D}'$ such that $\mathrm{TV}(\mathcal{D}', \mathcal{D}) > \varepsilon$.

A batch algorithm is one for which $\tau = \tau_1 = \tau_2$ is a constant random variable which only depends on δ, ε, n and $\mathcal{D}$.

2.2.2 Testing closeness

Given two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ on $\{1, \dots, n\}$ we want to distinguish between two hypothesis $H_1 : \mathcal{D}_1 = \mathcal{D}_2$ and $H_2 : \mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$. We call a stopping rule a function $T : \bigcup_{k \in \mathbb{N}} [n]^k \times [n]^k \to \{0, 1, 2\}$ such that if $T(x, y) \neq 0$ then $T(xz, yt) = T(x, y)$ for all strings x, y, z, t with $|x| = |y|$ and $|z| = |t|$. $T(x, y) = 1$ (resp. $T(x, y) = 2$) means that the rule accepts H_1 (resp. H_2) after seeing the sequences x and y while $T(x, y) = 0$

means the rule does not make a choice and continue sampling. We define two different stopping times, the first $\tau_1(T, \mathcal{D}_1, \mathcal{D}_2) = \inf\{t, T(x_1 \cdots x_t, y_1 \cdots y_t) = 1\}$ and the second $\tau_2(T, \mathcal{D}_1, \mathcal{D}_2) = \inf\{t, T(x_1 \cdots x_t, y_1 \cdots y_t) = 2\}$ where $x_1, \ldots$ are i.i.d. samples from $\mathcal{D}_1$ and $y_1, \ldots$ samples from $\mathcal{D}_2$. We want to find stopping rules satisfying

1. $\mathbb{P}\left(\tau_2(T, \mathcal{D}_1, \mathcal{D}_2) \leq \tau_1(T, \mathcal{D}_1, \mathcal{D}_2)\right) \leq \delta$ if $\mathcal{D}_1 = \mathcal{D}_2$ and
2. $\mathbb{P}\left(\tau_1(T, \mathcal{D}_1, \mathcal{D}_2) \leq \tau_2(T, \mathcal{D}_1, \mathcal{D}_2)\right) \leq \delta$ whenever $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$.

We call such stopping rules δ-correct. Our goal is to minimize the expected sample complexity $\mathbb{E}(\tau_1(T, \mathcal{D}_1, \mathcal{D}_2))$ in case of the input is from $\mathcal{D}_1$, $\mathcal{D}_2$ such that $\mathcal{D}_1 = \mathcal{D}_2$ and $\mathbb{E}(\tau_2(T, \mathcal{D}_1, \mathcal{D}_2))$ in case of the input is from $\mathcal{D}_1$, $\mathcal{D}_2$ such that $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$.

2.3 Testing identity for small n

In this section, we focus on small, constant values of n where $n \geq 2$ and we consider two distributions, $\mathcal{D} = U_n$ is the uniform distribution on $[n]$ and $\mathcal{D}'$ is an unknwon distribution on $[n]$. In this case, the hypothesis H_1 becomes $\mathcal{D}' = U_n$ and H_2 becomes $\mathrm{TV}(\mathcal{D}', U_n) > \varepsilon$. We are interested in precisely comparing the sample complexity of testing identity in the sequential versus the batch setting. In order to find the optimal constant, we first need to obtain a sharp lower bound in the batch setting, which is done directly by using Stirling's approximation. We then turn to the sequential case.

2.3.1 Batch setting

In the batch setting, the number of steps τ is fixed before the test. The tester samples $A_1, \ldots, A_\tau \sim \mathcal{D}'$ and decides according to the comparison between the empirical TV distance $\mathrm{TV}(\tilde{\mathcal{D}}'_\tau, U_n)$ and $\varepsilon/2$ where $\tilde{\mathcal{D}}'_\tau = \left\{\left(\sum_{j=1}^{\tau} A_j = i\right)/\tau\right\}_{i\in[n]}$. If $\mathrm{TV}(\tilde{\mathcal{D}}'_\tau, U_n) \leq \varepsilon/2$ it accepts H_1 and rejects it otherwise. In order to control the number of steps τ so that the error of this algorithm does not exceed δ, Chernoff–Hoeffding's inequality ([Hoe63]) writes for i.i.d. random variables $X_1, \ldots, X_\tau \sim \mathrm{Bern}(q)$:

$$\mathbb{P}\left(\frac{\sum_{i=1}^{\tau} X_i}{\tau} - q > \frac{\varepsilon}{2}\right) \leq e^{-\tau\,\mathrm{KL}(q+\varepsilon/2, q)} \text{ and } \mathbb{P}\left(\frac{\sum_{i=1}^{\tau} X_i}{\tau} - q < -\frac{\varepsilon}{2}\right) \leq e^{-\tau\,\mathrm{KL}(q-\varepsilon/2, q)}\,. \tag{C-H}$$

We use the following property of TV distance:

$$\mathrm{TV}(\mathcal{D}', U_n) = \max_{B\subset[\lfloor n/2\rfloor]} |\mathcal{D}'(B) - |B|/n| = |\mathcal{D}'(B_{opt}) - |B_{opt}|/n|, \tag{2.3}$$

and choose $X_i = \mathbf{1}_{A_i \in B_{opt}} \sim \mathrm{Bern}(\mathcal{D}'(B_{opt}))$.

Applying these inequalities for $\mathcal{D}' = U_n$ (to control the type I error) and for $\mathcal{D}' \neq U_n$ (to control the type II error) prove that this test is δ-correct if

$$\tau = \max_{b\in[n]} \left\{ \frac{\log(2/\delta)}{\mathrm{KL}(b/n \pm \varepsilon/2, b/n \pm \varepsilon)}, \frac{\log(2^{n+1}/\delta)}{\mathrm{KL}(b/n \pm \varepsilon/2, b/n)} \right\}, \tag{2.4}$$

where $\mathrm{KL}(p, q) = \mathrm{KL}(\mathrm{Bern}(p) \| \mathrm{Bern}(q))$ denotes the Kullback-Leibler divergence.

To see this, we analyze the three cases: $\mathcal{D}' = U_n$, $\mathcal{D}'(B_{opt}) - |B_{opt}|/n > \varepsilon$ and $\mathcal{D}'(B_{opt}) - |B_{opt}|/n < -\varepsilon$. They are all handled by a simple application of Chernoff–Hoeffding's inequality (C-H):

- If $\mathcal{D}' = U_n$, the probability of error is given by

$$\mathbb{P}\left(\left|\mathrm{TV}(\tilde{\mathcal{D}}', U_n)\right| > \frac{\varepsilon}{2}\right) = \mathbb{P}\left(\exists B \subset [n] : \left|\tilde{\mathcal{D}}'(B) - |B|/n\right| > \frac{\varepsilon}{2}\right) \tag{2.5}$$

$$\leq \sum_{B \subset [\lfloor n/2 \rfloor]} e^{-\tau \,\mathrm{KL}(|B|/n+\varepsilon/2, |B|/n)} + e^{-\tau \,\mathrm{KL}(1-|B|/n+\varepsilon/2, 1-|B|/n)} \tag{2.6}$$

$$\leq \delta\,. \tag{2.7}$$

- If $\mathcal{D}'(B_{opt}) - |B_{opt}|/n > \varepsilon$, the probability of error is given by

$$\mathbb{P}\left(\left|\mathrm{TV}(\tilde{\mathcal{D}}', U_n)\right| \leq \frac{\varepsilon}{2}\right) \leq \mathbb{P}\left(\tilde{\mathcal{D}}'(B_{opt}) - |B_{opt}|/n \leq \frac{\varepsilon}{2}\right) \tag{2.8}$$

$$\leq e^{-\tau \,\mathrm{KL}(|B_{opt}|/n+\varepsilon/2, \mathcal{D}'(B_{opt}))} \tag{2.9}$$

$$\leq e^{-\tau \,\mathrm{KL}(|B_{opt}|/n+\varepsilon/2|, B_{opt}|/n+\varepsilon)} \tag{2.10}$$

$$\leq \delta \tag{2.11}$$

  where we use the fact that the function $x \mapsto \mathrm{KL}(p, x)$ is increasing on $(p, 1)$.

- If $\mathcal{D}'(B_{opt}) - |B_{opt}|/n < -\varepsilon$, the probability of error is given by

$$\mathbb{P}\left(\left|\mathrm{TV}(\tilde{\mathcal{D}}', U_n)\right| \leq \frac{\varepsilon}{2}\right) \leq \mathbb{P}\left(\tilde{\mathcal{D}}'(B_{opt}) - |B_{opt}|/n \geq -\frac{\varepsilon}{2}\right) \tag{2.12}$$

$$\leq e^{-\tau \,\mathrm{KL}(|B_{opt}|/n-\varepsilon/2, \mathcal{D}'(B_{opt}))} \tag{2.13}$$

$$\leq e^{-\tau \,\mathrm{KL}(|B_{opt}|/n-\varepsilon/2|, B_{opt}|/n-\varepsilon)} \tag{2.14}$$

$$\leq \delta \tag{2.15}$$

  where we use the fact that the function $x \mapsto \mathrm{KL}(p, x)$ is decreasing on $(0, 1)$.

We show in the following theorem that this number of steps τ is necessary.

Theorem 2.3.1. *In the batch setting, any δ-correct algorithm testing identity to $\mathcal{D} = U_n$ requires at least τ samples, where*

$$\tau \geq \max_{b \in [n]} \min\left\{\frac{\log(1/\delta)}{\mathrm{KL}(b/n + \varepsilon/2, b/n + \varepsilon)}, \frac{\log(1/\delta)}{\mathrm{KL}(b/n + \varepsilon/2, b/n)}\right\} - \mathcal{O}\left(\frac{n \log\log 1/\delta}{\varepsilon^2}\right). \tag{2.16}$$

This lower bound has the simple equivalent $8\frac{\lfloor n^2/4 \rfloor}{n^2}\log(1/\delta)\varepsilon^{-2} - \mathcal{O}\left(n \log\log(1/\delta)\varepsilon^{-2}\right)$ when $\varepsilon \to 0$ (see Lemma 2.4.1 for the equivalent of KL divergence). To prove this lower bound, we show that every δ-correct tester can be transformed into a test which depends only on the numbers of $1's, 2's, \ldots, n's$ occurred on $\{A_1, \ldots, A_\tau\}$. We then consider the distribution $\mathcal{D}'$ with roughly half parts are $1/n + \varepsilon/\lfloor n/2 \rfloor$ and the others are $1/n - \varepsilon/\lceil n/2 \rceil$and derive tight lower bounds on the probability mass function of the multinomial distribution.

Proof. We consider such a δ-correct test $A : \{1, \ldots, n\}^\tau \to \{1, 2\}$, it sees a word consisting of τ samples either from a distribution ε-far from U_n or U_n and returns 1 if it thinks the

the samples come from U_n and 2 otherwise. We construct another test $B : \{1, \dots, n\}^\tau \to \{0, 1\}$ by the expression

$$B(x) = \mathbf{1}\left\{ \sum_{\sigma \in \mathfrak{S}_\tau} A(\sigma(x)) - 1 \geq \tau!/2 \right\}, \tag{2.17}$$

The test B has the property of invariance under the action of the symmetric group. Moreover, B is 2δ-correct. Indeed, if x represents τ i.i.d. samples from U_n, then for all $\sigma \in \mathfrak{S}_\tau$, $\sigma(x)$ represents also τ i.i.d. samples from U_n hence by Markov's inequality:

$$\begin{aligned}
\mathbb{P}\left(B(x) = 1\right) &= \mathbb{P}\left(\sum_{\sigma \in \mathfrak{S}_\tau} A(\sigma(x)) - 1 \geq \tau!/2 \right) && (2.18)\\
&\leq \frac{2}{\tau!} \sum_{\sigma \in \mathfrak{S}_\tau} \mathbb{E}\left(A(\sigma(x)) - 1\right) && (2.19)\\
&\leq \frac{2}{\tau!} \sum_{\sigma \in \mathfrak{S}_\tau} \mathbb{P}\left(A(\sigma(x)) = 2\right) && (2.20)\\
&\leq \frac{2}{\tau!} \sum_{\sigma \in \mathfrak{S}_\tau} \delta && (2.21)\\
&= 2\delta. && (2.22)
\end{aligned}$$

Similarly, if x represents τ i.i.d. samples from a distribution $\mathcal{D}$ that is ε-far from U_n, then for all $\sigma \in \mathfrak{S}_\tau$, $\sigma(x)$ represents also τ i.i.d. samples from $\mathcal{D}$ hence by Markov's inequality:

$$\begin{aligned}
\mathbb{P}\left(B(x) = 0\right) &= \mathbb{P}\left(\sum_{\sigma \in \mathfrak{S}_\tau} A(\sigma(x)) - 1 < \tau!/2 \right) && (2.23)\\
&\leq \frac{2}{\tau!} \sum_{\sigma \in \mathfrak{S}_\tau} \mathbb{E}\left(2 - A(\sigma(x))\right) && (2.24)\\
&\leq \frac{2}{\tau!} \sum_{\sigma \in \mathfrak{S}_\tau} \mathbb{P}\left(A(\sigma(x)) = 1\right) && (2.25)\\
&\leq \frac{2}{\tau!} \sum_{\sigma \in \mathfrak{S}_\tau} \delta && (2.26)\\
&= 2\delta. && (2.27)
\end{aligned}$$

Let $d \in [n]$ and suppose that $\tau\left(\frac{1}{n} + \frac{\varepsilon}{2d}\right)$ and $\tau\left(\frac{1}{n} - \frac{\varepsilon}{2(n-d)}\right)$ are integers. Consider the word

$$w = 1^{\tau\left(\frac{1}{n} + \frac{\varepsilon}{2d}\right)} \dots d^{\tau\left(\frac{1}{n} + \frac{\varepsilon}{2d}\right)} (d+1)^{\tau\left(\frac{1}{n} - \frac{\varepsilon}{2(n-d)}\right)} \dots n^{\tau\left(\frac{1}{n} - \frac{\varepsilon}{2(n-d)}\right)}, \tag{2.28}$$

we have two choices, either $B(w) = 0$ or $B(w) = 1$, we suppose the first and take q the distribution defined by $q_1 = \dots = q_d = \frac{1}{n} + \frac{\varepsilon}{d}$ and $q_{d+1} = \dots = q_n = \frac{1}{n} - \frac{\varepsilon}{n-d}$. It satisfies $\mathrm{TV}(q, U_n) = \varepsilon$ thus $\mathbb{P}_q(x_1 \cdots x_\tau = w) \leq \delta$ hence

$$\binom{\tau}{\tau_1 \cdots \tau_n} \left(\frac{1}{n} + \frac{\varepsilon}{d}\right)^{d\tau_1} \left(\frac{1}{n} - \frac{\varepsilon}{n-d}\right)^{(n-d)\tau_{d+1}} \leq \delta \tag{2.29}$$

where $\tau_1 = \cdots = \tau_d = \tau\left(\frac{1}{n} + \frac{\varepsilon}{2d}\right)$ and $\tau_{d+1} = \cdots = \tau_n = \tau\left(\frac{1}{n} - \frac{\varepsilon}{2(n-d)}\right)$ thus

$$\frac{\tau!}{(\tau_1!)^d(\tau_{d+1}!)^{n-d}}\left(\frac{1}{n}+\frac{\varepsilon}{d}\right)^{d\tau_1}\left(\frac{1}{n}-\frac{\varepsilon}{n-d}\right)^{(n-d)\tau_{d+1}} \leq \delta, \tag{2.30}$$

which implies by Stirling's approximation

$$\frac{e(\tau/e)^\tau}{(e\tau_1(\tau_1/e)^{\tau_1})^d(e\tau_{d+1}(\tau_{d+1}/e)^{\tau_{d+1}})^{n-d}}e^{-d\tau_1\log\frac{\tau_1}{q_1}}\left(\frac{1}{n}+\frac{\varepsilon}{d}\right)^{d\tau_1}\left(\frac{1}{n}-\frac{\varepsilon}{n-d}\right)^{(n-d)\tau_{d+1}} \leq \delta, \tag{2.31}$$

after simplifying we obtain

$$\frac{e}{(e\tau_1)^d(e\tau_{d+1})^{n-d}}e^{-d\tau_1\log\frac{\tau_1}{q_1}}e^{-(n-d)\tau_{d+1}\log\frac{\tau_{d+1}}{q_{d+1}}} \leq \delta, \tag{2.32}$$

or

$$\frac{e}{(e\tau_1)^d(e\tau_{d+1})^{n-d}}e^{-\tau\,\mathrm{KL}(d/n+\varepsilon/2,d/n+\varepsilon)} \leq \delta, \tag{2.33}$$

Finally

$$\tau \geq \frac{\log(1/\delta)+1-n}{\mathrm{KL}(d/n+\varepsilon/2,d/n+\varepsilon)} - \mathcal{O}\left(n\log\log(1/\delta)\varepsilon^{-2}\right). \tag{2.34}$$

If $B(w) = 1$, we consider $q = U_n$ and we obtain with the same approach

$$\tau \geq \frac{\log(1/\delta)+1-n}{\mathrm{KL}(d/n+\varepsilon/2,d/n)} - \mathcal{O}\left(n\log\log(1/\delta)\varepsilon^{-2}\right). \tag{2.35}$$

These lower bounds work for all $d \in [n]$ therefore

$$\tau \geq \max_{d\in[n]}\min\left\{\frac{\log(1/\delta)}{\mathrm{KL}(d/n+\varepsilon/2,d/n)}, \frac{\log(1/\delta)}{\mathrm{KL}(d/n+\varepsilon/2,d/n+\varepsilon)}\right\} - \mathcal{O}\left(\frac{n\log\log(1/\delta)}{\varepsilon^2}\right). \tag{2.36}$$

□

This simple analysis relies on well-known arguments for testing Bernoulli variables $\mathcal{D}_1 = \mathrm{Bern}(p)$ and $\mathcal{D}_2 = \mathrm{Bern}(q)$. For example, [AB09] and [KK07] test whether $q = 1/2 + \varepsilon$ or $q = 1/2 - \varepsilon$ with an error probability δ. [AB09] show that we need roughly $\log(1/\delta)\varepsilon^{-2}/4$ samples while [KK07] prove that $2\log(1/\delta)\varepsilon^{-2}$ samples are sufficient. If ε is not known to the tester, sequential algorithms prove to be essential. Indeed, [KK07] manage to prove that $\Theta(\log\log(1/|q-1/2|)|q-1/2|^{-2})$ is necessary and sufficient to test $q > 1/2$ vs $q < 1/2$ with an error probability $1/3$. In what follows, we use sequential algorithms to expose the dependency on $\mathrm{TV}(\mathcal{D}', U_n)$ for the testing identity problem.

2.3.2 Sequential setting

If one wants to leverage the sequential setting to improve the optimal sample complexity of testing identity, it is natural to first investigate how it can be improved by removing the batch assumption of the previous lower-bound in Theorem 2.3.1. We first state a new lower bound inspired by the work of [GK19].

Lemma 2.3.1. *Let $\mathcal{D} = U_n$ be the uniform distribution. Let T a stopping rule for testing identity: $\mathcal{D}' = U_n$ vs $\mathrm{TV}(\mathcal{D}', U_n) > \varepsilon$ with an error probability δ. Let τ_1 and τ_2 the associated stopping times. We have*

- $\mathbb{E}(\tau_1(T, U_n) \geq \frac{\log(1/3\delta)}{\min_{b\in[n]}\{\mathrm{KL}(b/n, b/n\pm\varepsilon)\}}$ *if* $\mathcal{D}' = U_n$.
- $\mathbb{E}(\tau_2(T, \mathcal{D}')) \geq \frac{\log(1/3\delta)}{\min\{\mathrm{KL}(|B_{opt}|/n\pm d, |B_{opt}|/n)\}}$ *if* $d = \mathrm{TV}(\mathcal{D}', U_n) = |\mathcal{D}'(B_{opt}) - |B_{opt}|/n| > \varepsilon$.

An average number of samples equivalent to $2\frac{\lfloor n^2/4\rfloor}{n^2}\log(1/3\delta)\varepsilon^{-2}$ (by Lemma 2.4.1) is thus necessary when the tester can access sequentially to the samples, which is roughly 4 times less than the complexity obtained in Theorem 2.3.1 for the batch setting. The proof, with a strong information-theoretic flavor, compares two situations: when the samples are from equal distributions and when they are from ε-far distributions. Those samples cannot be distinguished until their size is large enough, as can be proved by combining properties of Kullback-Leibler's divergence and Wald's lemma.

Proof. We apply the lower bounds of Section 2.7.1. If $\mathcal{D} = U_n$, we set $\mathcal{D}_b^+$ the distribution whose first b parts are equal to $1/n + \varepsilon/b$ and the others are equal to $1/n - \varepsilon/(n-b)$.

$$\mathbb{E}(\tau_1(\mathcal{D})) \geq \frac{\log(1/3\delta)}{\min \mathcal{D}'' \text{s.t. } \mathrm{TV}(\mathcal{D}'', \mathcal{D}) > \varepsilon\, \mathrm{KL}(\mathcal{D}, \mathcal{D}'')} \tag{2.37}$$

$$\geq \frac{\log(1/3\delta)}{\min_{b\in[n]} \mathrm{KL}(\mathcal{D}, \mathcal{D}_b^+)} \tag{2.38}$$

$$\geq \frac{\log(1/3\delta)}{\min_{b\in[n]} \mathrm{KL}(b/n, b/n + \varepsilon/n)}. \tag{2.39}$$

Likewise we can prove for $\mathcal{D}_b^-$ the distribution whose first b parts are equal to $1/n - \varepsilon/b$ and the others are equal to $1/n + \varepsilon/(n-b)$:

$$\mathbb{E}(\tau_1(\mathcal{D})) \geq \frac{\log(1/3\delta)}{\min_{\mathcal{D}'' \text{s.t. } \mathrm{TV}(\mathcal{D}'', \mathcal{D}) > \varepsilon} \mathrm{KL}(\mathcal{D}, \mathcal{D}'')} \tag{2.40}$$

$$\geq \frac{\log(1/3\delta)}{\min_{b\in[n]} \mathrm{KL}(\mathcal{D}, \mathcal{D}_b^-)} \tag{2.41}$$

$$\geq \frac{\log(1/3\delta)}{\min_{b\in[n]} \mathrm{KL}(b/n, b/n - \varepsilon/n)}. \tag{2.42}$$

Finally:

$$\mathbb{E}(\tau_1(\mathcal{D})) \geq \frac{\log(1/3\delta)}{\min_{b\in[n]} \mathrm{KL}(b/n, b/n \pm \varepsilon/n)}. \tag{2.43}$$

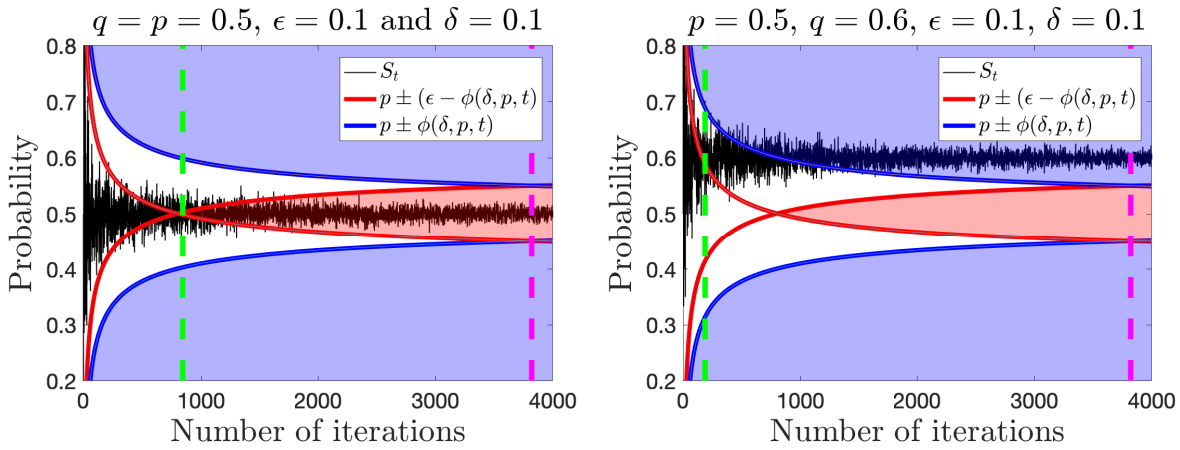

Figure 2.2: Testing identity for Bernoulli ($n = 2$). Left: $q = p = 0.5$ and $\varepsilon = 0.1$. Right: $p = 0.5$, $q = 0.6$ and $\varepsilon = 0.1$. The sequential tester stops as soon as S_t enters the red region (for H_1) or blue region (for H_2) whereas the batch tester waits for the red and blue regions to cover the whole segment $[0, 1]$. The green and magenta dashed lines represent respectively the stopping time of the sequential and batch algorithms. We note that, in both cases, the sequential tester stops long before the batch algorithm.

Now, in order to prove a lower bound on τ_2, we focus on distributions that are ε-far from U_n and have the same length of B_{opt}.

$$\sup_{\mathcal{D}:\mathrm{TV}(\mathcal{D},U_n)=d>\varepsilon,|B_{opt}(\mathcal{D})|=b} \mathbb{E}(\tau_2(\mathcal{D})) \geq \sup_{\mathcal{D}:\mathrm{TV}(\mathcal{D},U_n)=d>\varepsilon,|B_{opt}(\mathcal{D})|=b} \frac{\log(1/3\delta)}{\mathrm{KL}(\mathcal{D}, U_n)}. \tag{2.44}$$

$$\geq \frac{\log(1/3\delta)}{\mathrm{KL}(b/n \pm \varepsilon/n, b/n)}. \tag{2.45}$$

□

In the sequential testing, the tester chooses when to stop according to the previous observations $(A_1, \ldots, A_t)$, making comparisons at each step t. The key explanation of the sequential speedup is that the tester can stop as soon as it is sure that it can accept one of the hypothesis H_1 or H_2. On the contrary, in the batch setting it had to sample enough observation to be simultaneously sure that either H_1 or H_2 hold. In this aim, at each time step, after sampling a new observation X_t, it compares the updated empirical TV distance $S_t = \max_{B\subset[\lfloor n/2\rfloor]} |\tilde{\mathcal{D}'}_t(B) - |B|/n|$ to specific thresholds and sees if (a) S_t is sufficiently far from 0 to surely accept H_2, (b) S_t is sufficiently close to ε to surely accept H_1, (c) it is unsure and needs further sample to take a sound decision. This test is formally described in Algorithm 1 and its execution is illustrated in Figure 2.2 for $n = 2$.

To control the sample complexity of such sequential algorithms, we need here Chernoff-Hoeffding's inequality (C-H) and the union bound:

$$\mathbb{P}\left(\exists B \subset [\lfloor n/2\rfloor], \exists t \geq 1 : \frac{\sum_{i=1}^{t} \mathbf{1}_{A_i\in B}}{t} > \mathcal{D}'(B) + \phi(\delta, \mathcal{D}'(B), t)\right) \tag{2.46}$$

$$\leq \sum_{B\subset[\lfloor n/2\rfloor],t\geq 1} e^{-t\,\mathrm{KL}(\mathcal{D}'(B)+\phi,\mathcal{D}'(B))} = \sum_{B\subset[\lfloor n/2\rfloor],t\geq 1} \frac{\delta}{2^{n-1}t(t+1)} = \frac{\delta}{2}, \tag{2.47}$$

Algorithm 1 Distinguish between $\mathcal{D}' = U_n$ and $\mathrm{TV}(\mathcal{D}', U_n) > \varepsilon$ with high probability

Require: $A_1, \dots$ samples from $\mathcal{D}'$
Ensure: Accept if $\mathcal{D}' = U_n$ and Reject if $\mathrm{TV}(\mathcal{D}', U_n) > \varepsilon$ with probability of error less than δ
$t = 1$, $W = 1$
while $W = 1$ **do**
$\tilde{\mathcal{D}}'_t = \left\{ \left(\sum_{j=1}^{t} \mathbf{1}_{A_j = i} \right) / t \right\}_{i \in [n]}$
if $\exists B \subset [\lfloor n/2 \rfloor] : |\tilde{\mathcal{D}}'_t(B) - |B|/n| > \max\{\phi(\delta, |B|/n, t), \phi(\delta, 1 - |B|/n, t)\}$ **then**
$W = 0$
return 2
else if $\forall B \subset [\lfloor n/2 \rfloor] : (\tilde{\mathcal{D}}'_t(B) - |B|/n)^+ < \varepsilon - \phi(\delta, |B|/n + \varepsilon, t)$ and $(|B|/n - \tilde{\mathcal{D}}'_t(B))^+ < \varepsilon - \phi(\delta, |B|/n - \varepsilon, t)$ **then**
$W = 0$
return 1
else
$t = t + 1$
end if
end while

and

$$\mathbb{P}\left(\exists B \subset [\lfloor n/2 \rfloor], \exists t \geq 1 : \frac{\sum_{i=1}^{t} \mathbf{1}_{A_i \in B}}{t} < \mathcal{D}'(B) - \phi(\delta, 1 - \mathcal{D}'(B), t) \right) \tag{2.48}$$

$$\leq \sum_{B \subset [\lfloor n/2 \rfloor], t \geq 1} e^{-t\,\mathrm{KL}(\mathcal{D}'(B) - \phi, \mathcal{D}'(B))} = \sum_{B \subset [\lfloor n/2 \rfloor], t \geq 1} \frac{\delta}{2^{n-1} t(t+1)} = \frac{\delta}{2}\,, \tag{2.49}$$

where $\phi(\delta, p, t)$ is the real function[1] implicitly defined as the solution of the equation $\mathrm{KL}(p + \phi(\delta, p, t), p) = \log\left(\frac{2^{n-1}t(t+1)}{\delta}\right)/t$. The function ϕ is a key ingredient in designing the stopping rules of Algorithm 1 since it enables to directly bound its sample complexity in terms of the expression of the Kullback-Leibler's divergence. The stopping time of Algorithm 1 is $\tau = \min\{\tau_1, \tau_2\}$ where τ_1 and τ_2 are defined as follows:

$$\tau_1 = \inf\big\{ t \geq 1 : \forall B \subset [\lfloor n/2 \rfloor]\ (\tilde{\mathcal{D}}_t(B) - |B|/n)^+ < \varepsilon - \phi(\delta, |B|/n + \varepsilon, t), \tag{2.50}$$

$$(|B|/n - \tilde{\mathcal{D}}_t(B))^+ < \varepsilon - \phi(\delta, |B|/n - \varepsilon, t) \big\} \tag{2.51}$$

$$\tau_2 = \inf\big\{ t \geq 1 : \exists B \subset [\lfloor n/2 \rfloor]\ |\tilde{\mathcal{D}}_t(B) - |B|/n| > \max\{\phi(\delta, |B|/n, t), \phi(\delta, 1 - |B|/n, t)\} \big\}\,. \tag{2.52}$$

Note that the stopping time of the algorithm is random. Yet, we can show that this algorithm stops always before the batch one and give an upper bound on the expected stopping time τ (or expected sample complexity) using the inequality $\mathbb{E}(\tau) \leq N + \sum_{t \geq N} \mathbb{P}(\tau \geq t)$, where N is chosen so that $\mathbb{P}(\tau \geq t)$ is (exponentially) small for $t \geq N$. In the following theorem, we state an upper bound on the estimated sample complexity of this algorithm.

Theorem 2.3.2. *The Algorithm 1 is δ-correct and its stopping times can be bounded in*

[1] This function is well defined whenever $\log\left(\frac{2^{n-1}t(t+1)}{\delta}\right) < t \log(1/p)$ and can easily be approximated as a zero of a one-dimensional convex function

expectation for $n < \log^{2/3}(1/\delta)$ *as follows:*

$$\mathbb{E}(\tau_1(U_n)) \leq \frac{\log(2^{n-1}/\delta)}{\min_{b\in[n]}\{\mathrm{KL}(b/n, b/n \pm \varepsilon)\}} + \mathcal{O}\left(\frac{\log(2^{n-1}/\delta)^{2/3}}{\varepsilon^2}\right), \text{ and} \tag{2.53}$$

$$\mathbb{E}(\tau_2(\mathcal{D}')) \leq \frac{\log(2^{n-1}/\delta)}{\min\{\mathrm{KL}(|B_{opt}|/n \pm d, |B_{opt}|/n)\}} + \mathcal{O}\left(\frac{\log(2^{n-1}/\delta)^{2/3}}{d^2}\right) \tag{2.54}$$

if $d = \mathrm{TV}(\mathcal{D}', U_n) = |\mathcal{D}'(B_{opt}) - |B_{opt}|/n| > \varepsilon$.

These upper bounds are tight in the sense that they match the asymptotic lower bounds of Lemma 2.3.1 if $n \ll \log(1/\delta)$. We see here the many advantages of the sequential setting as shown in Figure 2.2: (a) the sequential algorithm stops always before the non-sequential algorithm since after the batch complexity the region of decisions of the sequential algorithm intersect, (b) the estimated sample complexity is 4 times less than the optimal complexity in the non sequential setting, (c) the sample complexity can be very small if the probability mass is concentrated on a small set i.e. $|B_{opt}| \ll n$ and (d) the sample complexity in the sequential setting depends on the unknown distribution $\mathcal{D}'$ through the distance $\mathrm{TV}(\mathcal{D}', U_n)$. Note that this cannot be the case in the batch setting as the number of sample should be fixed beforehand. This attribute makes a considerable difference when $\mathcal{D}'$ is very different from U_n. Nevertheless, the above lower bounds and upper bounds do not match exactly, the dependence on n cannot be avoided if n is of the order (or larger) of $\log(1/\delta)$. For this reason, we try in Section 2.5 to somehow truncate our algorithm in a way to get the best sample complexity in every regime. In the previous results the choice of $\mathcal{D} = U_n$ is not crucial, we can easily generalize them by replacing $|B_{opt}|/n$ by $\mathcal{D}(B_{opt})$.

Proof. We first prove the correctness of Algorithm 1, i.e., that it has an error probability less than δ. Let us recall the useful concentration inequalities which can be simply proven using Chernoff-Hoeffding's inequalities and union bounds.

Lemma 2.3.2. *If* $A_1, \ldots, A_t$ *are i.i.d. random variables with the law* $\mathcal{D}'$*, we have the following inequalities:*

$$\mathbb{P}\left(\exists B \subset [\lfloor n/2 \rfloor], \exists t \geq 1 : \frac{\sum_{i=1}^t \mathbf{1}_{A_i \in B}}{t} > \mathcal{D}'(B) + \phi(\delta, \mathcal{D}'(B), t)\right) \leq \frac{\delta}{2}, \tag{2.55}$$

$$\mathbb{P}\left(\exists B \subset [\lfloor n/2 \rfloor], \exists t \geq 1 : \frac{\sum_{i=1}^t \mathbf{1}_{A_i \in B}}{t} < \mathcal{D}'(B) - \phi(\delta, 1 - \mathcal{D}'(B), t)\right) \leq \frac{\delta}{2}, \tag{2.56}$$

Using this lemma we can conclude:

- If $\mathcal{D}' = U_n$, the probability of error is given by

$$\mathbb{P}(\tau_2 \leq \tau_1) \leq \mathbb{P}\left(\exists t \geq 1 : \max_{B\subset[\lfloor n/2 \rfloor]} |\tilde{\mathcal{D}'}_t(B) - |B|/n| > \max\{\phi(\delta, |B|/n, t), \phi(\delta, 1 - |B|/n, t)\}\right) \tag{2.57}$$

$$\leq \mathbb{P}\left(\exists t \geq 1, \exists B \subset [\lfloor n/2 \rfloor] : \tilde{\mathcal{D}'}_t(B) - |B|/n > \phi(\delta, |B|/n, t)\right) \tag{2.58}$$

$$+ \mathbb{P}\left(\exists t \geq 1, \exists B \subset [\lfloor n/2 \rfloor] : \tilde{\mathcal{D}'}_t(B) - |B|/n < -\phi(\delta, 1 - |B|/n, t)\right) \tag{2.59}$$

$$\leq \delta\,. \tag{2.60}$$

- If $\mathcal{D}'(B_{opt}) - |B_{opt}|/n > \varepsilon$, the probability of error is given by

$$\mathbb{P}(\tau_1 \leq \tau_2) \leq \mathbb{P}\left(\exists t \geq 1 : \tilde{\mathcal{D}}'_t(B_{opt}) - |B_{opt}|/n < \varepsilon - \phi(\delta, 1 - |B_{opt}| - \varepsilon, t)\right) \tag{2.61}$$

$$\leq \sum_{t\geq 1} e^{-t\,\mathrm{KL}(1-|B_{opt}|/n-\varepsilon+\phi(\delta,1-|B_{opt}|/n-\varepsilon,t),1-\mathcal{D}'(B_{opt}))} \tag{2.62}$$

$$\leq \sum_{t\geq 1} e^{-t\,\mathrm{KL}(|B_{opt}|/n+\varepsilon-\phi(\delta,1-|B_{opt}|/n-\varepsilon,t),|B_{opt}|/n+\varepsilon)} \tag{2.63}$$

$$\leq \sum_{t\geq 1} \frac{\delta}{2t(t+1)} \leq \delta \tag{2.64}$$

  where we use the fact that $x \mapsto \mathrm{KL}(p, p-x)$ is increasing on $(0, p)$.

- If $\mathcal{D}'(B_{opt}) - |B_{opt}|/n < -\varepsilon$, the probability of error is given by

$$\mathbb{P}(\tau_1 \leq \tau_2) \leq \mathbb{P}\left(\exists t \geq 1 : \tilde{\mathcal{D}}'_t(B_{opt}) - |B_{opt}|/n > -\varepsilon + \phi(\delta, 1 - |B_{opt}|/n - \varepsilon, t)\right) \tag{2.65}$$

$$\leq \sum_{t\geq 1} e^{-t\,\mathrm{KL}(|B_{opt}|/n-\varepsilon+\phi(\delta,|B_{opt}|/n-\varepsilon,t),\mathcal{D}'(B_{opt}))} \tag{2.66}$$

$$\leq \sum_{t\geq 1} e^{-t\,\mathrm{KL}(|B_{opt}|/n-\varepsilon+\phi(\delta,|B_{opt}|/n-\varepsilon,t),|B_{opt}|/n-\varepsilon)} \tag{2.67}$$

$$\leq \sum_{t\geq 1} \frac{\delta}{2t(t+1)} \leq \delta \tag{2.68}$$

  where we use the fact that $x \mapsto \mathrm{KL}(p, x)$ is decreasing on $(0, p)$.

This conclude the proof of the correctness of Algorithm 1.

Let us prove the upper bounds on the expected stopping times of Algorithm 1. We first prove the asymptotic bounds, then provide the proof for the non-asymptotic bounds. The proofs rely on the following well-known lemma:

Lemma 2.3.3. *Let T be a random variable taking values in $\mathbb{N}^*$, we have for all $N \in \mathbb{N}^*$*

$$\mathbb{E}(T) \leq N + \sum_{t\geq N} \mathbb{P}(T \geq t)\,. \tag{2.69}$$

Proof. Since the random variable T takes values in $\mathbb{N}^*$, we have:

$$\mathbb{E}(T) = \sum_{t\geq 1} \mathbb{P}(T \geq t) = \sum_{t=1}^{N-1} \mathbb{P}(T \geq t) + \sum_{t\geq N} \mathbb{P}(T \geq t) \tag{2.70}$$

$$\leq N + \sum_{t\geq N} \mathbb{P}(T \geq t)\,. \tag{2.71}$$

□

Recall that ϕ is defined by the relation $\mathrm{KL}(p + \phi(\delta, p, t), p) = \log\left(\frac{2^{n-1}t(t+1)}{\delta}\right)/t$. Pinsker's inequality ([RW09]) implies $0 < \phi(\delta, p, t) \leq \sqrt{\frac{1}{2t}\log\left(\frac{2^{n-1}t(t+1)}{\delta}\right)}$, hence

$\lim_{t\to\infty}\phi(\delta,p,t)=0$ and $\phi(\delta,p,t)$ exhibits the following asymptotic behavior:

$$\phi(\delta,p,t)\underset{t\to\infty}{\sim}\sqrt{\frac{2p(1-p)}{t}\log\left(\frac{2^{n-1}t(t+1)}{\delta}\right)} \tag{2.72}$$

since $\mathrm{KL}(p+x,p)\underset{x\to 0}{\sim}\frac{x^2}{2p(1-p)}$. Fix a parameter $0<\alpha<1$, and let N the minimum positive integer such that for all integers $t\geq N$, $\max_{b\in[n/2]}\{\phi(\delta,b/n-\varepsilon,t),\phi(\delta,1-b/n-\varepsilon,t)\}\leq\alpha\varepsilon$. The existence of N is guaranteed since $\lim_{t\to\infty}\phi(\delta,p,t)=0$. We have

$$\max_{b\in[n/2]}\left\{\phi(\delta,b/n-\varepsilon,N),\phi(\delta,1-b/n-\varepsilon,N)\right\}\leq\alpha\varepsilon, \tag{2.73}$$

$$\max_{b\in[n/2]}\left\{\phi(\delta,b/n-\varepsilon,N-1),\phi(\delta,1-b/n-\varepsilon,N-1)\right\}>\alpha\varepsilon. \tag{2.74}$$

Hence,

$$\min_{b\in[n/2]}\left\{\mathrm{KL}(b/n-\varepsilon+\alpha\varepsilon,b/n-\varepsilon),\mathrm{KL}(1-b/n-\varepsilon+\alpha\varepsilon,1-b/n-\varepsilon)\right\}\leq\frac{\log\left(\frac{2^{n-1}(N-1)N}{\delta}\right)}{N-1}. \tag{2.75}$$

Thus $\lim_{\delta\to 0}N=+\infty$ and from Lemma 2.4.2 we can deduce

$$N\leq\frac{\log(2^{n-1}/\delta)+4\log\left(\frac{\log(2^{n-1}/\delta)}{\min_{b\in[n/2]}\{\mathrm{KL}(b/n-\varepsilon+\alpha\varepsilon,b/n-\varepsilon),\mathrm{KL}(1-b/n-\varepsilon+\alpha\varepsilon,1-b/n-\varepsilon)\}}\right)}{\min_{b\in[n/2]}\{\mathrm{KL}(b/n-\varepsilon+\alpha\varepsilon,b/n-\varepsilon),\mathrm{KL}(1-b/n-\varepsilon+\alpha\varepsilon,1-b/n-\varepsilon)\}}+1\ . \tag{2.76}$$

Finally,

$$\limsup_{\delta\to 0}\frac{N}{\log(1/\delta)}\leq\frac{1}{\min_{b\in[n/2]}\left\{\mathrm{KL}(b/n-\varepsilon+\alpha\varepsilon,b/n-\varepsilon),\mathrm{KL}(1-b/n-\varepsilon+\alpha\varepsilon,1-b/n-\varepsilon)\right\}}. \tag{2.77}$$

Likewise, for $q=\mathcal{D}'(B_{opt})$ and $p=|B_{opt}|$ such that $|q-p|>\varepsilon$, we can define N_q the minimum positive integer such that for all $t\geq N_q$:

$$\max\left\{\phi(\delta,p,t),\phi(\delta,1-p,t)\right\}\leq\alpha|q-p|\ . \tag{2.78}$$

With the same analysis before, we can prove that

$$N_q\leq\frac{\log(2^{n-1}/\delta)+4\log\left(\frac{\log(2^{n_1}/\delta)}{\min\{\mathrm{KL}(p+\alpha|q-p\ \ ,p),\mathrm{KL}(1-p+\alpha|q-p|,1-p)\}}\right)}{\min\{\mathrm{KL}(p+\alpha|q-p|,p),\mathrm{KL}(1-p+\alpha|q-p|,1-p)\}}+1\ . \tag{2.79}$$

Finally,

$$\limsup_{\delta\to 0}\frac{N_q}{\log(1/\delta)}\leq\frac{1}{\min\{\mathrm{KL}(p+\alpha|q-p|,p),\mathrm{KL}(1-p+\alpha|q-p|,1-p)\}}\ . \tag{2.80}$$

Then we use Lemma 2.3.3 and make a case study on $\mathcal{D}'$:

- If $\mathcal{D}' = U_n$, the estimated stopping time can be bounded as

$$\mathbb{E}(\tau_1(U_n)) \le N + \sum_{t \ge N} \mathbb{P}(\tau_1(U_n) \ge t) \tag{2.81}$$

$$\le N + \sum_{s \ge N-1} \mathbb{P}(\exists B \subset [\lfloor n/2 \rfloor] : \tilde{\mathcal{D}}'_t(B) > |B|/n + \varepsilon - \phi \text{ or } \tilde{\mathcal{D}}'_t(B) < |B|/n - \varepsilon + \phi) \tag{2.82}$$

$$\le N + \sum_{B \subset [\lfloor n/2 \rfloor], s \ge N-1} \mathbb{P}(\tilde{\mathcal{D}}'_t(B) > |B|/n + \varepsilon - \alpha\varepsilon \text{ or } \tilde{\mathcal{D}}'_t(B) < |B|/n - \varepsilon + \alpha\varepsilon) \tag{2.83}$$

$$\le N + \sum_{B \subset [\lfloor n/2 \rfloor], s \ge N-1} \mathbb{P}(|\tilde{\mathcal{D}}'_t(B) - |B|/n| > (1-\alpha)\varepsilon) \tag{2.84}$$

$$\le N + \sum_{B \subset [\lfloor n/2 \rfloor], s \ge N-1} 2e^{-2s((1-\alpha)\varepsilon)^2} \tag{2.85}$$

$$\le N + \frac{2^{n/2+1} e^{-2(N-1)((1-\alpha)\varepsilon)^2}}{1 - e^{-2((1-\alpha)\varepsilon)^2}} \le N + \frac{2^{n/2+2} e^{-2(N-1)((1-\alpha)\varepsilon)^2}}{(1-\alpha)^2\varepsilon^2} , \tag{2.86}$$

  where we Chernoff-Hoeffding's inequality and the inequality $1 - e^{-x} \ge x/2$ for $0 < x < 1$ in the last line.

- If $q := \mathcal{D}'(B_{opt}) > |B_{opt}|/n + \varepsilon =: p + \varepsilon$, the estimated stopping time can be bounded as

$$\mathbb{E}(\tau_2) \le N_q + \sum_{t \ge N} \mathbb{P}(\tau_2(\mathcal{D}') \ge t) \tag{2.87}$$

$$\le N_q + \sum_{s \ge N-1} \mathbb{P}(|\tilde{\mathcal{D}}'_s(B_{opt}) - p| \le \max\{\phi(\delta, p, t), \phi(\delta, 1-p, t)\}) \tag{2.88}$$

$$\le N_q + \sum_{s \ge N-1} \mathbb{P}(\tilde{\mathcal{D}}'_s(B_{opt}) \le p + \max\{\phi(\delta, p, t), \phi(\delta, 1-p, t)\}) \tag{2.89}$$

$$\le N_q + \sum_{s \ge N-1} \mathbb{P}(\tilde{\mathcal{D}}'_s(B_{opt}) \le p + \alpha(q-p)) \text{ (by definition of } N_q) \tag{2.90}$$

$$\le N_q + \sum_{s \ge N-1} \mathbb{P}(\tilde{\mathcal{D}}'_s(B_{opt}) \le q - (1-\alpha)(q-p)) \tag{2.91}$$

$$\le N_q + \sum_{s \ge N-1} e^{-2s((1-\alpha)(q-p))^2} \text{ (Chernoff-Hoeffding's inequality)} \tag{2.92}$$

$$\le N_q + \frac{e^{-2(N-1)((1-\alpha)(q-p))^2}}{1 - e^{-2((1-\alpha)(q-p))^2}} \tag{2.93}$$

$$\le N_q + \frac{1}{(1-\alpha)^2(q-p)^2} . \tag{2.94}$$

- If $q := \mathcal{D}'(B_{opt}) < |B_{opt}|/n - \varepsilon =: p - \varepsilon$, the estimated stopping time can be bounded

as

$$\mathbb{E}(\tau_2(\mathcal{D}')) \leq N_q + \sum_{t\geq N} \mathbb{P}(\tau_2(\mathcal{D}') \geq t) \tag{2.95}$$

$$\leq N_q + \sum_{s\geq N-1} \mathbb{P}(|\tilde{\mathcal{D}}'_s(B_{opt}) - p| \leq \max\{\phi(\delta,p,t), \phi(\delta,1-p,t)\}) \tag{2.96}$$

$$\leq N_q + \sum_{s\geq N-1} \mathbb{P}(\tilde{\mathcal{D}}'_s(B_{opt}) \geq p - \max\{\phi(\delta,p,t), \phi(\delta,1-p,t)\}) \tag{2.97}$$

$$\leq N_q + \sum_{s\geq N-1} \mathbb{P}(\tilde{\mathcal{D}}'_s(B_{opt}) \geq p - \alpha(p-q)) \text{ (by definition of } N_q) \tag{2.98}$$

$$\leq N_q + \sum_{s\geq N-1} \mathbb{P}(\tilde{\mathcal{D}}'_s(B_{opt}) \geq q + (1-\alpha)(p-q)) \tag{2.99}$$

$$\leq N_q + \sum_{s\geq N-1} e^{-2s((1-\alpha)(q-p))^2} \text{ (Chernoff-Hoeffding's inequality)} \tag{2.100}$$

$$\leq N_q + \frac{e^{-2(N-1)((1-\alpha)(q-p))^2}}{1-e^{-2((1-\alpha)(q-p))^2}} \tag{2.101}$$

$$\leq N_q + \frac{1}{(1-\alpha)^2(q-p)^2} \,. \tag{2.102}$$

Dividing by $\log(1/\delta)$, taking the limits $\delta \to 0$ then $\alpha \to 1$ permit to deduce the asymptotic bounds. By choosing $\alpha = (1+\log(2^{n-1}/\delta)^{-1/3})^{-1}$, we conclude for $n < \log^{2/3}(1/\delta)$:

$$\mathbb{E}(\tau_1(U_n)) \leq \frac{\log(2^{n-1}/\delta)}{\min_{b\in[n]}\{\mathrm{KL}(b/n, b/n \pm \varepsilon)\}} + \mathcal{O}\left(\frac{\log(2^{n-1}/\delta)^{2/3}}{\varepsilon^2}\right), \text{ and} \tag{2.103}$$

$$\mathbb{E}(\tau_2(\mathcal{D}')) \leq \frac{\log(2^{n-1}/\delta)}{\min\{\mathrm{KL}(|B_{opt}|/n \pm d, |B_{opt}|/n)\}} + \mathcal{O}\left(\frac{\log(2^{n-1}/\delta)^{2/3}}{d^2}\right) \tag{2.104}$$

if $d = \mathrm{TV}(\mathcal{D}', U_n) = |\mathcal{D}'(B_{opt}) - |B_{opt}|/n| > \varepsilon$. □

2.4 Testing closeness for small n

In this section, $n \geq 2$ is still small and we consider this time two unknown distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ on $[n]$. We are testing two hypothesis H_1: $\mathcal{D}_1 = \mathcal{D}_2$ and H_2: $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$. Similar to the previous section, we are interested in precisely comparing the sample complexity of testing closeness in the sequential versus the batch setting. In order to find the optimal constant, we first need to obtain a sharp lower bound in the batch setting, which is done directly by using Stirling's approximation. We then turn to the sequential case.

2.4.1 Batch setting

In the batch setting, the number of steps τ is fixed before the test. The tester samples $A_1, \ldots, A_\tau \sim \mathcal{D}_1$ and $B_1, \ldots, B_\tau \sim \mathcal{D}_2$ then decides according to the comparison between the empirical TV distance $\mathrm{TV}(\tilde{\mathcal{D}}_{1\tau}, \tilde{\mathcal{D}}_{2\tau})$ and $\varepsilon/2$ where $\tilde{\mathcal{D}}_{1\tau} = \left\{\left(\sum_{j=1}^{\tau} \mathbf{1}_{A_j=i}\right)/\tau\right\}_{i\in[n]}$ and $\tilde{\mathcal{D}}_{2\tau} = \left\{\left(\sum_{j=1}^{\tau} \mathbf{1}_{B_j=i}\right)/\tau\right\}_{i\in[n]}$ are the empirical distributions. If $\mathrm{TV}(\tilde{\mathcal{D}}_{1\tau}, \tilde{\mathcal{D}}_{2\tau}) \leq$

$\varepsilon/2$, it accepts H_1 and rejects it otherwise. In order to control the number of steps τ so that the error of this algorithm does not exceed δ, McDiarmid's inequality ([HMRAR13]) writes for $\tau = \frac{4\log(2^{\lfloor n/2\rfloor}/\delta)}{\varepsilon^2}$:

$$\mathbb{P}\left(\exists B \subset [\lfloor n/2\rfloor] : \left|\tilde{\mathcal{D}}_{1,\tau}(B) - \mathcal{D}_1(B) - \tilde{\mathcal{D}}_{2,\tau}(B) + \mathcal{D}_2(B)\right| > \frac{\varepsilon}{2}\right) \leq \sum_{B\subset[\lfloor n/2\rfloor]} e^{-\tau\varepsilon^2/4} \leq \delta \,. \tag{M}$$

Using the concentration inequality (M) for $\mathcal{D}_1 = \mathcal{D}_2$ (to control the type I error) and for $\mathcal{D}_1 \neq \mathcal{D}_2$ (to control the type II error) we prove that this test is δ-correct. We show in the following theorem that this number of steps τ is necessary.

Proposition 2.4.1. *In the batch setting, the algorithm consisting of accepting H_1 when $\mathrm{TV}(\tilde{\mathcal{D}}_{1\tau}, \tilde{\mathcal{D}}_{2\tau}) \leq \varepsilon/2$ and rejecting it otherwise is δ-correct for $\tau = \frac{4\log(2^{\lfloor n/2\rfloor}/\delta)}{\varepsilon^2}$.*

Moreover, any δ-correct algorithm testing closeness requires at least τ samples, where

$$\tau \geq \min\left\{\frac{\log(1/2\delta)}{2\,\mathrm{KL}(1/2-\varepsilon/4, 1/2-\varepsilon/2)}, \frac{\log(1/2\delta)}{2\,\mathrm{KL}(1/2+\varepsilon/4, 1/2)}\right\} - \mathcal{O}\left(\frac{\log\log(1/\delta)}{\varepsilon^2}\right) . \tag{2.105}$$

For this proof, we show that every δ-correct tester can be transformed into a test which depends only on the numbers of $1's, 2's, \ldots, n's$ occurred on $\{A_1, \ldots, A_\tau\}$ and $\{B_1, \ldots, B_\tau\}$. We then consider the distributions $\mathcal{D}_{1,2} = \{1/2, 1/2, 0, \ldots, 0\}$ or $\mathcal{D}_{1,2} = \{1/2 \pm \varepsilon/2, 1/2 \mp \varepsilon/2, 0, \ldots, 0\}$ depending on the outcome of the algorithm when it sees two words of samples having respectively $\tau(1/2-\varepsilon/4)$ and $\tau(1/2+\varepsilon/4)$ ones (the rest of samples are equal to 2) and derive tight lower bounds on the probability mass function of the multinomial distribution.

Proof. We consider distributions supported only on $\{1, 2\}$, this is possible since we want that our algorithm would work for all distributions. We consider such a δ-correct test $A : \{1,2\}^\tau \times \{1,2\}^\tau \to \{1,2\}$, it sees two words consisting of τ samples either from equal distributions or ε-far ones and returns 1 if it thinks they are equal and 2 otherwise. We construct another test $B : \{1,2\}^\tau \times \{1,2\}^\tau \to \{0,1\}$ by the expression

$$B(x,y) = \mathbf{1}\left\{\sum_{\sigma,\rho\in\mathfrak{S}_\tau} A(\sigma(x), \rho(y)) - 1 \geq (\tau!)^2/2\right\} , \tag{2.106}$$

B can be proven to be 2δ-correct and have the property of invariance under the action of the symmetric group. This leads to an algorithm $C : \{0, \ldots, \tau\}^2 \to \{0,1\}$ which is 2δ correct and satisfies

$$C(i,j) = B(x_i, y_j) , \tag{2.107}$$

where $x_k = 1 \ldots 12 \ldots 2$ with k ones. We consider $i = [\tau(1/2-\varepsilon/4)]$ and $j = [\tau(1/2+\varepsilon/4)]$. We denote by $N_i(x)$ the number of i in a word x of length τ for $i = 1, 2$.

- If $C(i,j) = 0$, let x (resp. y) a word of length τ constituted of i.i.d samples from $\{1/2-\varepsilon/2, 1/2+\varepsilon/2, 0, \ldots, 0\}$ (resp. $\{1/2+\varepsilon/2, 1/2-\varepsilon/2, 0, \ldots, 0\}$), then

$\mathbb{P}_{1/2-\varepsilon/2,1/2+\varepsilon/2}(N_1(x) = i, N_1(y) = j) \leq 2\delta$ hence with Stirling's approximation ([Leu85])

$$\frac{e^{-2}}{2\pi\tau} e^{-\tau \operatorname{KL}(i/\tau,1/2-\varepsilon/2)} e^{-\tau \operatorname{KL}(1-j/\tau,1/2-\varepsilon/2)} \leq 2\delta. \tag{2.108}$$

Thus

$$2\tau \operatorname{KL}(1/2+\varepsilon/4-1/\tau, 1/2+\varepsilon/2) \geq \tau(\operatorname{KL}(i/\tau, 1/2-\varepsilon/2) + \operatorname{KL}(j/\tau, 1/2-\varepsilon/2)) \tag{2.109}$$

$$\geq \log(1/2\delta) - 2 - \log(2\pi) - \log(\tau)\ . \tag{2.110}$$

On the other hand, we have the following properties of the KL divergence:

Lemma 2.4.1 (Lemmas for KL-divergence)**.** *Let $q > p$ two numbers in $[0,1]$. Then*

- $2(p-q)^2 \leq \operatorname{KL}(p,q) \leq \frac{(p-q)^2}{q(1-q)}$,
- $\operatorname{KL}(p,q) \underset{q\to p}{\sim} \frac{(p-q)^2}{2q(1-q)}$,
- $\operatorname{KL}(q,p) = \int_p^q du \int_p^u dv \frac{1}{v(1-v)}$.

Sketch of the proof. The LHS of the first inequality is Pinsker's inequality, the RHS can be proven using the inequality $\log(1+x) \leq x$, the second equivalence can be found by developing the log function and the third equality is proven by computing the integral. □

Hence using Lemma 2.4.1, we have for $\tau > 2/\varepsilon$:

$$2\tau \operatorname{KL}(1/2+\varepsilon/4, 1/2+\varepsilon/2) \tag{2.111}$$

$$\geq -2\tau(\operatorname{KL}(1/2+\varepsilon/4-1/\tau, 1/2+\varepsilon/2) - \operatorname{KL}(1/2+\varepsilon/4, 1/2+\varepsilon/2)) \tag{2.112}$$

$$+ \log(1/2\delta) - 2 - \log(2\pi) - \log(\tau) \tag{2.113}$$

$$\geq -2\tau \int_{1/2+\varepsilon/4-1/\tau}^{1/2+\varepsilon/4} du \int_u^{1/2+\varepsilon/2} dv \frac{1}{v(1-v)} + \log(1/2\delta) \tag{2.114}$$

$$- 2 - \log(2\pi) - \log(\tau) \tag{2.115}$$

$$\geq -2(\varepsilon/4+1/\tau) \sup_{[1/2+\varepsilon/4-1/\tau, 1/2+\varepsilon/2]} \frac{1}{v(1-v)} + \log(1/2\delta) \tag{2.116}$$

$$- 2 - \log(2\pi) - \log(\tau) \tag{2.117}$$

$$\geq -2\varepsilon \sup_{[1/2,1/2+\varepsilon]} \frac{1}{v(1-v)} + \log(1/2\delta) - 2 - \log(2\pi) - \log(\tau)\ . \tag{2.118}$$

When we deal with inequalities involving τ and $\log\tau$ (or $\log\log\tau$) and want to deduce inequalities only on τ, the following lemma proves to be useful.

Lemma 2.4.2. *Let $t, a > 1$ and b real numbers. We have the following implications:*

- *If $b \geq a+1$:*

$$t \geq b + 2a\log(b) \Rightarrow t \geq b + a\log(t)\ , \tag{2.119}$$

– *If* $b \geq 1$:

$$t \geq b - a\log(t) \Rightarrow t \geq b - a\log(b)\ , \tag{2.120}$$

– *If* $b \geq 2a$:

$$t \geq b + 2a\log(\log(b)+1) \Rightarrow t \geq b + a\log(\log(t)+1)\ . \tag{2.121}$$

Proof. We prove only the first statement, the others being similar. Let $f(t) = t - b - a\log(t)$, we have $f'(t) = 1 - a/t$ thus f is increasing on $(a, +\infty)$. Let $t \geq b + 2a\log(b) > a$,

$$\begin{aligned} f(t) \geq f(b+2a\log(b)) &= b + 2a\log(b) - b - a\log(b + 2a\log(b)) && (2.122)\\ &= a\log(b) - a\log(1 + 2a\log(b)/b)) && (2.123)\\ &\geq a\log(1+a) - a\log(1+2ab/eb) \quad \text{because } \log(b) \leq b/e && (2.124)\\ &\geq 0\ . && (2.125) \end{aligned}$$

□

Then Lemma 2.4.2 implies:

$$\begin{aligned} \tau &\geq \frac{\log(1/2\delta) - 2\varepsilon \sup_{[1/2,1/2+\varepsilon]} \frac{1}{v(1-v)} - 2 - \log(2\pi)}{2\,\mathrm{KL}(1/2+\varepsilon/4, 1/2+\varepsilon/2)} && (2.126)\\ &\quad - \frac{\log\left(\frac{-2\varepsilon \sup_{[1/2,1/2+\varepsilon]} \frac{1}{v(1-v)} + \log(1/2\delta) - 2 - \log(2\pi)}{2\,\mathrm{KL}(1/2+\varepsilon/4, 1/2+\varepsilon/2)}\right)}{4\,\mathrm{KL}(1/2+\varepsilon/4, 1/2+\varepsilon/2)} && (2.127)\\ &\geq \frac{\log(1/2\delta)}{2\,\mathrm{KL}(1/2+\varepsilon/4, 1/2+\varepsilon/2)} - \mathcal{O}\left(\frac{\log\log(1/\delta)}{\mathrm{KL}(1/2+\varepsilon/4, 1/2+\varepsilon/2)}\right). && (2.128) \end{aligned}$$

Finally we get the asymptotic lower bound:

$$\liminf_{\delta\to 0} \frac{\tau}{\log(1/\delta)} \geq \frac{1}{2\,\mathrm{KL}(1/2-\varepsilon/4, 1/2-\varepsilon/2)}. \tag{2.129}$$

- If $C(i,j) = 1$, let x and y two words of length τ constituted of i.i.d samples from $\{1/2, 1/2, 0, \ldots, 0\}$, then $\mathbb{P}_{1/2,1/2}(N_1(x) = i, N_1(y) = j) \leq 2\delta$ hence with Stirling's approximation

$$\frac{e^{-2}}{2\pi\tau} e^{-\tau\,\mathrm{KL}(i/\tau,1/2)} e^{-\tau\,\mathrm{KL}(1-j/\tau,1/2)} \leq 2\delta\ . \tag{2.130}$$

Using the same lemmas as before, we get the following lower bound

$$\tau \geq \frac{\log(1/2\delta)}{2\,\mathrm{KL}(1/2+\varepsilon/4, 1/2)} - \mathcal{O}\left(\frac{\log\log(1/\delta)}{\mathrm{KL}(1/2+\varepsilon/4, 1/2)}\right). \tag{2.131}$$

Finally we get the asymptotic lower bound:

$$\liminf_{\delta\to 0} \frac{\tau}{\log(1/\delta)} \geq \frac{1}{2\,\mathrm{KL}(1/2+\varepsilon/4, 1/2)}\ . \tag{2.132}$$

□

2.4.2 Sequential setting

Inspired by testing identity for small alphabets results in Section 2.3, we would like to achieve an improvement of factor 4 in sample complexity for sequential strategies over batch ones for testing closeness problem. For this end, we start by stating a lower bound on the expected stopping times of a sequential algorithm for testing closeness.

Lemma 2.4.3. *Let T be a stopping rule for testing closeness: $\mathcal{D}_1 = \mathcal{D}_2$ vs $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$ with an error probability δ. Let τ_1 and τ_2 the associated stopping times. We have*

$$\sup_{\mathcal{D}} \mathbb{E}(\tau_1(\mathcal{D}, \mathcal{D})) \geq \frac{\log(1/3\delta)}{\mathrm{KL}(1/2, 1/2 + \varepsilon/2) + \mathrm{KL}(1/2, 1/2 - \varepsilon/2)} \underset{\varepsilon \to 0}{\sim} \frac{\log(1/3\delta)}{\varepsilon^2} \text{ and} \tag{2.133}$$

$$\sup_{\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) = d} \mathbb{E}(\tau_2(\mathcal{D}_1, \mathcal{D}_2)) \geq \frac{\log(1/3\delta)}{\mathrm{KL}(1/2 + d/2, 1/2) + \mathrm{KL}(1/2 - d/2, 1/2)} \underset{d \to 0}{\sim} \frac{\log(1/3\delta)}{d^2} \text{ if } d > \varepsilon\,. \tag{2.134}$$

An average number of samples equivalent to $\log(1/3\delta)(\varepsilon \vee \mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2))^{-2}$ (see Lemma 2.4.1) is thus necessary when the tester can access sequentially to the samples, which is roughly 4 times less than the complexity obtained in the batch setting.

Proof. The proof of this Lemma follows from Lemma 2.7.2 by choosing for the first point $\mathcal{D}_1 = \mathcal{D}_2 = \{1/2, 1/2, 0, \ldots, 0\}$ and $\mathcal{D}'_{1,2} = \{1/2 \pm \varepsilon/2, 1/2 \mp \varepsilon/2, 0, \ldots, 0\}$. For the second point, we use $\mathcal{D} = \{1/2, 1/2, 0, \ldots, 0\}$ and $\mathcal{D}_{1,2} = \{1/2 \pm d/2, 1/2 \mp d/2, 0, \ldots, 0\}$. □

In the sequential testing, the tester chooses when to stop according to the previous observations $((A_1, B_1), \ldots, (A_t, B_t))$, making comparisons at each step t. The tester can stop as soon as it is sure that it can accept one of the hypothesis H_1 or H_2. On the contrary, in the batch setting it had to sample enough observation to be simultaneously sure that either H_1 or H_2 hold. In this aim, at each time step, after sampling a new observation (A_t, B_t), it compares the updated empirical TV distance $S_t = \mathrm{TV}(\tilde{\mathcal{D}}_{1t}, \tilde{\mathcal{D}}_{2t})$ to specific thresholds and sees if (a) S_t is sufficiently far from 0 to surely accept H_2, (b) S_t is sufficiently close to ε to surely accept H_1, (c) it is unsure and needs further samples to take a sound decision. This test is formally described in Algorithm 2 and its execution is illustrated in Figure 2.1 for $n = 2$.

To show the correctness of such sequential algorithms, Chernoff-Hoeffding's inequality does not work since we have two unknwon distributions. It turns out that McDiarmid's inequality (M) is best suited in this situation:

$$\mathbb{P}\left(\exists t \geq 1, \exists B \subset [n/2] : \left|\tilde{\mathcal{D}}_{1,t}(B) - \mathcal{D}_1(B) - \tilde{\mathcal{D}}_{2,t}(B) + \mathcal{D}_2(B)\right| > \Phi_t\right) \leq \delta, \tag{2.135}$$

where Φ_t denote the constant $\Phi_t = \sqrt{\log\left(\frac{2^{n-1}t(t+1)}{\delta}\right)/t}$. On the other hand, to control the sample complexity, we prove upper bounds on the expected stopping times:

$$\tau_1 = \inf\left\{t \geq 1 : \mathrm{TV}\left(\tilde{\mathcal{D}}_{1,t}, \tilde{\mathcal{D}}_{2,t}\right) \leq \varepsilon - \Phi_t\right\}, \text{ and } \tau_2 = \inf\left\{t \geq 1 : \mathrm{TV}\left(\tilde{\mathcal{D}}_{1,t}, \tilde{\mathcal{D}}_{2,t}\right) > \Phi_t\right\}. \tag{2.136}$$

It is clear that the stopping time of the algorithm is random. Yet, we can show that this algorithm stops before the non sequential one and give an upper bound on the expected

Algorithm 2 Distinguish between $\mathcal{D}_1 = \mathcal{D}_2$ and $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$ with high probability

Require: $A_1, \dots$ samples from $\mathcal{D}_1$ and $B_1, \dots$ samples from $\mathcal{D}_2$
Ensure: Accept if $\mathcal{D}_1 = \mathcal{D}_2$ and Reject if $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$ with probability of error less than δ
$t = 1$, $W = 1$
while $W = 1$ **do**
$\tilde{\mathcal{D}}_{1,t} = \left\{ \left(\sum_{j=1}^{t} \mathbf{1}_{A_j = i} \right) / t \right\}_{i \in [n]}$, $\tilde{\mathcal{D}}_{2,t} = \left\{ \left(\sum_{j=1}^{t} \mathbf{1}_{B_j = i} \right) / t \right\}_{i \in [n]}$
if $\mathrm{TV}\left(\tilde{\mathcal{D}}_{1,t}, \tilde{\mathcal{D}}_{2,t} \right) > \sqrt{\frac{\log\left(\frac{2^{n-1} t(t+1)}{\delta}\right)}{t}}$ **then**
$W = 0$
return 2
else if $\mathrm{TV}\left(\tilde{\mathcal{D}}_{1,t}, \tilde{\mathcal{D}}_{2,t} \right) \leq \varepsilon - \sqrt{\frac{\log\left(\frac{2^{n-1} t(t+1)}{\delta}\right)}{t}}$ **then**
$W = 0$
return 1
else
$t = t + 1$
end if
end while

stopping time τ (or expected sample complexity). In the following theorem, we state an upper bound on the estimated sample complexity of this algorithm.

Theorem 2.4.1. *The Algorithm 2 is δ-correct and its stopping times verify for $n \leq \mathcal{O}(\log(1/\delta)^{1/3})$:*

$$\mathbb{E}(\tau_1(\mathcal{D}, \mathcal{D})) \leq \frac{\log(2^{n+1}/\delta)}{\varepsilon^2} + \mathcal{O}\left(\frac{\log(2^{n+1}/\delta)^{2/3}}{\varepsilon^2} \right) \text{ if } \mathcal{D}_1 = \mathcal{D}_2 = \mathcal{D} \text{ and} \tag{2.137}$$

$$\mathbb{E}(\tau_2(\mathcal{D}_1, \mathcal{D}_2)) \leq \frac{\log(2^{n+1}/\delta)}{\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2)^2} + \mathcal{O}\left(\frac{\log(2^{n+1}/\delta)^{2/3}}{\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2)^2} \right) \text{ if } \mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon\,. \tag{2.138}$$

These upper bounds are tight in the sense that they match the asymptotic lower bounds of Lemma 2.4.3 if $n \ll \log(1/\delta)$. The advantages of sequential strategies over batch ones for the testing closeness problem are the same as those for the testing identity problem.

Proof. We should prove that the Algorithm 2 has an error probability less than δ. We use the following lemma which can be proven using McDiarmid's inequality and union bounds.

Lemma 2.4.4. *If $\{A_1, \dots, A_t\}$ (resp $\{B_1, \dots, B_t\}$) i.i.d. with the law $\mathcal{D}_1$ (resp $\mathcal{D}_2$), we have the following inequality*

$$\mathbb{P}\left(\exists t \geq 1, \exists B \subset [n/2] : \left| \tilde{\mathcal{D}}_{1,t}(B) - \mathcal{D}_1(B) - \tilde{\mathcal{D}}_{2,t}(B) + \mathcal{D}_2(B) \right| > \sqrt{\log\left(\frac{2^{n-1} t(t+1)}{\delta} \right) / t} \right) \leq \delta.$$

Using this lemma we can conclude:

- If $\mathcal{D}_1 = \mathcal{D}_2$, the probability of error is given by

$$\mathbb{P}\left(\tau_2 \leq \tau_1\right) \leq \mathbb{P}\left(\exists t \geq 1 : \mathrm{TV}\left(\tilde{\mathcal{D}}_{1,t}, \tilde{\mathcal{D}}_{2,t}\right) > \sqrt{\log\left(\frac{2^{n-1}t(t+1)}{\delta}\right)/t}\right) \leq \delta \, .$$

- If $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) = |\mathcal{D}_1(B_{opt}) - \mathcal{D}_2(B_{opt})| > \varepsilon$, the probability of error is given by

$$\mathbb{P}\left(\tau_1 \leq \tau_2\right) \leq \mathbb{P}\left(\exists t \geq 1 : \mathrm{TV}\left(\tilde{\mathcal{D}}_{1,t}, \tilde{\mathcal{D}}_{2,t}\right) \leq \varepsilon - \sqrt{\log\left(\frac{2^{n-1}t(t+1)}{\delta}\right)/t}\right) \tag{2.139}$$

$$\leq \mathbb{P}\left(\exists t \geq 1 : \left|\tilde{\mathcal{D}}_{1,t}(B_{opt}) - \tilde{\mathcal{D}}_{2,t}(B_{opt}))\right| \leq \varepsilon - \sqrt{\log\left(\frac{2^{n-1}t(t+1)}{\delta}\right)/t}\right) \tag{2.140}$$

$$\leq \mathbb{P}\Bigg(\exists t \geq 1 : \left|\tilde{\mathcal{D}}_{1,t}(B_{opt}) - \mathcal{D}_1(B_{opt}) - \tilde{\mathcal{D}}_{2,t}(B_{opt}) + \mathcal{D}_2(B_{opt}))\right| \tag{2.141}$$

$$\geq |\mathcal{D}_1(B_{opt}) - \mathcal{D}_2(B_{opt})| - \varepsilon + \sqrt{\log\left(\frac{2^{n-1}t(t+1)}{\delta}\right)/t}\Bigg) \tag{2.142}$$

$$\leq \mathbb{P}\left(\exists t \geq 1 : \left|\tilde{\mathcal{D}}_{1,t}(B_{opt}) - \mathcal{D}_1(B_{opt}) - \tilde{\mathcal{D}}_{2,t}(B_{opt}) + \mathcal{D}_2(B_{opt}))\right| > \sqrt{\log\left(\frac{2^{n-1}t(t+1)}{\delta}\right)/t}\right) \tag{2.143}$$

$$\leq \delta \, . \tag{2.144}$$

These computations prove the correctness of Algorithm 2. It remains to study the complexity of Algorithm 2. To this aim, we make a case study and use Lemma 2.3.3 to upper bound the stopping rules.

Let us take $\alpha \in (0,1)$,

- If $\mathcal{D}_1 = \mathcal{D}_2$, we take $N = \left\lceil \frac{\log(2^{n+1}/\delta)}{(\alpha\varepsilon)^2} \right\rceil + 1$ and $\tilde{\alpha} \in (0,1)$[2] so that

$$\tilde{\alpha}^2 = \alpha^2 \left(\frac{\log\log(2^{n+1}/\delta) - \log((\alpha\varepsilon)^2)}{\log(2^{n+1}/\delta)} + 1\right).$$

[2] for fixed α we take δ small enough to have $\tilde{\alpha} < 1$.

The estimated stopping time can be bounded as

$$\mathbb{E}(\tau_1(\mathcal{D}_1,\mathcal{D}_2)) \leq N + \sum_{s\geq N} \mathbb{P}(\tau_1(\mathcal{D}_1,\mathcal{D}_2)\geq s) \tag{2.145}$$

$$\leq N + \sum_{t\geq N-1} \mathbb{P}\left(\mathrm{TV}\left(\tilde{\mathcal{D}}_{1,t},\tilde{\mathcal{D}}_{2,t}\right) > \varepsilon - \sqrt{\log\left(\frac{2^{n-1}t(t+1)}{\delta}\right)/t}\right) \tag{2.146}$$

$$\leq N + \sum_{t\geq N-1} \mathbb{P}\left(\mathrm{TV}\left(\tilde{\mathcal{D}}_{1,t},\tilde{\mathcal{D}}_{2,t}\right) > \varepsilon - \tilde{\alpha}\varepsilon\right) \tag{2.147}$$

$$\leq N + \sum_{t\geq N-1} \mathbb{P}\left(\mathrm{TV}\left(\tilde{\mathcal{D}}_{1,t},\tilde{\mathcal{D}}_{2,t}\right) > (1-\tilde{\alpha})\varepsilon\right) \tag{2.148}$$

$$\overset{(a)}{\leq} N + \sum_{t\geq N-1} 2^{n/2} e^{-t((1-\tilde{\alpha})\varepsilon)^2} \tag{2.149}$$

$$\leq N + \frac{2^{n/2}e^{-(N-1)((1-\tilde{\alpha})\varepsilon)^2}}{1-e^{-((1-\tilde{\alpha})\varepsilon)^2}} \tag{2.150}$$

where in (a) we used McDiarmid's inequality. Using the inequality ($1-e^{-x}\geq x/2$ for $0<x<1$) we deduce:

$$\mathbb{E}(\tau_1(\mathcal{D}_1,\mathcal{D}_2)) \leq N + \frac{2^{n/2}e^{-(N-1)((1-\tilde{\alpha})\varepsilon)^2}}{1-e^{-((1-\tilde{\alpha})\varepsilon)^2}} \tag{2.151}$$

$$\leq \frac{\log(2^{n+1}/\delta)}{(\alpha\varepsilon)^2} + 2\frac{2^{n/2}e^{-(N-1)((1-\tilde{\alpha})\varepsilon)^2}}{((1-\tilde{\alpha})\varepsilon)^2} + 1\,, \tag{2.152}$$

$$\leq \frac{\log(2^{n+1}/\delta)}{\varepsilon^2} + \frac{\log(2^{n+1}/\delta)^{2/3}}{\varepsilon^2} + \mathcal{O}\left(\frac{\log(2^{n+1}/\delta)^{2/3}}{\varepsilon^2}\right) \tag{2.153}$$

$$\leq \frac{\log(2^{n+1}/\delta)}{\varepsilon^2} + \mathcal{O}\left(\frac{\log(2^{n+1}/\delta)^{2/3}}{\varepsilon^2}\right)\,, \tag{2.154}$$

for $\alpha = (1+\log(2^{n+1}/\delta)^{-1/3})^{-2}$ so that $1-\tilde{\alpha}\geq C\log(2^{n+1}/\delta)^{-1/3}$ and we suppose here that $n < 2C^2\log(2^{n+1}/\delta)^{1/3}$.

- If $d = \mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2) = |\mathcal{D}_1(B_{opt}) - \mathcal{D}_2(B_{opt})| > \varepsilon$, we take $N = \left\lceil \frac{\log(2^{n+1}/\delta)}{(\alpha d)^2}\right\rceil + 1$. We take $\tilde{\alpha}\in(0,1)$ so that $\tilde{\alpha}^2 = \alpha^2\left(\frac{\log\log(2^{n+1}/\delta)-\log((\alpha d)^2)}{\log(2^{n+1}/\delta)}+1\right)$. The estimated stopping

time can be bounded as

$$\mathbb{E}(\tau_2(\mathcal{D}_1,\mathcal{D}_2)) \leq N + \sum_{s\geq N} \mathbb{P}(\tau_2(\mathcal{D}_1,\mathcal{D}_2)\geq s) \tag{2.155}$$

$$\leq N + \sum_{t\geq N-1} \mathbb{P}\left(\mathrm{TV}\left(\tilde{\mathcal{D}}_{1,t},\tilde{\mathcal{D}}_{2,t}\right) \leq \sqrt{\log\left(\frac{2^{n-1}t(t+1)}{\delta}\right)/t}\right) \tag{2.156}$$

$$\leq N + \sum_{t\geq N-1} \mathbb{P}\left(\mathrm{TV}\left(\tilde{\mathcal{D}}_{1,t},\tilde{\mathcal{D}}_{2,t}\right) \leq \sqrt{\log\left(\frac{2^{n-1}t(t+1)}{\delta}\right)/t}\right) \tag{2.157}$$

$$\leq N + \sum_{t\geq N-1} \mathbb{P}\left(\left|\tilde{\mathcal{D}}_{1,t}(B_{opt}) - \tilde{\mathcal{D}}_{2,t}(B_{opt})\right| \leq \sqrt{\log\left(\frac{2^{n-1}t(t+1)}{\delta}\right)/t}\right) \tag{2.158}$$

$$\leq N + \sum_{t\geq N-1} \mathbb{P}\Bigg(\left|\tilde{\mathcal{D}}_{1,t}(B_{opt}) - \mathcal{D}_1(B_{opt}) - \tilde{\mathcal{D}}_{2,t}(B_{opt}) + \mathcal{D}_2(B_{opt})\right| \tag{2.159}$$

$$> |\mathcal{D}_1(B_{opt}) - \mathcal{D}_2(B_{opt})| - \sqrt{\log\left(\frac{2^{n-1}t(t+1)}{\delta}\right)/t}\Bigg) \tag{2.160}$$

$$\leq N + \sum_{t\geq N-1} \mathbb{P}\left(\left|\tilde{\mathcal{D}}_{1,t}(B_{opt}) - \mathcal{D}_1(B_{opt}) - \tilde{\mathcal{D}}_{2,t}(B_{opt}) + \mathcal{D}_2(B_{opt})\right| > (1-\tilde{\alpha})d\right) \tag{2.161}$$

$$\leq N + \sum_{t\geq N-1} e^{-t((1-\tilde{\alpha})d)^2} \leq N + \frac{e^{-(N-1)((1-\tilde{\alpha})d)^2}}{1-e^{-((1-\tilde{\alpha})d)^2}} \tag{2.162}$$

$$\leq \frac{\log(2^{n+1}/\delta)}{(\alpha d)^2} + \frac{2}{(1-\tilde{\alpha})^2 d^2} + 1 \;\leq \frac{\log(2^{n+1}/\delta)}{d^2} + \mathcal{O}\left(\frac{\log(2^{n+1}/\delta)^{2/3}}{d^2}\right), \tag{2.163}$$

where we choose $\alpha = (1+\log(2^{n+1}/\delta)^{-1/3})^{-2}$ and we use the inequality $1-e^{-x} \geq x/2$ for $0 < x < 1$ in the last line.

Finally, we can deduce the limit when $\mathcal{D}_1 = \mathcal{D}_2$:

$$\limsup_{\delta\to 0} \frac{\mathbb{E}(\tau_1(\mathcal{D}_1,\mathcal{D}_2))}{\log(1/\delta)} \leq \limsup_{\delta\to 0} \frac{\log(2^{n+1}/\delta)}{\log(1/\delta)\varepsilon^2} + \mathcal{O}\left(\frac{\log(2^{n+1}/\delta)^{2/3}}{\log(1/\delta)\varepsilon^2}\right) \leq \frac{1}{\varepsilon^2}, \tag{2.164}$$

and when $d = \mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2) > \varepsilon$:

$$\limsup_{\delta\to 0} \frac{\mathbb{E}(\tau_2(\mathcal{D}_1,\mathcal{D}_2))}{\log(1/\delta)} \leq \limsup_{\delta\to 0} \frac{\log(2^{n+1}/\delta)}{\log(1/\delta)d^2} + \mathcal{O}\left(\frac{\log(2^{n+1}/\delta)^{2/3}}{\log(1/\delta)d^2}\right) \leq \frac{1}{d^2}. \tag{2.165}$$

This concludes the proof of the complexity of Algorithm 2. □

After dealing with small alphabets and understanding well the differences between sequential and batch strategies for testing identity/sequential problems, one could ask whether sequential strategies have different behaviour for general alphabets when n is greater than $\log(1/\delta)$. In the following, we try to transform batch algorithms to sequential ones in order to adapt to the actual TV distance.

2.5 Uniformity testing-the general case

[DGPP17] propose an algorithm for uniformity testing in the general case. The advantage of their algorithm is that it has the tight sample complexity (in batch setting) depending not only on n but also on the error probability δ. We capture the main ingredient of this article as a lemma:

Lemma 2.5.1. *Let $\mathcal{D}$ a distribution on $[n]$ such that $d = \mathrm{TV}(\mathcal{D}, U_n) > \varepsilon$. There is a universal constant C such that for all $t \geq 1$:*

$$\mathbb{E}(\mathrm{TV}(\tilde{\mathcal{D}}_t, U_n)) \geq \mu_t(U_n) + C \min\left(\frac{d^2t^2}{n^2}, d^2\sqrt{\frac{t}{n}}, d\right), \tag{2.166}$$

where $\mu_t(U_n) = \mathbb{E}(\mathrm{TV}(\tilde{U}_{n,t}, U_n)) = \frac{1}{2}\mathbb{E}_{X_1,\dots,X_t\sim U_n} \sum_{i=1}^n |\frac{1}{t}\sum_{j=1}^t \mathbf{1}_{X_j=i} - \frac{1}{n}|$ can be computed in $\mathcal{O}(t)$.

This Lemma (Lemma 4 from [DGPP17]) is used along with the ideas of Section 2.3 to design a sequential algorithm whose complexity is on the worst case of the order of batch complexity and improves in many situations. At step t, we have samples $A_1, \dots, A_t$ from $\mathcal{D}'$. For each subset $B \subset [n]$ we denote $\tilde{\mathcal{D}'}_t(B) = \frac{\sum_{j=1}^t \mathbf{1}_{A_j\in B}}{t}$ and by $\mu_t(U_n)$ the expected value of $\max_{B\subset[n]} \left|\tilde{U}_{n,t}(B) - |B|/n\right|$. The algorithm is formally described in Algorithm 3.

Algorithm 3 Distinguish between $\mathcal{D}' = U_n$ and $\mathrm{TV}(\mathcal{D}', U_n) > \varepsilon$ with high probability

Require: $A_1, \dots$ samples from $\mathcal{D}'$
Ensure: Accept if $\mathcal{D}' = U_n$ and Reject if $\mathrm{TV}(\mathcal{D}', U_n) > \varepsilon$ with probability of error less than δ
  $t = \min\{n, \sqrt{n\log(2/\delta)}\}$, $W = 1$
  while $W = 1$ **do**
    if $\exists B \subset [\lfloor n/2 \rfloor] : \left|\tilde{\mathcal{D}'}_t(B) - |B|/n\right| > \max\{\phi(\delta, |B|/n, t), \phi(\delta, 1 - |B|/n, t)\}$ or $\mathrm{TV}(\tilde{\mathcal{D}'}_t, U_n) > \mu_t(U_n) + 4\min\left(1, \frac{t^{3/2}}{n^{3/2}}\right)\sqrt{\frac{\log\left(\frac{2t(t+1)}{\delta}\right)}{2t}}$ **then**
      $W = 0$
      return 2
    else if $\mathrm{TV}(\tilde{\mathcal{D}'}_t, U_n) < \mu_t(U_n) + C\min\left(\frac{t^2\varepsilon^2}{n^2}, \varepsilon^2\sqrt{\frac{t}{n}}, \varepsilon\right) - 4\min\left(1, \frac{t^{3/2}}{n^{3/2}}\right)\sqrt{\frac{\log\left(\frac{2t(t+1)}{\delta}\right)}{2t}}$ **then**
      $W = 0$
      return 1
    else
      $t = t + 1$
    end if
  end while

Let $n' = \min\{n, \sqrt{n\log(2/\delta)}\}$ and $\Psi_t = 4\min\left(1, \frac{t^{3/2}}{n^{3/2}}\right)\sqrt{\log\left(\frac{2t(t+1)}{\delta}\right)/(2t)}$, the stopping times τ_1 and τ_2 of Algorithm 3 are then defined by :

$$\tau_1 = \inf\left\{t \geq n' : \mathrm{TV}(\tilde{\mathcal{D}'}_t, U_n) < \mu_t(U_n) + C\min\left(\frac{t^2\varepsilon^2}{n^2}, \varepsilon^2\sqrt{\frac{t}{n}}, \varepsilon\right) - \Psi_t\right\}$$

and

$$\tau_2 = \inf\left\{ t \geq n' : \exists B \subset [\lfloor n/2 \rfloor] : \left| \tilde{\mathcal{D}'}_t(B) - \frac{|B|}{n} \right| > \max\{\phi(\delta, |B|/n, t), \phi(\delta, 1 - |B|/n, t)\} \right.$$
$$\left. \text{or } \mathrm{TV}(\tilde{\mathcal{D}'}_t, U_n) > \mu_t(U_n) + \Psi_t \right\} .$$

In order to compare the sequential Algorithm 3 with the batch one of [DGPP17], we need to show first that it is indeed a δ-correct algorithm and show that its stopping times are smaller in expectation than the batch sample complexity. In the following theorem, an upper bound of the expected stopping times is given:

Theorem 2.5.1. *Algorithm 3 is δ-correct and its stopping times satisfy:*

$$\mathbb{E}(\tau_1(U_n)) \leq \max\left\{ \frac{2\log(1/\delta)}{C^2\varepsilon^2} + \frac{8}{C^2\varepsilon^2}\log\frac{2\log(1/\delta)}{C^2\varepsilon^2}, \left(\frac{2n\log(1/\delta)}{C^2\varepsilon^4} + \frac{4}{C^2\varepsilon^4}\log\frac{2n\log(1/\delta)}{C^2\varepsilon^4}\right)^{1/2} \right\}, \tag{2.167}$$

and for $d = \mathrm{TV}(\mathcal{D}', U_n) > \varepsilon$ and $B_d := \{i : \mathcal{D}'_i > (1+d)/n\}$:

$$\mathbb{E}(\tau_2(\mathcal{D}')) \leq \min\left\{ \max\left\{ \frac{3\log(1/\delta)}{C^2 d^2}, \left(\frac{3n\log(1/\delta)}{C^2 d^4}\right)^{1/2} \right\}, \right.$$
$$\left. \frac{\log(2^{n-1}/\delta)}{\min\{\mathrm{KL}(|B_d|/n \pm d/2, |B_d|/n)\}}, \frac{\log(2^{n-1}/\delta)}{\min\{\mathrm{KL}(|B_{opt}|/n \pm d, |B_{opt}|/n)\}} \right\}.$$

Proof. We prove here that Algorithm 3 has an error probability less than δ. The proof relies on the following uniform concentration lemma for $\mathrm{TV}(\tilde{\mathcal{D}}_t, U_n)$:

Lemma 2.5.2. *Let $\Psi_t = 4\min\left(1, \frac{t^{3/2}}{n^{3/2}}\right)\sqrt{\log\left(\frac{2t(t+1)}{\delta}\right)/(2t)}$, we have:*

$$\mathbb{P}\left(\exists t \geq \min\{n, \sqrt{n\log(2/\delta)}\} : \left|\mathrm{TV}(\tilde{\mathcal{D}}_t, U_n) - \mathbb{E}[\mathrm{TV}(\tilde{\mathcal{D}}_t, U_n)]\right| > \Psi_t\right) \leq \delta/2 . \tag{2.168}$$

Proof. If $n \leq \sqrt{n\log(2/\delta)}$, we apply the union bound along with McDiarmid's inequality on $\mathrm{TV}(\tilde{\mathcal{D}}_t, U_n)$ which is $(1/t, \ldots, 1/t)$-bounded to obtain:

$$\mathbb{P}\left(\exists t \geq n : \left|\mathrm{TV}(\tilde{\mathcal{D}}_t, U_n) - \mathbb{E}[\mathrm{TV}(\tilde{\mathcal{D}}_t, U_n)]\right| > 4\sqrt{\log\left(\frac{2t(t+1)}{\delta}\right)/(2t)}\right) \leq \sum_{t\geq n}\frac{\delta}{2t(t+1)} . \tag{2.169}$$

If $n > \sqrt{n\log(2/\delta)}$, since the last inequality it remains to consider $t < n$, an application

of the Bernstein form of McDiarmid's inequality (as detailed in [DK17]) permits to deduce

$$\mathbb{P}\left(\exists t \in [\sqrt{n\log(2/\delta)}, n] : \left|\mathrm{TV}(\tilde{\mathcal{D}}_t, U_n) - \mathbb{E}[\mathrm{TV}(\tilde{\mathcal{D}}_t, U_n)]\right| > 4\frac{t^{3/2}}{n^{3/2}}\sqrt{\log\left(\frac{2t(t+1)}{\delta}\right)/(2t)}\right) \tag{2.170}$$

$$\leq \sum_{\sqrt{n\log(2/\delta)}\leq t<n} \exp\left(\frac{-8\frac{t^2}{n^3}\log\left(\frac{2t(t+1)}{\delta}\right)}{4\frac{t^2}{n^3} + \frac{8t\sqrt{\log\left(\frac{2t(t+1)}{\delta}/2\right)}}{3n^{5/2}}}\right) \leq \sum_{\sqrt{n\log(2/\delta)}\leq t<n} \exp\left(-\log\left(\frac{2t(t+1)}{\delta}\right)\right) \tag{2.171}$$

$$\leq \sum_{\sqrt{n\log(2/\delta)}\leq t<n} \frac{\delta}{2t(t+1)}\,. \tag{2.172}$$

Combining both inequalities the lemma follows. □

The proof of correctness is detailed below:

- If $\mathcal{D}' = U_n$, using Lemma 2.5.2 and letting $n' = \min\{n, \sqrt{n\log(2/\delta)}\}$ and $\Psi_t = 4\min\left(1, \frac{t^{3/2}}{n^{3/2}}\right)\sqrt{\log\left(\frac{2t(t+1)}{\delta}\right)/(2t)}$, the probability of error can be bounded as :

$$\mathbb{P}\left(\tau_2 \leq \tau_1\right) \tag{2.173}$$

$$\leq \mathbb{P}\left(\exists t \geq n', \exists B \subset [\lfloor n/2 \rfloor] : \left|\tilde{\mathcal{D}'}_t(B) - \frac{|B|}{n}\right| > \max\{\phi(\delta, |B|/n, t), \phi(\delta, 1-|B|/n, t)\}\right) \tag{2.174}$$

$$+ \mathbb{P}\left(\exists t \geq n' : \mathrm{TV}(\tilde{\mathcal{D}'}_t, U_n) > \mu_t(U_n) + \Psi_t\right) \tag{2.175}$$

$$\leq \sum_{t\geq n', B\subset[\lfloor n/2\rfloor]} \mathbb{P}\left(\left|\tilde{\mathcal{D}'}_t(B) - \frac{|B|}{n}\right| > \max\{\phi(\delta, |B|/n, t), \phi(\delta, 1-|B|/n, t)\}\right) + \delta/2 \tag{2.176}$$

$$\leq \sum_{t\geq n', B\subset[\lfloor n/2\rfloor]} e^{-t\,\mathrm{KL}(|B|/n+\phi(\delta,|B|/n,t),|B|/n)} + e^{-t\,\mathrm{KL}(|B|/n-\phi(\delta,1-|B|/n,t),|B|/n)} + \delta/2 \tag{2.177}$$

$$\leq \delta\,. \tag{2.178}$$

- If $\mathrm{TV}(\mathcal{D}', U_n) > \varepsilon$, the probability of error can be bounded as:

$$\mathbb{P}\left(\tau_1 \leq \tau_2\right) \tag{2.179}$$

$$= \mathbb{P}\left(\exists t \geq n' : \mathrm{TV}(\tilde{\mathcal{D}'}_t, U_n) < \mu_t(U_n) + C\min\left(\frac{t^2\varepsilon^2}{n^2}, \varepsilon^2\sqrt{\frac{t}{n}}, \varepsilon\right) - \Psi_t\right) \tag{2.180}$$

$$\overset{(i)}{\leq} \mathbb{P}\left(\exists t \geq n' : \left|\mathrm{TV}(\tilde{\mathcal{D}'}_t, U_n) - \mathbb{E}(\mathrm{TV}(\tilde{\mathcal{D}'}_t, U_n))\right| \geq \Psi_t\right) \tag{2.181}$$

$$\overset{(ii)}{\leq} \delta\,. \tag{2.182}$$

where (i) follows from the triangle inequality and Lemma 2.5.1 and (ii) follows from Lemma 2.5.2.

The sample complexity of Algorithm 3 is given by the stopping time τ_1 if the input consists of samples from the uniform distribution and by the stopping time τ_2 if the samples are from a distribution ε-far from the uniform distribution. Let's start by upper bounding τ_1, the two regions related to the two stopping rules concur when

$$4\min\left(1, \frac{t^{3/2}}{n^{3/2}}\right)\sqrt{\frac{1}{2t}\log\left(\frac{t(t+1)}{\delta}\right)} \leq \frac{C}{2}\min\left\{\frac{\varepsilon^2 t^2}{n^2}, \varepsilon^2\sqrt{\frac{t}{n}}, \varepsilon\right\}, \tag{2.183}$$

and this latter condition is guaranteed by Lemma 2.4.2 if

$$t \geq N'_\varepsilon := \max\left\{\frac{2\log(1/\delta)}{C^2\varepsilon^2} + \frac{8}{C^2\varepsilon^2}\log\frac{2\log(1/\delta)}{C^2\varepsilon^2}, \tag{2.184}$$

$$\left(\frac{2n\log(1/\delta)}{C^2\varepsilon^4} + \frac{4}{C^2\varepsilon^4}\log\frac{2n\log(1/\delta)}{C^2\varepsilon^4}\right)^{1/2}\right\}. \tag{2.185}$$

Therefore,

$$\mathbb{E}(\tau_1(U_n)) \leq N'_\varepsilon. \tag{2.186}$$

It remains to upper bound $\mathbb{E}(\tau_2(\mathcal{D}'))$ for every distribution $\mathcal{D}'$ verifying $d = \mathrm{TV}(\mathcal{D}', U_n) > \varepsilon$. Let B_{opt} the smallest subset of $[n]$ such that $|B_{opt}| \leq n/2$ and $|\mathcal{D}'(B_{opt}) - U_n(B_{opt})| = d$. We make a case study on the size of $|B_{opt}|$:

- If $|B_{opt}| > \sqrt{n\log(1/\delta)}$, we have by letting $\Psi_t = \min\left(1, \frac{t^{3/2}}{n^{3/2}}\right)\sqrt{\log\left(\frac{t(t+1)}{\delta}\right)/(2t)}$:

$$\mathbb{E}(\tau_2) \leq N'_d + \sum_{s \geq N'_d}\mathbb{P}(\tau_2 > s) \leq N'_d + \sum_{t \geq N'_d - 1}\mathbb{P}\left(\mathrm{TV}(\tilde{\mathcal{D}'}_t, U_n) \leq \mu_t(U_n) + 4\Psi_t\right) \tag{2.187}$$

$$\leq N'_d + \sum_{t \geq N'_d - 1}\mathbb{P}\left(|\mathrm{TV}(\tilde{\mathcal{D}'}_t, U_n) - \mathbb{E}(\mathrm{TV}(\tilde{\mathcal{D}'}_t, U_n))| \geq C\min\left\{\frac{d^2t^2}{n^2}, d^2\sqrt{\frac{t}{n}}, d\right\} - 4\Psi_t\right) \tag{2.188}$$

$$\leq N'_d + \sum_{t \geq N'_d - 1}\mathbb{P}\left(|\mathrm{TV}(\tilde{\mathcal{D}'}_t, U_n) - \mathbb{E}(\mathrm{TV}(\tilde{\mathcal{D}'}_t, U_n))| \geq 4\Psi_t\right) \leq N'_d + \delta. \tag{2.189}$$

- If $|B_{opt}| \leq \sqrt{n\log(1/\delta)}$, we take $\alpha \in (0,1)$ and define N_α as the minimum positive integer such that for all integers $t \geq N_\alpha$, $\max\{\phi(\delta, |B_{opt}|/n, t), \phi(\delta, 1-|B_{opt}|/n, t)\} \leq \alpha d$. The existence of N_α is guaranteed since $\lim_{t\to\infty}\phi(B,t) = 0$. We have $\max\{\phi(\delta, |B_{opt}|/n, N_\alpha), \phi(\delta, 1-|B_{opt}|/n, N_\alpha)\} \leq \alpha d$ and $\max\{\phi(\delta, |B_{opt}|/n, N_\alpha - 1), \phi(\delta, 1-|B_{opt}|/n, N_\alpha - 1)\} > \alpha d$ hence,

$$\min\left\{\mathrm{KL}\left(\frac{|B_{opt}|}{n} + \alpha d, \frac{|B_{opt}|}{n}\right), \mathrm{KL}\left(\frac{|B_{opt}|}{n} - \alpha d, \frac{|B_{opt}|}{n}\right)\right\} \leq \frac{\log\left(\frac{2^{n-1}(N-1)N}{\delta}\right)}{N-1}. \tag{2.190}$$

Thus $\lim_{\delta\to 0} N = +\infty$ and from Lemma 2.4.2 we can deduce

$$N \leq \frac{\log(2^{n-1}/\delta) + 4\log\left(\frac{\log(2^{n-1}/\delta)}{\min\{\mathrm{KL}(|B_{opt}|/n\pm\alpha d, |B_{opt}|/n)\}}\right)}{\min\{\mathrm{KL}(|B_{opt}|/n \pm \alpha d, |B_{opt}|/n)\}} + 1\ . \tag{2.191}$$

Therefore,

$$\begin{aligned}
\mathbb{E}(\tau_2(\mathcal{D}')) &\leq N_\alpha + \sum_{t\geq N_\alpha} \mathbb{P}(\tau_2(\mathcal{D}') \geq t) && (2.192)\\
&\leq N_\alpha + \sum_{s\geq N_\alpha - 1} \mathbb{P}(|\tilde{\mathcal{D}'}_s(B_{opt}) - |B_{opt}|/n| \leq \phi(B_{opt}, t)) && (2.193)\\
&\leq N_\alpha + \sum_{s\geq N_\alpha - 1} \mathbb{P}(|\tilde{\mathcal{D}'}_s(B_{opt}) - \mathcal{D}'(B_{opt})| > d - \alpha d) \text{ (by definition of } N_\alpha) && (2.194)\\
&\leq N_\alpha + \sum_{s\geq N_\alpha - 1} e^{-2s((1-\alpha)d)^2} \text{ (Chernoff-Hoeffding's inequality)} && (2.195)\\
&\leq N_\alpha + \frac{e^{-2(N_\alpha-1)((1-\alpha)d)^2}}{1 - e^{-2((1-\alpha)d)^2}} && (2.196)\\
&\leq \frac{\log(2^{n-1}/\delta) + 4\log\left(\frac{\log(2^{n-1}/\delta)}{\min\{\mathrm{KL}(|B_{opt}|/n\pm\alpha d, |B_{opt}|/n)\}}\right)}{\min\{\mathrm{KL}(|B_{opt}|/n \pm \alpha d, |B_{opt}|/n)\}} + 1 + \frac{e^{-2(N_\alpha-1)((1-\alpha)d)^2}}{1 - e^{-2((1-\alpha)d)^2}}. && (2.197)
\end{aligned}$$

After dividing by $\log(1/\delta)$, taking the limit $\delta \to 0$ then $\alpha \to 1$ we deduce finally:

$$\limsup_{\delta\to 0} \frac{\mathbb{E}(\tau_2(\mathcal{D}'))}{\log(1/\delta)} \leq \frac{1}{\min\{\mathrm{KL}(|B_{opt}|/n \pm d, |B_{opt}|/n)\}} \underset{d\to 0}{\sim} \frac{2\frac{|B_{opt}|}{n}\left(1 - \frac{|B_{opt}|}{n}\right)}{d^2}$$

.

This is an interesting improvement especially when $|B_{opt}| \ll n$ but it is not sensitive to small fluctuations of $\mathcal{D}_i$ around $1/n$. For example the distribution $\{1/n + \varepsilon, 1/n, \ldots, 1/n, 1/n - \varepsilon\}$ whose B_{opt} has size 1 can easily transformed to a distribution with an optimal set of size $n/2$ by adding small noise $\eta \ll \varepsilon$ to $n/2-1$ parts among those of $1/n$ probability mass. Even though the transformed distribution has an optimal set of size $n/2$ (hence a large upper bound of the complexity), Algorithm 3 seems to stop on pretty the same time for both distributions. To overcome this inconvenient in this upper bound, we can use the same method to prove that for $B_d := \{i : \mathcal{D}_i > (1+d)/n\} \subset B_{opt}$ we have:

$$\limsup_{\delta\to 0} \frac{\mathbb{E}(\tau_2(\mathcal{D}'))}{\log(1/\delta)} \leq \frac{1}{\min\{\mathrm{KL}(|B_d|/n \pm d/2, |B_d|/n)\}} \underset{d\to 0}{\sim} \frac{8\frac{|B_d|}{n}\left(1 - \frac{|B_d|}{n}\right)}{d^2}.$$

This upper bound has the advantage to count only the bigger parts of the distributions for which the noise is of the order d/n at the cost of multiplying the upper bound by almost 4. □

We plot in the same Figure 2.3 the batch complexity and the sequential stopping time τ_2 if the statistic takes into account the optimal set B_{opt} and if not. It is clear that

the proposed Algorithm 3 is superior than the batch algorithm of [DGPP17] especially when $|B_{opt}|$ is smaller than $\sqrt{n\log(1/\delta)}$ which is also exhibited in Figure 2.3 for a specific example.

We've exposed so far the advantages of sequential procedures over batch algorithms, they share the idea that if the tested distributions are far away we can hope to stop earlier. However one can wonder if there is a tangible improvement independent of distributions. We focus on the worst case setting where we consider the eventual improvement of sequential procedures that works for all distributions. We show that we cannot hope to improve the dependency on n found in the batch setting more than a constant and replacing ε by $\varepsilon \vee \mathrm{TV}(\mathcal{D}, U_n)$. For instance we can prove the following lower bounds for uniformity testing.

Theorem 2.5.2. *There is no stopping rule T for the problem of testing $\mathcal{D} = U_n$ vs $\mathrm{TV}(\mathcal{D}, U_n) > \varepsilon$ with an error probability δ such that*

$$\mathbb{P}\left(\tau_2(T,\mathcal{D}) \leq c\frac{\sqrt{n\log(1/3\delta)}}{\mathrm{TV}(\mathcal{D},U_n)^2}\right) \geq 1-\delta \quad \textit{if } \mathrm{TV}(\mathcal{D},U_n) > \varepsilon \textit{ and} \tag{2.198}$$

$$\mathbb{P}\left(\tau_1(T,U_n) \leq c\frac{\sqrt{n\log(1/3\delta)}}{\varepsilon^2}\right) \geq 1-\delta\,, \tag{2.199}$$

where c a universal constant. We have similar statement if we replace $\sqrt{n\log(1/3\delta)}$ by $\log(1/3\delta)$.

This can be proven using pretty the same construction of distributions as for the batch lower bounds along with Wald's lemma.

Proof. We prove only the first statement, the others being similar. Suppose that such a stopping rule exists. Let $d > \varepsilon$ and $m = c\frac{\sqrt{n\log(1/3\delta)}}{d^2}$. Let U_n the uniform distribution and D a uniformly chosen distribution where $\mathcal{D}_i = \frac{1\pm 2d}{n}$ with probability $1/2$ each. With the work of [DK16] (Section 3), we can show that $\mathrm{KL}(\mathcal{D}^{\times \mathrm{Poi}(m)}, U_n^{\times \mathrm{Poi}(m)}) \leq C\frac{m^2d^4}{n}$ where C is a constant. Therefore

$$\begin{aligned}
\mathrm{KL}(\mathcal{D}^{\times m}, U_n^{\times m}) &= m\,\mathrm{KL}(\mathcal{D}, U_n) && (2.200)\\
&= \mathbb{E}(\mathrm{Poi}(m))\,\mathrm{KL}(\mathcal{D}, U_n) && (2.201)\\
&= \mathrm{KL}(\mathcal{D}^{\times \mathrm{Poi}(m)}, U_n^{\times \mathrm{Poi}(m)}) \quad \text{(Wald's lemma)} && (2.202)\\
&\leq C\frac{m^2d^4}{n}\,. && (2.203)
\end{aligned}$$

But

$$\begin{aligned}
\mathrm{KL}(\mathcal{D}^{\times m}, U_n^{\times m}) &\geq \mathrm{KL}(\mathbb{P}_{\mathcal{D}}(\tau_2 \leq m), \mathbb{P}_{U_n}(\tau_2 \leq m)) && (2.204)\\
&\geq \mathrm{KL}(1-\delta, \delta) && (2.205)\\
&\geq \log(1/3\delta)\,, && (2.206)
\end{aligned}$$

since $\mathbb{P}_D(\tau_2 \leq m) \geq 1-\delta$ and $\mathbb{P}_{U_n}(\tau_2 \leq m) = \mathbb{P}_{U_n}(\tau_2 \leq m, \tau_1 < \tau_2) + \mathbb{P}_{U_n}(\tau_2 \leq m, \tau_1 \geq \tau_2) \leq \delta$. Hence

$$C\frac{\left(c\frac{\sqrt{n\log(1/3\delta)}}{d^2}\right)^2 d^4}{n} \geq \log(1/3\delta)\,, \tag{2.207}$$

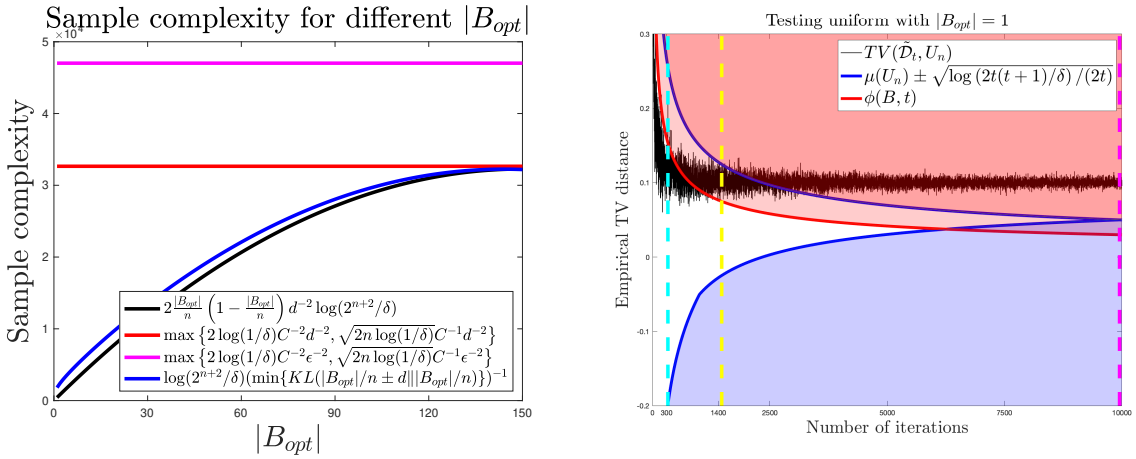

Figure 2.3: Left: We compare different upper bounds on the sample complexity for uniformity testing with $n = 300, \varepsilon = 0.05, \delta = 10^{-10}$ and $d = \mathrm{TV}(\mathcal{D}, U_n) = 0.06$. We remark that for $|B_{opt}| < n/2$ we have better upper bounds. Right: Uniformity testing for $n = 10, \delta = 10^{-10}, \mathrm{TV}(\mathcal{D}', U_n) = \varepsilon = 0.1$ and $|B_{opt}| = 1$. We compare the empirical TV distance with different thresholds. The red zone corresponds to accepting H_2 while the blue one corresponds to accepting H_1. The magenta line identifies the batch threshold, the yellow line designates the sequential stopping time without looking on B_{opt} and finally the cyan line defines the actual stopping time of Algorithm 3.

which gives the contradiction if $c < 1/\sqrt{C}$.

□

2.6 Testing closeness-the general case

In this section we consider testing closeness in the general case $n \geq 3$. Let us recall that we have $\mathcal{D}_1$ and $\mathcal{D}_2$ two unknown distributions on $[n]$ and we want to distinguish $\mathcal{D}_1 = \mathcal{D}_2$ and $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$ with high probability $1 - \delta$. Inspired by the case of the Bernoulli distribution, we describe how to transform a batch algorithm to a sequential one with better expected sample complexity. In that case, however, identifying the sample complexity exactly remains out of reach: the dependency on ε, δ and n can be computed only up to a multiplicative constant.

2.6.1 Batch setting

Recently, [DGKPP20] have shown that the dependence on the error probability in the sample complexity of the closeness problem could be better than the $\log 1/\delta$ found by repeating $\log 1/\delta$ times the classical algorithm of [CDVV14] and accepting or rejecting depending on the majority test. More precisely:

Theorem 2.6.1 ([DGKPP20]). $\Theta\left(\max\left(\frac{n^{2/3}\log^{1/3}(1/\delta)}{\varepsilon^{4/3}}, \frac{n^{1/2}\log^{1/2}(1/\delta)}{\varepsilon^2}, \frac{\log(1/\delta)}{\varepsilon^2}\right)\right)$ *samples are necessary and sufficient to test whether* $\mathcal{D}_1 = \mathcal{D}_2$ *or* $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$ *with an error* $\delta > 0$.

The main ingredient of a closeness tester is an efficient test statistic which can distinguish between the two hypothesis. Let us define by X_i (resp. Y_i) the number of samples

from $\mathcal{D}_1$ (resp. $\mathcal{D}_2$) whose values are equal to $i \in [n]$. Thinking to the TV distance we use, we could be tempted to take a decision based on the statistic $\sum_{i=1}^{n} |X_i - Y_i|$. However this simple statistic suffers from a principal caveat: its expected value is neither zero nor easily lower bounded when $\mathcal{D}_1 = \mathcal{D}_2$. As a remedy, [DGKPP20] propose to use the following statistic: $Z = \sum_{i=1}^{n} |X_i - Y_i| + |X'_i - Y'_i| - |X_i - X'_i| - |Y_i - Y'_i|$, where X'_i and Y'_i correspond to a second set of independent samples. The expected value of the estimator Z is obviously 0 when the distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ are equal. On the other hand when $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$, they provide a lower bound on the expected value of the estimator Z which enable to test closeness between $\mathcal{D}_1$ and $\mathcal{D}_2$. Since these results turn out to be similarly useful in our subsequent analysis, we summarized them in the following lemma.

Lemma 2.6.1 ([DGKPP20]). *Let $d = \mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2)$. Let $k \geq 1$ and $(k_1, k_2, k'_1, k'_2) \sim Multinom(4k, (1/4, 1/4, 1/4, 1/4))$. Let $(X_i)_{i=1}^{k_1}$ and $(X'_i)_{i=1}^{k'_1}$ two sets of i.i.d. samples from $\mathcal{D}_1$ and $(Y_i)_{i=1}^{k_2}$ and $(Y'_i)_{i=1}^{k'_2}$ two sets of i.i.d. samples from $\mathcal{D}_2$. Then there are universal constants c and C such that*

- *If $\mathcal{D}_1 = \mathcal{D}_2$, $\mathbb{E}[Z] = 0$.*
- *If $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$, $\mathbb{E}[Z] \geq C \min\left\{kd, \frac{k^2 d^2}{n}, \frac{k^{3/2} d^2}{\sqrt{n}}\right\} - c\sqrt{k}$.*

The lower bound on the expectation of Z is obtained by a technique of Poissonization. Note that this lower bound is stronger than the one obtained for the *chi*-square estimator; $\sum_{i=1}^{n} \frac{(X_i - Y_i)^2 - X_i - Y_i}{X_i + Y_i}$, used by [CDVV14]. In fact, for far distributions the lower bound on the expected value of the *chi*-square estimator does not allow the best dependency on n, ε and δ. This lemma is the key ingredient behind the batch algorithm. Indeed, for sufficiently large $k = \Omega\left(\max\left(\frac{n^{2/3}\log^{1/3}(1/\delta)}{\varepsilon^{4/3}}, \frac{n^{1/2}\log^{1/2}(1/\delta)}{\varepsilon^2}, \frac{\log(1/\delta)}{\varepsilon^2}\right)\right)$, [DGKPP20] show that $\mathbb{E}[Z] \geq C''\sqrt{k \log 1/\delta}$ for a universal constant C'' if $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$, then by applying McDiarmid's inequality they prove that the algorithm consisting of returning H_2 if $Z \geq C''\sqrt{k \log 1/\delta}/2$ and returning H_1 otherwise is δ-correct. In the following we draw our inspiration from them to design a sequential algorithm for testing closeness.

2.6.2 Sequential setting

In this section we present how the sequential setting can improve the sample complexity found in the batch setting. We base our sequential tester on the same test statistic Z as [DGKPP20], but we allow the stopping rules of this new algorithm to be time-dependent. When the distributions to be tested $\mathcal{D}_1$ and $\mathcal{D}_2$ are equal, the estimator Z_t cannot be very large, and if they are ε-far the estimator cannot be very small: at each step, the tester compares Z_t to some well chosen thresholds. If it cannot decide with sufficient confidence, it asks for more samples. This can possibly last until the two regions of decision meet. This time has the order of the complexity of the batch algorithm. This tester is formally defined in Algorithm 4. For sake of simplicity, let us denote by $\Delta_t = C \min\left\{t\varepsilon, \frac{t^2\varepsilon^2}{n}, \frac{t^{3/2}\varepsilon^2}{\sqrt{n}}\right\} - c\sqrt{t}$ and by $\Psi_t = 2\sqrt{2t \log\left(\frac{\pi^2}{3\delta}\right) + 4et \log(\log(4t) + 1)}$. The stopping times τ_1 and τ_2 of Algorithm 4 are then defined by

$$\tau_1 = \inf\left\{t \geq 1 : |Z_t| \leq \Delta_t - \Psi_t\right\}, \text{ and } \tau_2 = \inf\left\{t \geq 1 : |Z_t| > \Psi_t\right\}. \tag{2.208}$$

We prove now that Algorithm 4 is δ-correct and then study its sample complexity.

Algorithm 4 Distinguish between $\mathcal{D}_1 = \mathcal{D}_2$ and $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$ with high probability

Require: $A_1, \dots$ samples from $\mathcal{D}_1$ and $B_1, \dots$ samples from $\mathcal{D}_2$
Ensure: Accept if $\mathcal{D}_1 = \mathcal{D}_2$ and Reject if $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$ with probability of error less than δ
$t = 1$, $W = 1$
while $W = 1$ **do**
$(m_{1,t}, m'_{1,t}, m_{2,t}, m'_{2,t}) \sim \mathrm{Multinom}(4t, (1/4, 1/4, 1/4, 1/4))$

$$Z_t = \sum_{i=1}^{n} |X_i - Y_i| + |X'_i - Y'_i| - |X_i - X'_i| - |Y_i - Y'_i| \,, \tag{2.209}$$

where X_i(resp. X'_i, Y_i, Y'_i) are the numbers of i's in the word formed with $m_{1,t}$ (resp. $m'_{1,t}, m_{2,t}, m'_{2,t}$) samples from $\mathcal{D}_1$ (resp. $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_2$) . We need only to sample the difference of $(m_{1,t} - m_{1,t-1})^+ + (m'_{1,t} - m'_{1,t-1})^+$ from $\mathcal{D}_1$ and $(m_{2,t} - m_{2,t-1})^+ + (m'_{2,t} - m'_{2,t-1})^+$ from $\mathcal{D}_2$.
if $|Z_t| > 2\sqrt{2t \log\left(\frac{\pi^2}{3\delta}\right) + 4et \log(\log(4t) + 1)}$ **then**
$W = 0$
return 2
else if $|Z_t| \leq C \min\left\{t\varepsilon, \frac{t^2\varepsilon^2}{n}, \frac{t^{3/2}\varepsilon^2}{\sqrt{n}}\right\} - c\sqrt{t} - 2\sqrt{2t \log\left(\frac{\pi^2}{3\delta}\right) + 4et \log(\log(4t) + 1)}$ **then**
$W = 0$
return 1
else
$t = t + 1$
end if
end while

Correctness

We prove here that Algorithm 4 has an error probability less than δ. The proof relies on the following uniform concentration lemma for Z_t:

Lemma 2.6.2. *For $\eta, s > 1$, let $J(\eta, s, t) = \sqrt{2\eta ts \log\left(\frac{\log(t)}{\log(\eta)} + 1\right) + 2t \log\left(\frac{2\zeta(s)}{\delta}\right)}$, where $\zeta(s) = \sum_{n \geqslant 1} \frac{1}{n^s}$. Then*

$$\mathbb{P}\left(\exists t \geq 1 : |Z_t - \mathbb{E}[Z_t]| > J(\eta, s, 4t)\right) \leq \delta \,. \tag{2.210}$$

The proof of this lemma is inspired from [HRMS18] and relies on dividing the set of integers into some well chosen subsets, applying union bound and finally invoking McDiarmid's inequality with specific arguments for each interval. Note that Lemma 2.6.2 yields the best second order term in the complexity (up to constant factor), in contrast to a simple union bound on McDiarmid's inequality. We do not use this feature in our study of the sample complexity of the testing closeness problem as we are interested here in leading terms only (see Theorem 2.6.2). However, the $\log - \log$ dependency proves useful when showing that Algorithm 4 used with $\varepsilon = 0$ obtains the optimal sample complexity for testing $\mathcal{D}_1 = \mathcal{D}_2$ vs $\mathcal{D}_1 \neq \mathcal{D}_2$ (see Theorem 2.6.4).

Proof. The proof uses similar arguments of [HRMS18]. Actually Z_t is a function of $4t$ variables (the samples from the distributions) and has the property $(2,\dots,2)$-bounded differences. McDiarmid's inequality implies $\mathbb{P}\left(\exists t \geq 1 : |Z_t - \mathbb{E}[Z_t]| \geq a + 4bt/a\right) \leq 2e^{-2b}$, taking the intervals $I_k = [\eta^k, \eta^{k+1})$ for k integer we deduce for $b_k = \frac{1}{2}\log\left(\frac{2(k+1)^s}{\zeta(s)^{-1}\delta}\right)$ and $a_k = \frac{b_k}{a_k}\eta^{k+1}$ that

$$\mathbb{P}\left(\exists t \geq 1 : |Z_t - \mathbb{E}[Z_t]| \geq J(\eta, s, 4t)\right) \leq \sum_{k\geq 0} \mathbb{P}\left(\exists t \in I_k : |Z_t - \mathbb{E}[Z_t]| \geq J(\eta, s, 4t)\right) \tag{2.211}$$

$$\leq \sum_{k\geq 0} \mathbb{P}\left(\exists t \in I_k : |Z_t - \mathbb{E}[Z_t]| \geq a_k + 4b_k t/a_k\right) \tag{2.212}$$

$$\leq \sum_{k\geq 0} 2e^{-2b_k} \leq \sum_{k\geq 0} \delta \frac{\zeta(s)^{-1}}{(k+1)^s} \leq \delta\ . \tag{2.213}$$

□

For $\eta = e$ and $s = 2$, the function J becomes $J(e, 2, 4t) = \Psi_t$ and we can use Lemma 2.6.2 to prove the correctness of Algorithm 4 as sketched below:

- If $\mathcal{D}_1 = \mathcal{D}_2$, using Lemma 2.6.2, the probability of error can be bounded as:

$$\mathbb{P}\left(\tau_2 \leq \tau_1\right) \leq \mathbb{P}\left(\exists t \geq 1 : |Z_t| > \Psi_t\right) \leq \delta\ . \tag{2.214}$$

- If $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$, the probability of error can be bounded as:

$$\mathbb{P}\left(\tau_1 \leq \tau_2\right) = \mathbb{P}\left(\exists t \geq 1 : |Z_t| \leq \Delta_t - \Psi_t\right) \tag{2.215}$$

$$\overset{(i)}{\leq} \mathbb{P}\left(\exists t \geq 1 : |Z_t - \mathbb{E}(Z_t)| \geq \mathbb{E}(Z_t) - \Delta_t + \Psi_t\right) \tag{2.216}$$

$$\overset{(ii)}{\leq} \mathbb{P}\left(\exists t \geq 1 : |Z_t - \mathbb{E}(Z_t)| \geq \Psi_t\right) \overset{(iii)}{\leq} \delta\ . \tag{2.217}$$

  where (i) follows from the triangle inequality $|Z_t - \mathbb{E}(Z_t)| \geq \mathbb{E}(Z_t) - Z_t$, (ii) follows by the fact that $\mathbb{E}(Z_t) \geq \Delta_t$ from Lemma 2.6.1 and (iii) follows from Lemma 2.6.2.

Complexity

In order to show the advantage of our sequential algorithm, we need to upper bound the expectations of the stopping times τ_1 and τ_2. This is done in the following theorem:

Theorem 2.6.2. *Let $d = \mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2)$. The sample complexity of Algorithm 4 satisfies*

- *If $\mathcal{D}_1 = \mathcal{D}_2$, $\mathbb{E}(\tau_1(T, \mathcal{D}_1, \mathcal{D}_2)) \leq 2N_\varepsilon$.*
- *If $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$, $\mathbb{E}(\tau_2(T, \mathcal{D}_1, \mathcal{D}_2)) \leq 2N_d$.*

where for all $\eta > 0$, N_η is defined by

$$N_\eta = \max\left\{ \frac{128}{C^2}\frac{\log(\frac{\pi^2}{3\delta})}{\eta^2} + \frac{512e}{C^2\eta^2}\log\left(\log\left(\frac{128\log(\frac{\pi^2}{3\delta})}{\eta^2 C^2}\right)+1\right) + \frac{16c^2}{C^2\eta^2}, \right. \tag{2.218}$$

$$\left(\frac{128}{C^2}\frac{n^2\log(\frac{\pi^2}{3\delta})}{\eta^4} + \frac{512en^2}{C^2\eta^4}\log\left(\log\left(\frac{128}{C^2}\frac{n^2\log(\frac{\pi^2}{3\delta})}{\eta^4}\right)+1\right) + \frac{16c^2n^2}{\eta^4C^2}\right)^{1/3}, \tag{2.219}$$

$$\left.\left(\frac{128}{C^2}\frac{n\log(\frac{\pi^2}{3\delta})}{\eta^4} + \frac{512en}{C^2\eta^4}\log\left(\log\left(\frac{128}{C^2}\frac{n\log(\frac{\pi^2}{3\delta})}{\eta^4}\right)+1\right) + \frac{16c^2n}{\eta^4C^2}\right)^{1/2}\right\}, \tag{2.220}$$

and the constants c and C come from Lemma 2.6.1.

This theorem states that $\mathcal{O}\left(\max\left(\frac{n^{2/3}\log^{1/3}(1/\delta)}{(\varepsilon\vee \mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2))^{4/3}}, \frac{n^{1/2}\log^{1/2}(1/\delta)}{(\varepsilon\vee \mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2))^{2}}, \frac{\log(1/\delta)}{(\varepsilon\vee \mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2))^{2}}\right)\right)$ samples are sufficient to distinguish between $\mathcal{D}_1 = \mathcal{D}_2$ and $\mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2) > \varepsilon$ with high probability. We remark that after N_ε steps, the two stopping conditions of Algorithm 4 cannot be both unsatisfied. Therefore, the Algorithm 4 stops surely before N_ε hence it has at least a comparable complexity, in the leading terms, of the batch algorithm of [DGKPP20] when $\mathcal{D}_1 = \mathcal{D}_2$. Moreover, Algorithm 4 has the advantage of stopping rapidly when $\mathcal{D}_1$ and $\mathcal{D}_2$ are far away.

Proof. We start by the case $\mathcal{D}_1 = \mathcal{D}_2$, we know that $\mathbb{E}(\tau_1) \le \sum_{s\le N_\varepsilon}\mathbb{P}(\tau_1 \ge s) + \sum_{s>N_\varepsilon}\mathbb{P}(\tau_1 \ge s) \le N_\varepsilon + \sum_{s>N_\varepsilon}\mathbb{P}(\tau_1 \ge s)$ so it suffices to prove that $\sum_{s>N_\varepsilon}\mathbb{P}(\tau_1 \ge s) \le N_\varepsilon$. By the definition of τ_1, $\tau_1 \ge s$ implies $|Z_{s-1}| > \Delta_{s-1} - \Psi_{s-1}$ but we have chosen N_ε so that if $t = s-1 \ge N_\varepsilon$, $\Delta_{s-1} - \Psi_{s-1} \ge \frac{C}{2}\min\left\{(s-1)\varepsilon, \frac{(s-1)^2\varepsilon^2}{n}, \frac{(s-1)^{3/2}\varepsilon^2}{\sqrt{n}}\right\}$. This last claim follows from Lemma 2.6.1 and Lemma 2.4.2. Finally

$$\sum_{s>N_\varepsilon}\mathbb{P}(\tau_1 \ge s) \le \sum_{t\ge N_\varepsilon}\mathbb{P}\left(|Z_t| > \frac{C}{2}\min\left\{t\varepsilon, \frac{t^2\varepsilon^2}{n}, \frac{t^{3/2}\varepsilon^2}{\sqrt{n}}\right\}\right) \tag{2.221}$$

$$\overset{\text{(McDiarmid's inequality)}}{\le} \sum_{t\ge N_\varepsilon - 1} e^{-\frac{C^2}{16}\min\left\{t\varepsilon^2, \frac{t^3\varepsilon^4}{n^2}, \frac{t^2\varepsilon^4}{n}\right\}} \le N_\varepsilon\,. \tag{2.222}$$

The last inequality is proven in the following lemma.

Lemma 2.6.3. *We have for all $d > 0$: $\sum_{t\ge N_d} e^{-\frac{C^2}{16}\min\left\{td^2, \frac{t^3d^4}{n^2}, \frac{t^2d^4}{n}\right\}} \le N_d$.*

Proof. We have

$$\sum_{t\geq N_d} e^{-\frac{C^2}{16}\min\left\{td^2,\frac{t^3d^4}{n^2},\frac{t^2d^4}{n}\right\}} \leq \sum_{t\geq nd^{-2}} e^{-\frac{C^2}{16}td^2} + \sum_{n\geq t\geq N_d-1} e^{-\frac{C^2}{16}\frac{t^3d^4}{n^2}} + \sum_{nd^{-2}>t>n} e^{-\frac{C^2}{16}\frac{t^2d^4}{n}} \tag{2.223}$$

$$\leq \sum_{t\geq nd^{-2}} e^{-\frac{C^2}{16}td^2} + \sum_{n\geq t\geq N_d-1} e^{-2C^{1/3}\frac{td^{4/3}}{n^{2/3}}} + \sum_{nd^{-2}>t>n} e^{-\frac{C}{2}\frac{td^2}{\sqrt{n}}} \tag{2.224}$$

$$\leq \frac{1}{1-e^{-\frac{C^2}{16}d^2}} + \frac{1}{1-e^{-2C^{1/3}\frac{d^{4/3}}{n^{2/3}}}} + \frac{1}{1-e^{-\frac{C}{2}\frac{d^2}{\sqrt{n}}}} \tag{2.225}$$

$$\leq \frac{32}{C^2d^2} + \frac{n^{2/3}}{C^{1/3}d^{4/3}} + \frac{4\sqrt{n}}{Cd^2} \quad \text{since } 1-e^{-x}\geq x/2 \text{ for } 0<x<1 \tag{2.226}$$

$$\leq N_d\ . \tag{2.227}$$

□

For the case $d=\mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2)>\varepsilon$. By the definition of τ_2, $\tau_2\geq s$ implies $|Z_{s-1}|\leq\Psi_{s-1}$ hence by triangle inequality $|Z_{s-1}-\mathbb{E}(Z_{s-1})|\geq\mathbb{E}(Z_{s-1})-\Psi_{s-1}\geq\Delta_{s-1}-\Psi_{s-1}$ therefore $|Z_{s-1}-\mathbb{E}(Z_{s-1})|\geq\frac{C}{2}\min\left\{(s-1)d,\frac{(s-1)^2d^2}{n},\frac{(s-1)^{3/2}d^2}{\sqrt{n}}\right\}$ by Lemma 2.6.1. Hence

$$\sum_{s>N_\varepsilon}\mathbb{P}(\tau_2\geq s)\leq\sum_{t\geq N_\varepsilon}\mathbb{P}\left(|Z_t-\mathbb{E}(Z_{s-1})|>\frac{C}{2}\min\left\{td,\frac{t^2d^2}{n},\frac{t^{3/2}d^2}{\sqrt{n}}\right\}\right) \tag{2.228}$$

$$\overset{\text{(McDiarmid's inequality)}}{\leq}\sum_{t\geq N_d-1} e^{-\frac{C^2}{16}\min\left\{td^2,\frac{t^3d^4}{n^2},\frac{t^2d^4}{n}\right\}}\leq N_d\ . \tag{2.229}$$

The latter inequality is proven in Lemma 2.6.3. Finally $\mathbb{E}(\tau_2)\leq N_d+\sum_{s>N_d}\mathbb{P}(\tau_2\geq s)\leq 2N_d$.

□

Similar to uniformity testing, we show that we cannot improve the dependency on n found in the batch setting more than a constant and replacing ε by $\varepsilon\vee\mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2)$. We can prove the following lower bounds for testing closeness in the worst case setting.

Theorem 2.6.3. *There is no stopping rule T for the problem of testing $\mathcal{D}_1=\mathcal{D}_2$ vs $\mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2)>\varepsilon$ with an error probability δ such that*

$$\mathbb{P}\left(\tau_2(T,\mathcal{D}_1,\mathcal{D}_2)\leq c\frac{\sqrt{n\log(1/3\delta)}}{\mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2)^2}\right)\geq 1-\delta \quad \textit{if } \mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2)>\varepsilon \textit{ and} \tag{2.230}$$

$$\mathbb{P}\left(\tau_1(T,\mathcal{D}_1,\mathcal{D}_2)\leq c\frac{\sqrt{n\log(1/3\delta)}}{\varepsilon^2}\right)\geq 1-\delta \quad \textit{if } \mathcal{D}_1=\mathcal{D}_2, \tag{2.231}$$

where c a universal constant. We have similar statement if we replace $\frac{\sqrt{n\log(1/3\delta)}}{(\varepsilon\vee\mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2))^2}$ by $\frac{\log(1/3\delta)}{(\varepsilon\vee\mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2))^2}$ or $\frac{n^{2/3}\log(1/3\delta)^{1/3}}{(\varepsilon\vee\mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2))^{4/3}}$.

Proof. We can use the same techniques as the previous proof of Theorem 2.5.2, by taking the KL between samples from $U_n \times U_n$ and samples from $U_n \times D$. □

On the other hand, we can deduce from Theorem 2.6.2's proof that with high probability we have $\tau_2 \leq N_{\mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2)}$ and this upper bound has the equivalent $\mathcal{O}\left(\frac{\log\log(1/d)}{d^2} \vee \frac{n^{2/3}\log\log(1/d)^{1/3}}{d^{4/3}} \vee \frac{n^{1/2}\log\log(1/d)^{1/2}}{d^2}\right)$ for $d = \mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2) \to 0$. If we take $\varepsilon = 0$ the Algorithm 4 provides stopping rules for which it does not stop if $\mathcal{D}_1 = \mathcal{D}_2$ and rejects if $\mathcal{D}_1 \neq \mathcal{D}_2$ with probability at least $1 - \delta$.

Theorem 2.6.4. *There is a stopping rule that can decide $\mathcal{D}_1 \neq \mathcal{D}_2$ with probability at least $9/10$ using at most $\mathcal{O}\left(\frac{\log\log(1/d)}{d^2} \vee \frac{n^{2/3}\log\log(1/d)^{1/3}}{d^{4/3}} \vee \frac{n^{1/2}\log\log(1/d)^{1/2}}{d^2}\right)$ samples where $d = \mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2)$.*

This improves the results of [DK17] where the dependency in n is $n/\log n$. Furthermore, we cannot find stopping rules whose sample complexity is tighter than this upper bound as stated in the following theorem.

Theorem 2.6.5. *There is no stopping rule T for the problem of testing $\mathcal{D}_1 = \mathcal{D}_2$ vs $\mathcal{D}_1 \neq \mathcal{D}_2$ with an error probability $1/16$ such that*

$$\mathbb{P}\left(\tau_2(T,\mathcal{D}_1,\mathcal{D}_2) \leq C\frac{n^{1/2}\log\log(1/d)^{1/2}}{d^2}\right) \geq \frac{15}{16}\,, \tag{2.232}$$

where $d = \mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2)$ and C a universal constant. We have similar statements if we replace $\frac{n^{1/2}\log\log(1/d)^{1/2}}{d^2}$ by $\frac{\log\log(1/d)}{d}$ or $\frac{n^{2/3}\log\log(1/d)^{1/3}}{d^{4/3}}$.

To sum up, a number $\Theta\left(\frac{\log\log(1/d)}{d^2} \vee \frac{n^{2/3}\log\log(1/d)^{1/3}}{d^{4/3}} \vee \frac{n^{1/2}\log\log(1/d)^{1/2}}{d^2}\right)$ of samples is necessary and sufficient to decide whether $\mathcal{D}_1 = \mathcal{D}_2$ or $\mathcal{D}_1 \neq \mathcal{D}_2$ with probability $15/16$.

Proof. We use ideas similar to [KK07]. We prove only the first statement, the others being similar. Let's start by a lemma:

Lemma 2.6.4. *Let X and Y two random variables and E some event verifying $\mathbb{P}_X(E) \geq 1/3$ and $\mathbb{P}_Y(E) < 1/3$, we have*

$$\mathrm{KL}(\mathbb{P}_X,\mathbb{P}_Y) \geq -\frac{1}{3}\log(3\mathbb{P}_Y(E)) - \frac{1}{e}. \tag{2.233}$$

Proof. By data processing property of Kullback-Leibler's divergence:

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_X,\mathbb{P}_Y) &\geq \mathrm{KL}(\mathbb{P}_X(E),\mathbb{P}_Y(E)) && (2.234)\\
&\geq \mathbb{P}_X(E)\log\left(\frac{\mathbb{P}_X(E)}{\mathbb{P}_Y(E)}\right) + (1-\mathbb{P}_X(E))\log\left(\frac{1-\mathbb{P}_X(E)}{1-\mathbb{P}_Y(E)}\right) && (2.235)\\
&\geq -\frac{1}{3}\log(3\mathbb{P}_Y(E)) + (1-\mathbb{P}_X(E))\log(1-\mathbb{P}_X(E)) && (2.236)\\
&\geq -\frac{1}{3}\log(3\mathbb{P}_Y(E)) - \frac{1}{e}\,. && (2.237)
\end{aligned}$$

□

Suppose by contradiction that there is a stopping rule such that

$$\mathbb{P}\left(\tau_2(T,\mathcal{D}_1,\mathcal{D}_2) > \frac{n^{1/2}\log\log(1/d)^{1/2}}{Cd^2}\right) \leq \frac{1}{16}\,, \tag{2.238}$$

whenever $d = \mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2) > 0$. Let $\varepsilon_1 = 1/3$, we construct recursively $T_k = \left\lceil \frac{n^{1/2}\log\log(1/\varepsilon_k)^{1/2}}{C\varepsilon_k^2}\right\rceil = \frac{C'\sqrt{n}}{\varepsilon_{k+1}^2}$ where C and C' are constants defined later. For each integer j, we take $m_j \sim \mathrm{Poi}(j)$. Let U_n the uniform distribution and $\mathcal{D}_k$ a uniformly chosen distribution where $\mathcal{D}_{k,i} = \frac{1\pm 2\varepsilon_k}{n}$ with probability $1/2$ each. With the work of [DK16] (Section 3), we can show that $\mathrm{KL}(U_n^{\times m_j}\times \mathcal{D}_k^{\times m_j}, U_n^{\times m_j}\times U_n^{\times m_j}) \leq C''\frac{j^2\varepsilon_k^4}{n}$ where C'' is a constant. Since $\mathrm{TV}(U_n,\mathcal{D}_k) = \varepsilon_k > 0$, $\mathbb{P}\left(\tau_2(T,U_n,\mathcal{D}_k) > T_k\right) \leq 1/16$. Let E_k be the event that the stopping rule decides that the distributions are not equal between T_{k-1} and T_k. We have $\mathbb{P}\left(\tau_2(T,U_n,\mathcal{D}_k) \leq T_{k-1}\right) \leq 1/3$ since otherwise Lemma 2.6.4 implies:

$$-\frac{1}{3}\log\left(3\mathbb{P}\left(\tau_2(T,U_n,U_n)\leq T_{k-1}\right)\right) - \frac{1}{e} \leq \mathrm{KL}(U_n^{\times m_{T_{k-1}}}\times \mathcal{D}_k^{\times m_{T_{k-1}}}, U_n^{\times m_{T_{k-1}}}\times U_n^{\times m_{T_{k-1}}}) \tag{2.239}$$

$$\leq C''\frac{T_{k-1}^2\varepsilon_k^4}{n} \leq C''C'\,, \tag{2.240}$$

thus

$$\mathbb{P}\left(\tau_2(T,U_n,U_n)\leq T_{k-1}\right) \geq e^{-3C''C'-3/e}/3 > 0.1, \tag{2.241}$$

for good choice of C' and this contradicts the fact the the stopping rule is infinite with a probability at least 0.9. The stopping rule is 0.1 correct so $\mathbb{P}\left(\tau_2(T,U_n,\mathcal{D}_k) < +\infty\right) \geq 0.9$ then

$$\mathbb{P}\left(T_{k-1} < \tau_2(T,U_n,\mathcal{D}_k) \leq T_k\right) \geq 0.9 - 1/3 - 1/16 > 0.5. \tag{2.242}$$

The same inequalities for the Kullback-Leibler's divergence as above permits to deduce:

$$1 \geq \sum_{k\geq 1}\mathbb{P}\left(T_{k-1} < \tau_2(T,U_n,U_n)\leq T_k\right) \geq \sum_{k\geq 1}\frac{1}{3}e^{-3C''T_k^2\varepsilon_k^4/n - 3/e} \tag{2.243}$$

$$\geq \sum_{k\geq 1}\frac{1}{3e^2}e^{-3C''/C^2\log\log(1/\varepsilon_k)} \geq \sum_{k\geq 1}\frac{1}{3e^2}\frac{1}{\sqrt{\log(1/\varepsilon_k)}} \tag{2.244}$$

where we choose C such that $3C''/C^2 = 1/2$. But the latter sum is divergent because if we denote $a_k = \log(1/\varepsilon_k)$, we have $a_{k+1} \leq a_k + \frac{1}{4}\log\log a_k + \mathcal{O}(1)$ thus $a_k = \mathcal{O}(k\log\log k)$ therefore $\frac{1}{\sqrt{\log(1/\varepsilon_k)}} \geq \frac{c}{k}$ which is divergent. □

Remark 2.6.1. *We note that we can transform the batch algorithms to sequential ones using the doubling search technique. For instance, we can use the algorithm of [DGKPP20] as a black box and test sequentially for $1 \leq t \leq \log(1/\varepsilon)$ whether $\mathcal{D}_1 = \mathcal{D}_2$ or $\mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2) > 2^{-t}$ with a probability of failure no more than $\delta_t = \delta/t^2$. If at some step the batch-algorithm rejects we reject and halt, otherwise we accept. Note that one could think, at first sight, that this reduction even permits to estimate $\mathrm{TV}(\mathcal{D}_1,\mathcal{D}_2)$; it is however not the case, since we cannot ensure for different distributions of TV distance less than ε that the proposed algorithm will respond with the right answer (the black box algorithm can return any hypothesis H_1 or H_2 if the TV distance is strictly between 0 and ε) hence the doubling*

search algorithm (as described) cannot be used for tolerant testing. On the other hand, the doubling search method can lead to the same order of sample complexity as Theorem 2.6.2 for testing closeness problem. Nevertheless, when it comes to multiplicative constants, the two algorithms appear to have significantly different behaviours. In all experiments the actual sample complexity of Algorithm 4 appears to be better by an important constant factor. This does not show off in the bounds (since the multiplicative constants are not known), but this can be understood at least when the TV *distance d satisfies $\varepsilon < 2^{-k-1} < d = 2^{-k}(1-\eta) < 2^{-k}$, when the doubling search obviously requires ≈ 4 times more samples than necessary. Actually, even without this discretization effect, the difference is significant. This sub-optimality of the doubling search algorithm can be seen clearly for small alphabets where we can characterize the sample complexity to the constant: We gain a factor 4 using our approach while using doubling search algorithm requires up to 4 times more than the batch sample complexity when the* TV *distance is strictly greater than ε.*

2.7 General lower bounds and their proofs

In this section we present lower bounds for testing identity and testing closeness in the general case of $n \geq 2$ and provide the proofs of the lower bounds presented in this chapter.

2.7.1 Lower bound for testing identity in the general case $n \geq 2$

We first provide and prove a lower bound result for testing identity.

Lemma 2.7.1. *Let $\mathcal{D}$ be a known distribution on $[n]$. Let T a stopping rule for testing identity: $\mathcal{D}' = \mathcal{D}$ vs $\mathrm{TV}(\mathcal{D}', \mathcal{D}) > \varepsilon$ with an error probability δ. Let τ_1 and τ_2 the associated stopping times. We have*

- $\mathbb{E}(\tau_1(T, \mathcal{D})) \geq \frac{\log(1/3\delta)}{\inf_{\mathcal{D}'' s.t.\ \mathrm{TV}(\mathcal{D}'', \mathcal{D}) > \varepsilon} \mathrm{KL}(\mathcal{D}, \mathcal{D}'')}$ *if* $\mathcal{D}' = \mathcal{D}$.
- $\mathbb{E}(\tau_2(T, \mathcal{D}')) \geq \frac{\log(1/3\delta)}{\mathrm{KL}(\mathcal{D}', \mathcal{D})}$ *if* $\mathrm{TV}(\mathcal{D}', \mathcal{D}) > \varepsilon$.

Proof. We consider the two different cases $\mathcal{D}' = \mathcal{D}$ and $\mathrm{TV}(\mathcal{D}', \mathcal{D}) > \varepsilon$.

The case $\mathcal{D}' = \mathcal{D}$. We denote by $\mathbb{P}_{\mathcal{D}}$ the probability distribution on $[n]^{\mathbb{N}}$ with independent marginals X_i of distribution $\mathcal{D}$. Let $Z = (X_1, \ldots, X_{\tau_1})$ and $\mathcal{D}''$ be a distribution such that $\mathrm{TV}(\mathcal{D}'', \mathcal{D}) > \varepsilon$. The data processing property of the Kullback-Leibler divergence implies

$$\mathrm{KL}\left(\mathbb{P}^Z_{\mathcal{D}}, \mathbb{P}^Z_{\mathcal{D}''}\right) \geq \mathrm{KL}\left(\mathbb{P}_{\mathcal{D}}(\tau_1 < \infty), \mathbb{P}_{\mathcal{D}''}(\tau_1 < \infty)\right) . \tag{2.245}$$

But $\mathbb{P}_{\mathcal{D}}(\tau_1 < \infty) \geq 1 - \delta$ and $\mathbb{P}_{\mathcal{D}''}(\tau_1 < \infty) \leq \delta$. Moreover, $x \mapsto \mathrm{KL}(x, y)$ is increasing on $(y, 1)$ and $y \mapsto \mathrm{KL}(x, y)$ is decreasing on $(0, x)$ hence $\mathrm{KL}\left(\mathbb{P}_X(E), \mathbb{P}_Y(E)\right) \geq \mathrm{KL}(1-\delta, \delta)$. Tensorization property and Wald's lemma (Lemma 1.2.1) lead to

$$\mathrm{KL}\left(\mathbb{P}^Z_{\mathcal{D}}, \mathbb{P}^Z_{\mathcal{D}''}\right) = \mathbb{E}(\tau_1(T, \mathcal{D}))\, \mathrm{KL}(\mathcal{D}, \mathcal{D}'') . \tag{2.246}$$

The inequality Equation (2.245) becomes

$$\mathbb{E}(\tau_1(T, \mathcal{D})\, \mathrm{KL}(\mathcal{D}, \mathcal{D}'') \geq \mathrm{KL}(1-\delta, \delta) \geq \log(1/3\delta) , \tag{2.247}$$

which is valid for all distribution $\mathcal{D}''$, consequently

$$\mathbb{E}(\tau_1(T, \mathcal{D}) \geq \frac{\log(1/3\delta)}{\inf_{\mathcal{D}'' : \mathrm{TV}(\mathcal{D}, \mathcal{D}'') > \varepsilon} \mathrm{KL}(\mathcal{D}, \mathcal{D}'')} . \tag{2.248}$$

The case $\mathrm{TV}(\mathcal{D}', \mathcal{D}) > \varepsilon$**.** With similar notations and techniques we find for $Z = (X_1, \dots, X_{\tau_2})$

$$\begin{aligned}
\mathbb{E}(\tau_2(T, \mathcal{D}')\,\mathrm{KL}(\mathcal{D}', \mathcal{D}) &= \mathrm{KL}\left(\mathbb{P}^Z_{\mathcal{D}'}, \mathbb{P}^Z_{\mathcal{D}}\right) && (2.249)\\
&\geq \mathrm{KL}\left(\mathbb{P}_{\mathcal{D}'}(\tau_2 < \infty), \mathbb{P}_{\mathcal{D}}(\tau_2 < \infty)\right) && (2.250)\\
&\geq \mathrm{KL}(1-\delta, \delta) && (2.251)\\
&\geq \log(1/3\delta)\ . && (2.252)
\end{aligned}$$

Finally we can deduce

$$\mathbb{E}(\tau_2(T, \mathcal{D}')) \geq \frac{\log(1/3\delta)}{\mathrm{KL}(\mathcal{D}', \mathcal{D})}\ . \tag{2.253}$$

□

2.7.2 Lower bound for testing closeness in the general case $n \geq 2$

We propose the following lower bounds for testing closeness in general case

Lemma 2.7.2. *Let T a stopping rule for testing $\mathcal{D}_1 = \mathcal{D}_2$ vs $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$ with an error probability δ. Let τ_1 and τ_2 the associated stopping times. We have*

- $\mathbb{E}(\tau_1(T, \mathcal{D}_1, \mathcal{D}_2)) \geq \frac{\log(1/3\delta)}{\inf_{\mathcal{D}'_{1,2}{}^{s.t.}\ \mathrm{TV}(\mathcal{D}'_1, \mathcal{D}'_2) > \varepsilon} \mathrm{KL}(\mathcal{D}_1, \mathcal{D}'_1) + \mathrm{KL}(\mathcal{D}_2, \mathcal{D}'_2)}$ *if* $\mathcal{D}_1 = \mathcal{D}_2$.
- $\mathbb{E}(\tau_2(T, \mathcal{D}_1, \mathcal{D}_2)) \geq \frac{\log(1/3\delta)}{\inf_{\mathcal{D}} \mathrm{KL}(\mathcal{D}_1, \mathcal{D}) + \mathrm{KL}(\mathcal{D}_2, \mathcal{D})}$ *if* $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$.

Proof. Similarly as in the previous proof, we consider the two different cases $\mathcal{D}' = \mathcal{D}$ and $\mathrm{TV}(\mathcal{D}', \mathcal{D}) > \varepsilon$.

The case $\mathcal{D}_1 = \mathcal{D}_2$**.** We denote by $\mathbb{P}_{\mathcal{D}_1, \mathcal{D}_2}$ the probability distribution on $([n] \times [n])^{\mathbb{N}}$ with independent marginals (X_i, Y_i) of distribution $\mathcal{D}_1 \times \mathcal{D}_2$. Let $Z = (X_1, Y_1 \dots, X_{\tau_1}, Y_{\tau_1})$. Let $\mathcal{D}'_1, \mathcal{D}'_2$ be two distributions such that $\mathrm{TV}(\mathcal{D}'_1, \mathcal{D}'_2) > \varepsilon$. Data processing property of Kullback-Leibler's divergence implies

$$\mathrm{KL}\left(\mathbb{P}^Z_{\mathcal{D}_1, \mathcal{D}_2}, \mathbb{P}^Z_{\mathcal{D}'_1, \mathcal{D}'_2}\right) \geq \mathrm{KL}\left(\mathbb{P}_{\mathcal{D}_1, \mathcal{D}_2}(\tau_1 < \infty), \mathbb{P}_{\mathcal{D}'_1, \mathcal{D}'_2}(\tau_1 < \infty)\right)\ . \tag{2.254}$$

By definition of τ_1 we have $\mathbb{P}_{\mathcal{D}_1, \mathcal{D}_2}(\tau_1 < \infty) \geq 1 - \delta$ and $\mathbb{P}_{\mathcal{D}'_1, \mathcal{D}'_2}(\tau_1 < \infty) \leq \delta$. Tensorization property and Wald's lemma (Lemma 1.2.1) lead to

$$\mathrm{KL}\left(\mathbb{P}^Z_{\mathcal{D}_1, \mathcal{D}_2}, \mathbb{P}^Z_{\mathcal{D}'_1, \mathcal{D}'_2}\right) = \mathbb{E}(\tau_1(T, \mathcal{D}_1, \mathcal{D}_1))\,\mathrm{KL}(\mathcal{D}_1, \mathcal{D}'_1) + \mathbb{E}(\tau_1(T, \mathcal{D}_1, \mathcal{D}_2))\,\mathrm{KL}(\mathcal{D}_2, \mathcal{D}'_2)\ . \tag{2.255}$$

The inequality Equation (2.254) becomes

$$\mathbb{E}(\tau_1(T, \mathcal{D}_1, \mathcal{D}_2))\,\mathrm{KL}(\mathcal{D}_1, \mathcal{D}'_1) + \mathbb{E}(\tau_1(T, \mathcal{D}_1, \mathcal{D}_2))\,\mathrm{KL}(\mathcal{D}_2, \mathcal{D}'_2) \geq \mathrm{KL}(1-\delta, \delta) \geq \log(1/3\delta)\ , \tag{2.256}$$

which is valid for all distribution $\mathcal{D}'_1$ and $\mathcal{D}'_2$ such that $\mathrm{TV}(\mathcal{D}'_1, \mathcal{D}'_2) > \varepsilon$, consequently

$$\mathbb{E}(\tau_1(T, \mathcal{D}_1, \mathcal{D}_2)) \geq \frac{\log(1/3\delta)}{\inf_{\mathcal{D}'_{1,2}\text{s.t. }\ \mathrm{TV}(\mathcal{D}'_1, \mathcal{D}'_2) > \varepsilon} \mathrm{KL}(\mathcal{D}_1, \mathcal{D}'_1) + \mathrm{KL}(\mathcal{D}_2, \mathcal{D}'_2)}\ . \tag{2.257}$$

The case $\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) > \varepsilon$**.** Likewise we prove for $Z = (X_1, Y_1 \ldots, X_{\tau_2}, Y_{\tau_2})$ and $\mathcal{D}$ a distribution on $[n]$.

$$\mathbb{E}(\tau_2(T, \mathcal{D}_1, \mathcal{D}_2)) \,\mathrm{KL}(\mathcal{D}_1, \mathcal{D}) + \mathbb{E}(\tau_2(T, \mathcal{D}_1, \mathcal{D}_2)) \,\mathrm{KL}(\mathcal{D}_2, \mathcal{D}) = \mathrm{KL}\left(\mathbb{P}^Z_{\mathcal{D}_1,\mathcal{D}_2}, \mathbb{P}^Z_{\mathcal{D},\mathcal{D}}\right) \quad (2.258)$$

$$\geq \mathrm{KL}\left(\mathbb{P}_{\mathcal{D}_1,\mathcal{D}_2}(\tau_2 < \infty), \mathbb{P}_{\mathcal{D},\mathcal{D}}(\tau_2 < \infty)\right) \geq \mathrm{KL}(1-\delta, \delta) \geq \log(1/3\delta)\ . \quad (2.259)$$

which is valid for all distribution $\mathcal{D}$, consequently

$$\mathbb{E}(\tau_2(T, \mathcal{D}_1, \mathcal{D}_2)) \geq \frac{\log(1/3\delta)}{\inf_{\mathcal{D}} \mathrm{KL}(\mathcal{D}_1, \mathcal{D}) + \mathrm{KL}(\mathcal{D}_2, \mathcal{D})}. \quad (2.260)$$

□

2.8 Conclusion

We have provided a tight analysis of the complexity of testing identity and closeness for small n, where the importance of sequential procedures is clearly exhibited.
For the general case, we proposed tight algorithms for testing identity and closeness where the complexity can depend on the actual TV distance between the two distributions. We note that for some specific families of distributions the improvement can be much more than the general one. This is the case of distributions concentrated in small sets which can be tested rapidly by sequential strategies.

Chapter 3

On adaptivity in quantum testing

3.1 Introduction

We consider the hypothesis selection problem using independent measurements, where the tester is asked to determine the hypothesis set containing the unknown quantum state ρ with high probability. This problem is ubiquitous in the quantum learning theory literature, and several variants are considered: testing identity [OW15; BCL20; CLO22], testing closeness [Yu20], binary hypothesis testing [HP91], [ACMTBMAV07; NS09], composite quantum hypothesis testing [BDKSSSS05]. If the tester is limited to independent measurements, the problem is very related to classical testing problems. Indeed, on the one hand, every classical testing problem on discrete distributions can be cast into a quantum testing problem by taking diagonal quantum states corresponding to the discrete distributions. Measuring these quantum states is equivalent to sampling from the classical distributions. On the other hand, the quantum hypothesis selection problem can be seen as a bandit problem (see e.g. [GK19; LHT22b; BLT23]). Born's rule defines exactly the classical distribution of the reward when pulling a particular arm (performing a measurement). Note that these probability distributions are not arbitrary: they are governed by the unknown quantum state. This connection leads to an important question: Can sequential strategies outperform non-sequential ones for some hypothesis selection problem with independent measurements? In other words, if the tester is allowed to choose the measurement device at a given step depending on the previous observations, would it require fewer copies of the unknown quantum state?

Moreover, measurements come with a considerable cost, so we would like to reduce the number of measurements. Besides using entangled measurements which require a large quantum memory, one idea is to focus on independent measurements and try to adapt the new devices according to the accumulated information given by the previous outcomes.

Classically, sequential strategies prove to have an advantage over non-sequential ones for instance for binary hypothesis testing problems (see [Wal45]), testing continuous distributions (see [ZZSE16; BR15]), testing identity and closeness problems with small alphabet size (see Chapter 2). This speedup comes, mainly, from the fact that a sequential algorithm can make comparisons at each step and can respond earlier once it has the enough confidence. However, sequential strategies in the quantum setting have not only the capacity to choose the stopping time, but also to change the measurement devices adaptively. We expect then a larger gap between sequential and non-sequential strategies. To avoid confusion, *sequential* strategies can choose the stopping time according to the previous observations and thus they have random stopping times, while *adaptive*

strategies are allowed to adapt their measurement devices at each step according to past observations. With these definitions, a strategy can be sequential and adaptive, sequential and non-adaptive, non-sequential and adaptive, or non-sequential and non-adaptive (see Section 1.3.4). When we do not specify whether the strategy is non-sequential or sequential (resp. non-adaptive or adaptive), it can be either and the statement remains true.

On the other hand, non-adaptive strategies have been shown to be optimal for many interesting quantum testing problems, including testing identity by [CHLL22], purity testing and shadow tomography by [CCHL22], tomography by [CHLLS22]. These works suggest that adaptive/sequential strategies cannot outperform non-adaptive non-sequential ones. The goal of the chapter is to show the contrary: there are some situations where sequential or adaptive strategies require fewer measurements than non-adaptive non-sequential ones.

Let d the dimension of quantum states, $\varepsilon > 0$ the precision parameter and $\delta \in (0, 1/2)$ the error probability.

Contributions When the number of hypotheses m is equal to 2 and the hypotheses are simple (i.e., only one possible state), we can precisely characterize the optimal worst case complexity for non-sequential and sequential strategies. We show that sequential strategies outperform non-sequential ones by a factor 4. For the lower bounds, we show how to reduce this problem to the classical testing identity problem, then apply the lower bounds of Chapter 2. For the sequential upper bound, we design stopping rules inspired by time uniform concentration inequalities.

Moreover, we show that sequential algorithms can adapt to the actual difficulty for the testing mixedness and testing closeness problems. For this, we show a lower bound on the TV-distance between the probability distributions after measurement depending on the actual 1-norm between the quantum states (see Lemma 3.3.3). This inequality helps to reduce quantum testing to classical testing at the cost of a factor $1/\sqrt{d}$ and can be useful for other applications.

For a number of hypotheses $m \geq 2$, we prove a separation between adaptive and non-adaptive strategies for a specific problem. The learner has the information that the unknown quantum state can be diagonalised in a basis amongst m known orthonormal bases and would like to approximate it. We show that this problem can be solved by adaptive algorithms using $\mathcal{O}(d \log(m)/\varepsilon^2)$ copies of ρ. On the other hand, every non-adaptive algorithm solving this problem will require $\Omega(\min\{md/\log(m)\varepsilon^2, d^2/\varepsilon^2\})$ copies of ρ. The upper bounds follows from the shadow tomography algorithm of [HKP20]. For the lower bounds, we construct an ε-separated family of quantum states close to the maximally mixed state $(\mathbb{I}/d)$ and use it to encode a message from $[me^{\Omega(d)}]$. A learning algorithm can be used to decode this message with the same success probability. Hence, the encoder and decoder should share at least $\Omega(\log(m) + d)$ bits of information (Fano's inequality [Fan61]). On the other hand, after each step, we show that the correlation between the encoder and decoder can only increase by at most $\mathcal{O}(\varepsilon^2 \log(m)/m + \varepsilon^2/d)$ bits for non-adaptive strategies and it can only increase by at most $\mathcal{O}(\varepsilon^2)$ bits for adaptive strategies. We obtain an improvement by a factor d or $m/\log(m)$ for non-adaptive strategies by exploiting the randomness in the construction and the independence of the observations at different steps.

Related work Quantum testing identity using entangled measurements is well understood [OW15] and [BOW19]: it is known that $\Theta(d/\varepsilon^2)$ copies are necessary and sufficient. For independent measurements, it starts with the work of [BCL20] where we have two

different lower bounds for testing mixedness problem using independent adaptive and non-adaptive measurements. This result is generalized for general testing identity to some quantum state σ by [CLO22]. Recently [CHLL22] show that adaptive algorithms cannot significantly outperform non-adaptive ones neither for testing mixedness nor testing identity.

If entangled measurements are allowed, the quantum hypothesis selection problem can be solved using $\text{poly}(\log(m))$ copies of ρ (see [BO21]). This poly-logarithmic complexity in m can be explained by the fact that $\rho^{\otimes N}$ can be reused after measurement. In contrast, this is not possible using independent measurement for which the state collapses after performing the measurement. In general, the quantum hypothesis selection problem, where each hypothesis contains only one quantum state, is highly related to the shadow tomography problem where the learner is asked to uniformly approximate the expected values $\{\text{Tr}(\rho O_i)\}_{i\in[m]}$ of m known observables $\{O_i\}_{i\in[m]}$ by measuring the unknown quantum state ρ. A popular algorithm for the shadow tomography problem is given by [HKP20] and uses at most $\mathcal{O}(\log(m)d/\varepsilon^2)$ non-sequential non-adaptive independent measurements. On the other hand, independent adaptive strategies are shown to be useless for shadow tomography (and purity testing) by [CCHL22].

Moreover, sequential adaptive strategies have been used by [LTT22] (see [LHT22a] for quantum channel discrimination) to achieve the optimal rates given by the quantum relative entropy for both type I and type II errors at the same time for binary hypothesis testing problem using entangled measurements.

Adaptive strategies have been considered for testing quantum channels in [HHLW10; PLLP19; SHW22]. In particular, [HHLW10] and [SHW22] provide examples for which adaptive strategies outperform non-adaptive ones for testing quantum channels. We note that for channels, one has the possibility to adapt the *input* of the channel to the previous observations, but this is not the case for testing states. As such, it is more challenging to find a separation between adaptive and non-adaptive strategies for testing quantum states than it is for channels.

Finally, for the tomography problem, [CHLLS22] shows that adaptive independent strategies cannot beat non-sequential non-adaptive ones and thus need at least $\Omega(d^3/\varepsilon^2)$ copies to learn the quantum state ρ. However, it is unclear whether adaptivity can help for learning restricted families of states such as graph states [OT22]. On the other hand, sequential strategies were used for (online) state tomography by [KF15; YFT19; SMPE+22; RYTR22].

3.2 Preliminaries

Throughout the chapter, d is the dimension of the quantum states. An observable is a Hermitian matrix O satisfying $O \succcurlyeq 0$ and $\mathbb{I} - O \succcurlyeq 0$ where $\mathbb{I}$ is the identity matrix.

All the problems discussed in this chapter are special cases of the general hypothesis selection problem. Given an unknown quantum state $\rho \in \mathbb{C}^{d\times d}$ and m hypothesis classes $\{H_i\}_{i\in[m]}$, the learner is asked to find one of the hypothesis classes containing ρ with high probability. Formally, we have the promise that at least one of the following assertions is satisfied:

$$\rho \in H_1, \rho \in H_2, \ldots, \rho \in H_m. \tag{3.1}$$

An algorithm $\mathcal{A}$ is δ-correct for this problem if it verifies the following property:

$$\forall i \in [m] : \rho \notin H_i \implies \mathbb{P}(\mathcal{A} = i) \leq \delta. \tag{3.2}$$

The difference between quantum and classical testing is that in the quantum case we have the possibility to choose a measurement (given by positive operators summing to the identity). If the quantum states are restricted to be diagonal, we may assume the measurement is always the same and so the problem becomes a classical testing problem (see Lemma 3.3.1).

The quantum state ρ is unknown, but the learner can extract classical information from it by performing a measurement. The way the unknown quantum state ρ is measured is important and can lead to different results about the number of copies needed for this task. Recall that for testing states, we can distinguish between two types of measurements depending on the considered Hilbert space:

1. An **entangled measurement** is given by a POVM on the Hilbert space $\mathcal{H} = (\mathbb{C}^d)^{\otimes N}$, where N is the number of copies available of the quantum state ρ. We can measure the whole state $\rho^{\otimes N}$ at once. An interesting POVM related to the observable O on $\mathbb{C}^d$ is given by $\mathcal{M}(O) = \{M_k\}_{0 \leq k \leq N}$ where $M_k = \sum_{x \in \{0,1\}^N, |x|=k} O^{x_1} \otimes \cdots \otimes O^{x_N}$. Measuring $\rho^{\otimes N}$ with the POVM $\mathcal{M}(O)$ outputs a sample from the binomial distribution $\mathrm{Bin}(n, \mathrm{Tr}(\rho O))$.

2. An **independent (or incoherent) measurement** is given by a sequence of POVMs $\{\mathcal{M}_t\}_{t \in [N]}$, each of them acts on the Hilbert space $\mathcal{H} = \mathbb{C}^d$. In this case, we measure at step t the quantum state ρ using the POVM $\mathcal{M}_t$. For instance, for an observable O, measuring ρ with the POVM $\mathcal{M}(O) = (\mathbb{I} - O, O)$ outputs a sample from the Bernoulli distribution $\mathrm{Bern}(\mathrm{Tr}(\rho O))$. If the POVMs $\{\mathcal{M}_t\}_t$ are fixed in advance (i.e., do not depend on the outcomes of the previous measurements), the procedure is called *non-adaptive*; when $\mathcal{M}_t$ can be chosen depending on the results of the previous measurements with the $\{\mathcal{M}_s\}_{s<t}$, we call it an *adaptive* algorithm. If the number of measurements is not fixed beforehand and can be chosen as a function of the previous measurement outcomes, the algorithm is called *sequential* and has a random stopping time N. In this case, the *expected copy complexity* of the procedure is $\mathbb{E}(N)$. Otherwise, the algorithm has a fixed number of measurements N and is called *non-sequential*.

In this chapter, we focus on algorithms using independent measurements and our goal is to assess the potential improvement of sequential/adaptive algorithms over non-adaptive non-sequential ones.

3.3 Sequential improvement for problems involving two hypotheses

In this section, we focus on sequential algorithms for problems having only two hypotheses ($m = 2$), which can be simple or not.

3.3.1 Provable constant improvement of sequential strategies

The simplest case for hypothesis selection problem with $m = 2$ corresponds to hypothesis sets containing only one known quantum state. Formally, the learner would like

to distinguish two hypothesis: $H_1 = \{\sigma_1\}$ and $H_2 = \{\sigma_2\}$. We want to characterize the exact number of copies the learner needs to solve this problem using sequential and non-sequential independent measurements.

Non-adaptive strategies

The tester knows the quantum states σ_1 and σ_2 and can hence calculate the actual 1-norm between them, denoted by $\varepsilon = \|\sigma_1 - \sigma_2\|_{\mathrm{Tr}}$. The optimal POVM to distinguish between σ_1 and σ_2 is thus given by $\mathcal{M} = (\mathbb{I} - O, O)$ (Holevo-Helstrom theorem, see [Wat18]) where $0 \preccurlyeq O \preccurlyeq \mathbb{I}$ satisfies

$$\varepsilon = \|\sigma_1 - \sigma_2\|_{\mathrm{Tr}} = \mathrm{Tr}((\sigma_1 - \sigma_2)O). \tag{3.3}$$

Let $X_1, \ldots, X_N$ be the outcomes of measuring ρ by the POVM $\mathcal{M}$. By Born's rule, they follow the Bernoulli distribution of parameter $\mathrm{Tr}(\rho O)$. Let S be the statistic given by the difference between the empirical mean and the actual mean under H_2: $S = \frac{1}{N}\sum_{i=1}^{N} X_i - \mathrm{Tr}(\sigma_2 O)$. Its expected value is $\mathrm{Tr}((\rho - \sigma_2)O)$ which is ε under H_1 and 0 under H_2. The learner can measure ρ a sufficient number of times, compare the statistic S with $\varepsilon/2$ and decide accordingly: If $S \geq \varepsilon/2$ it accepts H_1, otherwise it accepts H_2. Following the Chernoff-Hoeffding inequality ([Hoe63]), the sufficient number of measurement for the learner to be δ-correct is

$$\max\left\{\frac{\log(1/\delta)}{\mathrm{KL}(\mathrm{Tr}((\sigma_1+\sigma_2)O)/2\|\mathrm{Tr}(\sigma_1 O))}, \frac{\log(1/\delta)}{\mathrm{KL}(\mathrm{Tr}((\sigma_1+\sigma_2)O)/2\|\mathrm{Tr}(\sigma_2 O))}\right\} \leq \frac{2\log(1/\delta)}{\varepsilon^2}. \tag{3.4}$$

The latter inequality follows from Pinsker's inequality ([FHT03]). Note that this previous upper bound is optimal in the worst case setting where we fix ε and take the infimum over all σ_1 and σ_2 satisfying $\|\sigma_1 - \sigma_2\|_{\mathrm{Tr}} = \varepsilon$. This first result is summarized in the following proposition:

Proposition 3.3.1. *There is a non sequential algorithm for testing $H_1 : \rho = \sigma_1$ vs $H_2 : \rho = \sigma_2$ using $\frac{2\log(1/\delta)}{\varepsilon^2}$ measurements. Moreover, there exists two quantum states σ_1 and σ_2 satisfying $\|\sigma_1 - \sigma_2\|_{\mathrm{Tr}} = \varepsilon$ so that every non sequential algorithm distinguishing between $H_1 : \rho = \sigma_1$ and $H_2 : \rho = \sigma_2$ with high probability needs an equivalent of $\frac{2\log(1/\delta)}{\varepsilon^2}$ measurements.*

Proof. The correctness of the batch algorithm presented above can be done using Chernoff-Hoeffding inequality, if $\rho = \sigma_1$ the error probability can be upper bounded as follows:

$$\begin{aligned}
\mathbb{P}\left(S - \mathrm{Tr}(\sigma_2 O) \leq \varepsilon/2\right) &= \mathbb{P}\left(S - \mathrm{Tr}(\sigma_1 O) \leq \varepsilon/2 - \varepsilon\right) && (3.5)\\
&\leq \mathbb{P}\left(S - \mathrm{Tr}(\sigma_1 O) \leq -\varepsilon/2\right) && (3.6)\\
&\leq \exp(-N\,\mathrm{KL}(\mathrm{Tr}(\sigma_1 O) - \varepsilon/2\|\mathrm{Tr}(\sigma_1 O))). && (3.7)
\end{aligned}$$

On the other hand, if $\rho = \sigma_2$:

$$\mathbb{P}\left(S - \mathrm{Tr}(\sigma_2 O) \geq \varepsilon/2\right) \leq \exp(-N\,\mathrm{KL}(\mathrm{Tr}(\sigma_2 O) + \varepsilon/2\|\mathrm{Tr}(\sigma_2 O))). \tag{3.8}$$

Therefore to ensure that the batch algorithm is δ-correct we need N to satisfy

$$N \geq \max\left\{\frac{\log(1/\delta)}{\mathrm{KL}(\mathrm{Tr}(\sigma_1 O) - \varepsilon/2\|\mathrm{Tr}(\sigma_1 O))}, \frac{\log(1/\delta)}{\mathrm{KL}(\mathrm{Tr}(\sigma_2 O) + \varepsilon/2\|\mathrm{Tr}(\sigma_2 O))}\right\}. \tag{3.9}$$

Moreover by Pinsker's inequality ([FHT03]), the right hand side is upper bounded by:

$$\max\left\{\frac{\log(1/\delta)}{\mathrm{KL}(\mathrm{Tr}(\sigma_1 O)-\varepsilon/2\|\mathrm{Tr}(\sigma_1 O))},\frac{\log(1/\delta)}{\mathrm{KL}(\mathrm{Tr}(\sigma_2 O)+\varepsilon/2\|\mathrm{Tr}(\sigma_2 O))}\right\}\leq\frac{2\log(1/\delta)}{\varepsilon^2}. \tag{3.10}$$

For the lower bound, we construct two quantum states, $\sigma_1=\mathbb{I}_2/d$ and $\sigma_2=\mathrm{diag}((1+2\varepsilon)/2,(1-2\varepsilon)/2)=\mathbb{I}/2+\varepsilon O$ where $O=\mathrm{diag}(1,-1)$. The reduction to classical testing can be proven using the following lemma on measurements of diagonal quantum states.

Lemma 3.3.1. *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two discrete distributions and ρ_1 and ρ_2 their corresponding diagonal quantum states. Let $\mathcal{M}$ be a POVM. Measuring the quantum state ρ_1 (resp. ρ_2) with the POVM $\mathcal{M}$ can be seen as post-processing (independent of the quantum states) of samples from the distribution $\mathcal{D}_1$ (resp. $\mathcal{D}_2$).*

Proof. Let $\mathcal{M}=\{M^i\}_{i\in[k]}$. For each $i\in[k]$, we can write

$$M^i=\sum_{x,y}M^i_{x,y}\left|x\right\rangle\left\langle y\right|. \tag{3.11}$$

By Born's rule, the probability distribution of the outcomes of the measurement of ρ by the POVM $\mathcal{M}$ is:

$$\mathcal{M}(\rho)=\{\mathrm{Tr}(\rho M^i)\}_{i\in d}=\left\{\mathrm{Tr}\left(\sum_x\mathcal{D}_x\left|x\right\rangle\!\left\langle x\right|\sum_{x,y}M^i_{x,y}\left|x\right\rangle\left\langle y\right|\right)\right\}_{i\in d} \tag{3.12}$$

$$=\left\{\sum_x M^i_{x,x}\mathcal{D}_x\right\}_{i\in d}=\mathcal{P}\mathcal{D}, \tag{3.13}$$

where $\mathcal{P}=(M^i_{x,x})_{i,x}$ is a stochastic matrix. Indeed, $M^i\succcurlyeq 0$ implies $M^i_{x,x}=\left\langle x\right|M^i\left|x\right\rangle\geq 0$ and $\sum_i M^i=\mathbb{I}$ implies

$$\sum_i M^i_{x,x}=\sum_i\left\langle x\right|M^i\left|x\right\rangle=\left\langle x|x\right\rangle=1. \tag{3.14}$$

□

Now we move to show how the reduction works for entangled strategies. We have $\sigma_1^{\otimes N}=\frac{\mathbb{I}}{2^N}$ and $\sigma_2^{\otimes N}=\frac{1}{2^N}\,\mathrm{diag}\left((1+2\varepsilon)^{|i|}(1-2\varepsilon)^{N-|i|}\right)_{i\in\{0,1\}^N}$ where $|i|=i_1+\cdots+i_N$. By Lemma 3.3.1, measuring the quantum states $\sigma_1^{\otimes N}$ (resp. $\sigma_2^{\otimes N}$) can be seen as post-processing of samples from the distribution $\mathcal{D}_1=\{1/2^N\}_{i\in\{0,1\}^N}$ (resp. $\mathcal{D}_2=\left\{(1/2-\varepsilon)^{|i|}(1/2+\varepsilon)^{N-|i|}\right\}_{i\in\{0,1\}^N}$). Observe that a sample $i=(i_1,\ldots,i_N)\sim\mathcal{D}_1$ is given by N i.i.d. random variables $\{i_k\sim\mathrm{Bern}(1/2)\}_{k\in[N]}$. Similarly, a sample $i=(i_1,\ldots,i_N)\sim\mathcal{D}_2$ is given by N i.i.d. random variables $\{i_k\sim\mathrm{Bern}(1/2-\varepsilon)\}_{k\in[N]}$. Therefore, distinguishing σ_1 from σ_2 using N entangled copies can be reduced to testing $\mathrm{Bern}(1/2)$ vs $\mathrm{Bern}(1/2-\varepsilon)$ using N samples. Once the reduction to classical testing identity is done, we can invoke the lower bound of Theorem 2.3.1.

□

Sequential strategies

If we allow the tester to adapt the measurements and choose its stopping time according to previous observations, it can outperform (in expectation) every non-sequential algorithms by a factor 4. Precisely, it can be proven that an expected number of measurements equivalent to $\frac{\log(1/\delta)}{2\varepsilon^2}$ is sufficient to distinguish between $H_1 : \rho = \sigma_1$ and $H_2 : \rho = \sigma_2$ with probability at least $1-\delta$. We use again the optimal POVM $\mathcal{M}$ defined in Equation (3.3) to distinguish between σ_1 and σ_2. Let $X_1, \dots, X_t \sim \mathrm{Bern}(\mathrm{Tr}(\rho O))$ the outcomes of measuring ρ by the POVM $\mathcal{M}$. Let $S_t = \frac{1}{t}\sum_{i=1}^{t} X_i$ the empirical mean until the time t. Contrary to the algorithm described in the previous subsection, a sequential algorithm can make comparisons at each time t until the tester is confident enough to answer the correct answer H_1 or H_2. Under H_1, the statistic S_t has an expected value $\mathrm{Tr}(\sigma_1 O)$. On the other hand, under H_1, the statistic S_t has an expected value $\mathrm{Tr}(\sigma_2 O)$. These expected values are known to the tester, so it can compare at each time the statistic S_t with two thresholds: $\mathrm{Tr}(\sigma_1 O) - \phi(\delta, t)$ and $\mathrm{Tr}(\sigma_2 O) + \phi(\delta, t)$ where $\phi(\delta, t)^2 = \log\left(\frac{2t(t+1)}{\delta}\right)/2t$. If $S_t \leq \mathrm{Tr}(\sigma_1 O) - \phi(\delta, t)$, the tester can answer H_2 confidently. Similarly, it would answer H_1 if $S_t \geq \mathrm{Tr}(\sigma_2 O) + \phi(\delta, t)$. However if none of these inequalities is verified it does not answer and makes a new measurement, and so forth until the regions defined by the thresholds coincide. The idea of comparing the statistic with time dependent thresholds has been previously used for classical testing identity and closeness in Chapter 2, where it is proven that in expectation this algorithm outperform the non sequential one by a factor 4. We adapt this result to the quantum setting in the following proposition.

Proposition 3.3.2. *There is a sequential algorithm for testing $H_1 : \rho = \sigma_1$ vs $H_2 : \rho = \sigma_2$ using an expected number of measurements:*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}(N)}{\log(1/\delta)} \leq \frac{1}{2\varepsilon^2}. \tag{3.15}$$

Moreover, for small enough ε, there are two quantum states σ_1 and σ_2 satisfying $\|\sigma_1 - \sigma_2\|_{\mathrm{Tr}} = \varepsilon$ so that every sequential algorithm distinguishing between $H_1 : \rho = \sigma_1$ and $H_2 : \rho = \sigma_2$ with high probability needs, in expectation, an equivalent a number $\frac{\log(1/\delta)}{2\varepsilon^2}$ of measurements.

Note that we can obtain non asymptotic upper bound depending on σ_1 and σ_2 by carefully choosing the thresholds of the algorithm. For the sake of simplicity, we prefer to present the asymptotic worst case complexities that can be easily compared. The correctness of the algorithm presented here is proved using the following time uniform concentration inequality which is an application of union bound and Hoeffding's inequality ([Hoe63]):

$$\mathbb{P}\left(\exists t \geq 1 : |S_t - \mathbb{E}(S_t)| > \phi(\delta, t)\right) \leq \delta. \tag{3.16}$$

The lower bound follows from the previous proof's reduction and the lower bound on the expected number of samples for testing uniform using sequential algorithms: $\mathrm{Bern}(1/2)$ vs $\mathrm{Bern}(1/2 \pm \varepsilon)$ (see Lemma 2.3.1). We note that [LTT22; LHT22a] have also established an advantage of sequential strategies over non-adaptive ones in terms of the error exponents.

Proof. We use a similar approach as in Chapter 2:

Correctness. Let's start by showing that the algorithm presented above is δ-correct. To this end, we need a time uniform concentration inequality which can be obtained by Hoeffding inequality along with the union bound, recall that $S_t = \frac{1}{t}\sum_{i=1}^{t} X_i$ and $X_i \sim \text{Bern}(\text{Tr}(\rho O))$:

$$\mathbb{P}\left(\exists t \geq 1 : |S_t - \mathbb{E}\left(S_t\right)| > \phi(\delta, t)\right) \leq \sum_{t\geq 1} \mathbb{P}\left(|S_t - \mathbb{E}\left(S_t\right)| > \phi(\delta, t)\right) \tag{3.17}$$

$$\leq \sum_{t\geq 1} \exp(-2t\phi(\delta, t)^2) \tag{3.18}$$

$$\leq \sum_{t\geq 1} \frac{\delta}{t(t+1)} \tag{3.19}$$

$$\leq \delta. \tag{3.20}$$

Complexity. To obtain an upper bound on the complexity, we use the following lemma:

Lemma 3.3.2. *N a random variable taking values in $\mathbb{N}^*$, we have for all $k \in \mathbb{N}^*$*

$$\mathbb{E}(N) \leq k + \sum_{t\geq k} \mathbb{P}(N \geq t)\ . \tag{3.21}$$

This inequality can be proved by writing $\mathbb{E}(N) = \sum_{t\geq 1} \mathbb{P}(N \geq t)$ then upper bounding the first $k-1$ terms by 1.
Let $\alpha \in (0,1)$ and k the smallest integer so that for all $t \geq k : \phi(\delta, t) \leq \alpha\varepsilon$. We focus only on the case $\rho = \sigma_1$ (the other being similar), the expected stopping time of the algorithm can be controlled as follows:

$$\mathbb{E}\left(N\right) \leq k + \sum_{t\geq k} \mathbb{P}(N \geq t) \leq k + \sum_{t\geq k} \mathbb{P}(S_{t-1} < \text{Tr}(\sigma_2 O) + \phi(\delta, t-1)) \tag{3.22}$$

$$\leq k + \sum_{t\geq k-1} \mathbb{P}(S_t - \text{Tr}(\sigma_1 O) < -\varepsilon + \alpha\varepsilon) \leq k + \sum_{t\geq k-1} \mathbb{P}(S_t - \text{Tr}(\sigma_1 O) < -(1-\alpha)\varepsilon) \tag{3.23}$$

$$\leq k + \sum_{t\geq k-1} 2\exp(-2t(1-\alpha)^2\varepsilon^2) \leq k + \frac{2\exp(-2(k-1)(1-\alpha)^2\varepsilon^2)}{1-\exp(-2(1-\alpha)^2\varepsilon^2)} \tag{3.24}$$

$$\leq k + \frac{2\exp(-2(k-1)(1-\alpha)^2\varepsilon^2)}{(1-\alpha)^2\varepsilon^2}. \tag{3.25}$$

On the other hand we have $\phi(\delta, k) \leq \alpha\varepsilon$ and $\phi(\delta, k-1) \geq \alpha\varepsilon$ so

$$\log\left(\frac{(k-1)k}{\delta}\right) \geq 2(k-1)\alpha^2\varepsilon^2. \tag{3.26}$$

Therefore:

$$k - 1 \leq \frac{\log(1/\delta)}{2\alpha^2\varepsilon^2} + 2\frac{\log(\log(1/\delta)/(\alpha\varepsilon)^2)}{\alpha^2\varepsilon^2}. \tag{3.27}$$

Hence:

$$\frac{\mathbb{E}\left(N\right)}{\log(1/\delta)} \leq \frac{1}{2\alpha^2\varepsilon^2} + 2\frac{\log(\log(1/\delta)/(\alpha\varepsilon)^2)}{\log(1/\delta)\alpha^2\varepsilon^2} + \frac{1}{\log(1/\delta)} + \frac{2\exp(-2(k-1)(1-\alpha)^2\varepsilon^2)}{\log(1/\delta)(1-\alpha)^2\varepsilon^2}, \tag{3.28}$$

and by taking $\delta \to 0$, then $\alpha \to 1$ we obtain:

$$\limsup_{\delta \to 0} \frac{\mathbb{E}(N)}{\log(1/\delta)} \leq \frac{1}{2\varepsilon^2}. \tag{3.29}$$

□

3.3.2 Sequential strategies adapt on the actual difficulty of the problem without prior knowledge

In this section, we change the previous setting by letting the second hypothesis be multiple. Precisely, we consider the problem of testing identity with $H_1 = \{\mathbb{I}/d\}$ and $H_2 = \{\rho : \|\rho - \mathbb{I}/d\|_{\mathrm{Tr}} \geq \varepsilon\}$ where ε is a positive parameter. [CHLL22] has proved that the optimal adaptive copy complexity is $\Theta(d^{3/2}/\varepsilon^2)$. We show that while adaptive algorithms cannot improve the copy complexity, sequential algorithms can be used to adapt to the actual difficulty of the problem. Mainly we show the following result:

Proposition 3.3.3. *There is a sequential algorithm for testing identity problem using a number of measurements satisfying:*

$$\mathbb{E}(N) = \mathcal{O}\left(\min\left\{\frac{d^{3/2}\log(1/\delta)}{\varepsilon^2}, \frac{d^{1/2}\log(1/\delta)}{\|\rho - \mathbb{I}/d\|_2^2}\right\}\right). \tag{3.30}$$

In particular, the expected copy complexity can be reduced to $\mathcal{O}(rd^{1/2}\log(1/\delta))$ if the quantum state ρ has low rank $r \leq d/2$ or $\mathcal{O}\left(\frac{rd^{1/2}\log(1/\delta)}{\|\rho - \mathbb{I}/d\|_{\mathrm{Tr}}^2}\right)$ if the trace-less matrix $\rho - \mathbb{I}/d$ has low rank r even if the algorithm does not have any information about these ranks (see Section 3.3.2). The algorithm uses random measurements and a time-dependent stopping rule. Since we have already sequential algorithms for the classical testing identity problem, it is sufficient to show how to reduce the quantum problem to the classical one. For a POVM $\mathcal{M}$ and a quantum state ρ, let $\rho(\mathcal{M})$ denotes the classical probability distribution $\{\mathrm{Tr}(\rho M_i)\}_i$. The following lemma captures the main ingredient of the reduction:

Lemma 3.3.3. *For all $\delta > 0$, let $l = 3072\log(2/\delta)$ and $U^1, U^2, \ldots, U^l \in \mathbb{C}^{d\times d}$ be* Haar-*random unitary matrices of columns $\{|U_i^j\rangle\}_{1\leq i\leq d, 1\leq j\leq l}$, $\mathcal{M} = \{\frac{1}{l}|U_i^j\rangle\langle U_i^j|\}_{i,j}$ is a POVM and for all quantum states ρ and σ we have with a probability at least $1-\delta$:*

$$\mathrm{TV}(\rho(\mathcal{M}), \sigma(\mathcal{M})) \geq \frac{\|\rho - \mathbb{I}/d\|_2}{16} \geq \frac{\|\rho - \sigma\|_{\mathrm{Tr}}}{16\sqrt{r}}, \tag{3.31}$$

where r is the rank of $(\rho - \sigma)$.

A similar statement can be found in [MWW09] where the authors analyze the uniform POVM and a POVM defined by a spherical 4-designs. However, for our reduction, it is important to minimize the number of outcomes of the POVM. Performing measurment with a random basis is well known in the quantum learning literature (see e.g., [EFHKPVZ23]).

Proof. Let $\xi = \rho - \sigma$, we have $U|e_i\rangle = |U_i\rangle$ and we use Weingarten calculus [Gu13;

CMS12] to calculate

$$\mathbb{E}\left[\langle U_i|\,\xi\,|U_i\rangle^2\right] = \mathbb{E}\left[\langle U_i|\,\xi\,|U_i\rangle\,\langle U_i|\,\xi\,|U_i\rangle\right] = \mathbb{E}\left[\mathrm{Tr}(\xi\,|U_i\rangle\langle U_i|\,\xi\,|U_i\rangle\langle U_i|)\right] \tag{3.32}$$

$$= \mathbb{E}\left[\mathrm{Tr}(\xi U\,|e_i\rangle\langle e_i|\,U^*\xi U\,|e_i\rangle\langle e_i|\,U^*)\right] = \mathbb{E}\left[\mathrm{Tr}(U^*\xi U\,|e_i\rangle\langle e_i|\,U^*\xi U\,|e_i\rangle\langle e_i|)\right] \tag{3.33}$$

$$= \sum_{\alpha,\beta\in\mathfrak{S}_2} \mathrm{Wg}(\beta\alpha^{-1}, d)\mathrm{Tr}_{\beta^{-1}}(\xi,\xi)\mathrm{Tr}_\alpha(|e_i\rangle\langle e_i|\,,|e_i\rangle\langle e_i|) = \frac{1}{d(d+1)}\mathrm{Tr}(\xi^2). \tag{3.34}$$

Similarly

$$\mathbb{E}\left[\langle U_i|\,\xi\,|U_i\rangle^4\right] = \mathbb{E}\left[\langle U_i|\,\xi\,|U_i\rangle\,\langle U_i|\,\xi\,|U_i\rangle\,\langle U_i|\,\xi\,|U_i\rangle\,\langle U_i|\,\xi\,|U_i\rangle\right] \tag{3.35}$$

$$= \mathbb{E}\left[\mathrm{Tr}(\xi\,|U_i\rangle\langle U_i|\,\xi\,|U_i\rangle\langle U_i|\,\xi\,|U_i\rangle\langle U_i|\,\xi\,|U_i\rangle\langle U_i|)\right] \tag{3.36}$$

$$= \mathbb{E}\left[\mathrm{Tr}(\xi U\,|e_i\rangle\langle e_i|\,U^*\xi U\,|e_i\rangle\langle e_i|\,U^*\xi U\,|e_i\rangle\langle e_i|\,U^*\xi U\,|e_i\rangle\langle e_i|\,U^*)\right] \tag{3.37}$$

$$= \mathbb{E}\left[\mathrm{Tr}(U^*\xi U\,|e_i\rangle\langle e_i|\,U^*\xi U\,|e_i\rangle\langle e_i|\,U^*\xi U\,|e_i\rangle\langle e_i|\,U^*\xi U\,|e_i\rangle\langle e_i|)\right] \tag{3.38}$$

$$= \sum_{\alpha,\beta\in\mathfrak{S}_4} \mathrm{Wg}(\beta\alpha^{-1}, d)\mathrm{Tr}_{\beta^{-1}}(\xi,\xi,\xi,\xi)\mathrm{Tr}_\alpha(|e_i\rangle\langle e_i|\,,|e_i\rangle\langle e_i|\,,|e_i\rangle\langle e_i|\,,|e_i\rangle\langle e_i|) \tag{3.39}$$

$$= \frac{1}{d(d+1)(d+2)(d+3)}(6\mathrm{Tr}(\xi^2)^2 + 6\mathrm{Tr}(\xi^4)). \tag{3.40}$$

$$\leq \frac{12}{d(d+1)(d+2)(d+3)}\mathrm{Tr}(\xi^2)^2. \tag{3.41}$$

We can now conclude by Hölder's inequality:

$$2\mathbb{E}\left[\mathrm{TV}(\rho(\mathcal{M}),\sigma(\mathcal{M}))\right] = \sum_{i=1}^{d}\mathbb{E}\left[|\,\langle U_i|\,\xi\,|U_i\rangle\,|\right] \geq \sum_{i=1}^{d}\sqrt{\frac{\left(\mathbb{E}\left[\langle U_i|\,\xi\,|U_i\rangle^2\right]\right)^3}{\mathbb{E}\left[\langle U_i|\,\xi\,|U_i\rangle^4\right]}} \tag{3.42}$$

$$\geq \sum_{i=1}^{d}\sqrt{\frac{(d^{-1}(d+1)^{-1}\mathrm{Tr}(\xi^2))^3}{12d^{-1}(d+1)^{-1}(d+2)^{-1}(d+3)^{-1}\mathrm{Tr}(\xi^2)^2}} \tag{3.43}$$

$$\geq \sum_{i=1}^{d}\frac{\sqrt{\mathrm{Tr}(\xi^2)}}{4d} \geq \frac{1}{4}\sqrt{\mathrm{Tr}(\rho-\sigma)^2}. \tag{3.44}$$

Let $f(U) = \mathrm{TV}(\rho(\mathcal{M}),\sigma(\mathcal{M}))$, we first show that f is Lipschitz by using the triangular

and Cauchy Schwarz inequality:

$$2|f(U)-f(V)| = \left| \sum_{1\le i\le d, 1\le j\le l} \frac{1}{l} |\mathrm{Tr}(|U_i^j\rangle\langle U_i^j|\,\xi)| - |\mathrm{Tr}(|V_i^j\rangle\langle V_i^j|\,\xi)| \right| \tag{3.45}$$

$$\le \sum_{1\le i\le d, 1\le j\le l} \frac{1}{l} \left|\mathrm{Tr}((|U_i^j\rangle\langle U_i^j| - |V_i^j\rangle\langle V_i^j|)\xi)\right| \tag{3.46}$$

$$\le \sum_{1\le i\le d, 1\le j\le l} \frac{1}{l} \sqrt{\mathrm{Tr}(\xi^2)}\sqrt{\mathrm{Tr}((|U_i^j\rangle\langle U_i^j| - |V_i^j\rangle\langle V_i^j|)^2)} \tag{3.47}$$

$$\le \sqrt{\frac{d}{l}}\sqrt{\mathrm{Tr}(\xi^2)}\sqrt{\sum_{1\le i\le d, 1\le j\le l} \mathrm{Tr}((|U_i^j\rangle\langle U_i^j| - |V_i^j\rangle\langle V_i^j|)^2)} \tag{3.48}$$

$$\le \sqrt{\frac{d}{l}}\sqrt{\mathrm{Tr}(\xi^2)}\sqrt{\sum_{1\le j\le l} \mathrm{Tr}((U^j - V^j)^2)} \tag{3.49}$$

$$\le \sqrt{\frac{d}{l}}\sqrt{\mathrm{Tr}(\xi^2)}\|U-V\|_{2,\mathrm{HS}}, \tag{3.50}$$

hence f is $L = \sqrt{\frac{d}{4l}}\sqrt{\mathrm{Tr}(\xi^2)}$-Lipschitz. Therefore by Theorem 1.4.2:

$$\mathbb{P}\left(|f(U) - \mathbb{E}\left(f(U)\right)| > \frac{1}{16}\sqrt{\mathrm{Tr}(\xi^2)}\right) \le e^{-\frac{d\mathrm{Tr}(\xi^2)}{48\cdot 16^2 L^2}} = e^{-l/3072} = \delta/2, \tag{3.51}$$

for $l = 3072\log(2/\delta)$. Finally with high probability (at least $1-\delta/2$) we have

$$\mathrm{TV}(\rho(\mathcal{M}),\sigma(\mathcal{M})) \ge \mathbb{E}\left(\mathrm{TV}(\rho(\mathcal{M}),\sigma(\mathcal{M}))\right) - |\,\mathrm{TV}(\rho(\mathcal{M}),\sigma(\mathcal{M})) - \mathbb{E}\left(\mathrm{TV}(\rho(\mathcal{M}),\sigma(\mathcal{M}))\right)| \tag{3.52}$$

$$\ge \frac{1}{8}\sqrt{\mathrm{Tr}(\xi^2)} - \frac{1}{16}\sqrt{\mathrm{Tr}(\xi^2)} \ge \frac{1}{16}\sqrt{\mathrm{Tr}(\xi^2)} \ge \frac{1}{16}\frac{\|\rho-\sigma\|_{\mathrm{Tr}}}{\sqrt{r}}, \tag{3.53}$$

where r is the rank of $(\rho - \mathbb{I}/d)$. □

Let $\eta = \|\rho - \mathbb{I}/d\|_2$. Under the alternative hypothesis H_2, the TV distance between P and U_n can be lower bounded by $\mathrm{TV}(P,U_n) \ge \frac{1}{16}\|\rho - \mathbb{I}/d\|_2$. So Lemma 3.3.3 gives a POVM for which our problem reduces to testing identity: $P = U_n$ vs $\mathrm{TV}(P,U_n) \ge \frac{\eta}{16}$ with high probability, where $n = \frac{1}{4}d\log(2/\delta)$ and $P = \mathcal{M}(\rho)$. Therefore we can apply the classical testing uniform result of [DGPP17] to obtain a non sequential algorithm for testing identity in the 2-norm with a copy complexity:

$$\mathcal{O}\left(\frac{\sqrt{d}\log(1/\delta)}{\eta^2}\right). \tag{3.54}$$

Moreover we can apply the sequential upper bound for classical testing uniform (see Algorithm 3 and Theorem 2.5.1) to obtain a sequential testing identity in the 1-norm with an expected copy complexity:

$$\tilde{\mathcal{O}}\left(\frac{d^{3/2}\log(1/\delta)}{\max\{\varepsilon^2, d\|\rho - \mathbb{I}/d\|_2^2\}}\right). \tag{3.55}$$

Algorithm 5 Testing whether $\rho = \mathbb{I}/d$ or $\|\rho - \mathbb{I}/d\|_2 \geq \eta$ with an error probability at most δ.

$l = 3072 \log(2/\delta)$.
Sample $U^1, U^2, \ldots, U^l \in \mathbb{C}^{d\times d}$ from Haar(d) distribution.
Let $\left\{\left|U_i^j\right\rangle\right\}_{1\leq i\leq d, 1\leq j\leq l}$ be the columns of the unitary matrices $U^1, U^2, \ldots, U^l$
Measure the quantum state ρ using the POVM $\mathcal{M} = \left\{\frac{1}{l}\left|U_i^j\right\rangle\!\left\langle U_i^j\right|\right\}_{i,j}$ and observe $\mathcal{O}\left(\sqrt{d}\log(1/\delta)/\eta^2\right)$ samples from $\rho(\mathcal{M})$.
Test whether $h_0 : \rho(\mathcal{M}) = U_{ld}$ or $h_1 : \mathrm{TV}(\rho(\mathcal{M}), U_{ld}) \geq \eta/16$ using the testing identity of discrete distributions of [DGPP17], with an error probability δ, and answer accordingly.

A matching lower bound can be obtained in the worst case setting where we are interested only in the parameters d, ε and $\|\rho - \mathbb{I}/d\|_{\mathrm{Tr}}$. This can be done using Markov's inequality to transform the algorithm to a deterministic-time one then invoking the lower bound of [CHLL22]: Any adaptive algorithm for testing identity would require $\Omega(d^{3/2}/\varepsilon^2)$ copies of ρ.
Once we have the lower bound on the TV distance between the distributions obtained after performing the measurements, we can deduce upper bounds on sequential algorithms for testing identity depending on the rank of ρ or $\rho - \mathbb{I}/d$.

Dependence in the rank of $\rho - \mathbb{I}/d$: From the previous lower bound on the TV-distance, we can achieve an upper bound using the sequential tester of uniform (Theorem 2.5.1):

$$\tilde{\mathcal{O}}\left(\min\left\{\frac{n^{1/2}\log(1/\delta)^{1/2}}{(\max\{\varepsilon/\sqrt{d}, \|\rho - \mathbb{I}/d\|_2\})^2}, \frac{\log(1/\delta)}{(\max\{\varepsilon/\sqrt{d}, \|\rho - \mathbb{I}/d\|_2\})^2}\right\}\right) \tag{3.56}$$

$$= \tilde{\mathcal{O}}\left(\frac{d^{3/2}\log(1/\delta)}{\max\{\varepsilon^2, d\|\rho - \mathbb{I}/d\|_2^2\}}\right) = \tilde{\mathcal{O}}\left(\min\left\{\frac{d^{3/2}\log(1/\delta)}{\varepsilon^2}, \frac{rd^{1/2}\log(1/\delta)}{\|\rho - \mathbb{I}/d\|_{\mathrm{Tr}}^2}\right\}\right), \tag{3.57}$$

where r is the rank of $(\rho - \mathbb{I}/d)$ and we use Cauchy Schwarz to obtain the latter inequality.

Dependence in the rank of ρ The proof of Lemma 3.3.3 permits to deduce that with high probability:

$$\mathrm{TV}(P, U_n) \geq \frac{1}{16}\|\rho - \mathbb{I}/d\|_2 \geq \frac{1}{16}\sqrt{\sum_{i=1}^{r}\left(\lambda_i - \frac{1}{d}\right)^2 + \frac{d-r}{d^2}} \tag{3.58}$$

$$\geq \frac{1}{16}\sqrt{\sum_{i=1}^{r}\lambda_i^2 - \frac{1}{d}} \geq \frac{1}{16}\sqrt{\frac{1}{r} - \frac{1}{d}} \geq \frac{1}{16}\sqrt{\frac{1}{2r}}, \tag{3.59}$$

where r is the rank of ρ supposed to be less than $d/2$ and we use Cauchy Schwarz inequality. Therefore we can test whether $\rho = \sigma$ or $\|\rho - \sigma\|_{\mathrm{Tr}} > \varepsilon$ with probability at least $1 - \delta$ using

$$\tilde{\mathcal{O}}\left(\frac{d^{1/2}\log(1/\delta)}{\left(\max\left\{\varepsilon/\sqrt{d}, 1/\sqrt{2r}\right\}\right)^2}\right) = \tilde{\mathcal{O}}\left(\min\left\{\frac{d^{3/2}\log(1/\delta)}{\varepsilon^2}, rd^{1/2}\log(1/\delta)\right\}\right) \tag{3.60}$$

copies of ρ.

Observe that, using Lemma 3.3.3 and the sequential tester of closeness for classical distributions (see Algorithm 4 and Theorem 2.6.2), we obtain the same copy complexity for testing closeness (i.e., testing $\rho = \sigma$ vs $\|\rho - \sigma\|_{\mathrm{Tr}} \geq \varepsilon$ where we can measure the unknown quantum states ρ and σ) as for testing identity. This is different from the classical case where testing identity can be done with much less copies than testing closeness (see [DGPP17] and [DGKPP20]).

3.4 Provable separation between adaptive and non-adaptive strategies

In this section, we focus only on adaptive algorithms meaning that the number of measurements is always deterministic.

We construct a problem for which we have a separation between adaptive and non-adaptive algorithms. Let $\{\sigma_1, \ldots, \sigma_m\}$ be a set of ε-separated known quantum states. The unknown quantum state ρ is $\varepsilon/3$-close to (at most) one of the quantum states $\sigma_{i^\star} \in \{\sigma_1, \ldots, \sigma_m\}$ and has the same diagonalisation basis than $\sigma_{i^\star}$. We aim to learn the quantum state ρ to within $\varepsilon/10$ with high probability. Formally, the goal is to design an algorithm that measures a number of copies of ρ and returns a quantum state $\tilde{\rho}$ (an $\varepsilon/10$-approximation of ρ) such that with probability (the randomness comes from the measurements and possibly the algorithm) at least $1 - \delta$:

$$\|\tilde{\rho} - \rho\|_{\mathrm{Tr}} \leq \varepsilon/10. \tag{3.61}$$

The problem described above is not a hypothesis selection problem in the strict sense of the term. However it is equivalent to the following hypothesis selection problem which has the same order of copy complexity. For $i \in [m]$, let $\sigma_i = \sum \lambda_k |\phi_k^i\rangle\langle\phi_k^i|$ and $\{\sigma_{i,j}\}_{j\in[M]}$ an $\varepsilon/10$-covering of the set $\{\rho = \sum_k \mu_k |\phi_k^i\rangle\langle\phi_k^i| : 2\,\mathrm{TV}(\lambda, \mu) \leq \varepsilon/3\}$. Our problem is equivalent to the hypothesis selection problem for $\{H_{i,j} = \{B(\sigma_{i,j}, \varepsilon/10)\} \cap \{\rho : \rho\sigma_{i,j} = \sigma_{i,j}\rho\}\}_{i\in[m],j\in[M]}$. For simplicity, we use the first formulation of the problem and refer to it as (P).

3.4.1 Upper bound

In this section, we present an adaptive algorithm for the problem (P) achieving a copy complexity strictly less than the lower bound which holds for all non-adaptive algorithms. The first step is to determine with high probability the closest quantum state $\sigma_{i^\star}$ to ρ, then it remains to approximate ρ by measuring it in its basis of diagonalization.

Adaptive strategies.

The quantum state $\sigma_{i^\star}$ has the property to minimize the 1-norm between ρ and $\{\sigma_i\}_i$, so it is natural to take the state minimizing the statistics of expected value roughly $\max_{i,j} \mathrm{Tr}\, O_{i,j}(\rho - \sigma_l)$ for $l \in [m]$. To do this, we need to approximate $\mathrm{Tr}\, \rho O_{i,j}$ for all $i \neq j$. We can use the classical shadow tomography algorithm of [HKP20] to predict all these events using a few number of copies of ρ:

Theorem 3.4.1 ([HKP20]). *Let $(O_1, \ldots, O_m)$ be a tuple of observables. There is an*

algorithm using non-adaptive independent measurements requiring:

$$N = \mathcal{O}\left(\frac{d\log(m/\delta)}{\varepsilon^2}\right) \tag{3.62}$$

copies of ρ to predict $\mathrm{Tr}(\rho O_i)$ to within ε-error for all $i = 1, \dots, m$ with at most an error probability of δ.

Once we find the quantum state $\sigma_{i^\star}$, we know the basis of diagonalization of ρ and we can learn the eigenvalues using $\mathcal{O}(d/\varepsilon^2)$ independent copies. The algorithm is summarized in Algorithm 6. This algorithm is δ-correct. We need to show that with probability at

Algorithm 6 Hypothesis selection problem (P).

Require: $N = \mathcal{O}(d\log(m/\delta)/\varepsilon^2)$ independent measurement on ρ and m quantum states $\sigma_1, \dots, \sigma_m$.

Ensure: Two quantum states $\sigma_{i^\star}$ and $\tilde{\rho}$ satisfying with a probability at least $1-\delta$: $\|\sigma_{i^\star} - \rho\|_{\mathrm{Tr}} \leq \varepsilon/3$ and $\|\tilde{\rho} - \rho\|_{\mathrm{Tr}} \leq \varepsilon/10$.
For all $i \neq j \in [m]$, let $O_{i,j}$ an observable satisfying $\|\sigma_i - \sigma_j\|_{\mathrm{Tr}} = \mathrm{Tr}\ O_{i,j}(\sigma_i - \sigma_j)$.
For all $i \neq j \in [m]$, let $\mu_{i,j}$ an $\varepsilon/10$ approximation of $\mathrm{Tr}(\rho O_{i,j})$ given by classical shadow tomography of [HKP20].
Let $k^\star = \arg\min_l \max_{i,j} \mu_{i,j} - \mathrm{Tr}(\sigma_l O_{i,j})$.
Let $\mathcal{M} = \{|\phi_i\rangle\langle\phi_i|\}_{i\in[d]}$ the POVM corresponding to the basis of diagonalisation of $\sigma_{k^\star}$.

Measure ρ independently $M = 200\log(2^{d+2}/\delta)/\varepsilon^2$ times using the POVM $\mathcal{M}$ and denote the outcomes $\{E_i\}_{1\leq i\leq M}$.
Return $\tilde{\rho} = \sum_{i\in[d]} \left(\frac{\sum_{j\in[M]} \mathbf{1}_{E_j=i}}{M}\right) |\phi_i\rangle\langle\phi_i|$.

least $1-\delta/2$, Algorithm 6 finds the closest quantum state $\sigma_{i^\star}$ to ρ.

Lemma 3.4.1. *For all $i \neq j \in [m]$, let $\mu_{i,j}$ an $\varepsilon/10$ approximation of $\mathrm{Tr}(\rho O_{i,j})$ given by classical shadow tomography of [HKP20]. Let $k^\star = \arg\min_l \max_{i,j} \mu_{i,j} - \mathrm{Tr}(\sigma_l O_{i,j})$. We have with at least a probability $1-\delta/2$:*

$$\|\rho - \sigma_{k^\star}\|_{\mathrm{Tr}} \leq \varepsilon/3. \tag{3.63}$$

Proof. Classical shadow tomography of [HKP20] permits to have the following approximations

$$\forall i \neq j \in [m] : |\mu_{i,j} - \mathrm{Tr}(\rho O_{i,j})| \leq \varepsilon/10, \tag{3.64}$$

with a probability at least $1-\delta/2$ using only $N = \mathcal{O}(d\log(m)/\varepsilon^2)$ copies of ρ.
Let $\sigma_{i^\star}$ the closest quantum state to ρ. We want to prove that with high probability $k^\star = i^\star$. We have for all $l \neq i^\star$: $\|\sigma_{i^\star} - \sigma_l\|_{\mathrm{Tr}} > \varepsilon$ hence:

$$\max_{i,j} \mu_{i,j} - \mathrm{Tr}(\sigma_l O_{i,j}) \geq \mu_{i^\star,l} - \mathrm{Tr}(\sigma_l O_{i^\star,l}) \geq \mathrm{Tr}(\rho O_{i^\star,l}) - \mathrm{Tr}(\sigma_l O_{i^\star,l}) - \varepsilon/10 \tag{3.65}$$

$$\geq \mathrm{Tr}(\sigma_{i^\star} O_{i^\star,l}) - \mathrm{Tr}(\sigma_l O_{i^\star,l}) + \mathrm{Tr}(\rho O_{i^\star,l}) - \mathrm{Tr}(\sigma_{i^\star} O_{i^\star,l}) - \varepsilon/10 \tag{3.66}$$

$$\geq \|\sigma_{i^\star} - \sigma_l\|_{\mathrm{Tr}} - \|\rho - \sigma_{i^\star}\|_{\mathrm{Tr}} - \varepsilon/10 \geq \varepsilon - \varepsilon/3 - \varepsilon/10 > \varepsilon/2. \tag{3.67}$$

On the other hand

$$\max_{i,j} \mu_{i,j} - \mathrm{Tr}(\sigma_{k^\star} O_{i,j}) \leq \max_{i,j} \mu_{i,j} - \mathrm{Tr}(\sigma_{i^\star} O_{i,j}) \tag{3.68}$$

$$\leq \max_{i,j} \mathrm{Tr}(\rho O_{i,j}) - \mathrm{Tr}(\sigma_{i^\star} O_{i,j}) + \varepsilon/10 \tag{3.69}$$

$$\leq \|\rho - \sigma_{i^\star}\|_{\mathrm{Tr}} + \varepsilon/10 \leq \varepsilon/3 + \varepsilon/10 < \varepsilon/2. \tag{3.70}$$

Therefore, with high probability, $k^\star$ cannot be different from $i^\star$. □

Once we know, with high probability, the closest quantum state to ρ we can read its basis and use it to to learn ρ. The following lemma indicates how to construct this approximation along with the number of copies/measurements needed for this learning task.

Lemma 3.4.2. *Let $\rho = \sum_{i=1}^d \lambda_i |\phi_i\rangle\langle\phi_i|$. Let $A_1, \ldots, A_N$ the outcomes of the measurement of ρ independently by the POVM $\mathcal{M} = \{|\phi_i\rangle\langle\phi_i|\}_i$. The quantum state*

$$\tilde{\rho} = \sum_{i=1}^d \left(\frac{\sum_{j=1}^N \mathbf{1}_{A_j=i}}{N} \right) |\phi_i\rangle\langle\phi_i| \tag{3.71}$$

is $\varepsilon/10$-close in 1-norm to ρ with a probability at least $1 - \delta/2$ if $N = 200 \log(2^{d+2}/\delta)/\varepsilon^2$.

Proof. ρ is a quantum state so it is a Hermitian matrix positive semi definite of trace 1. Hence, we can write $\rho = \sum_{i=1}^d \lambda_i |\phi_i\rangle\langle\phi_i|$ where$\{\lambda_i\}_i$ is a probability distribution and $\{\phi_i\}_i$ is an orthonormal basis. Therefore $\sum_{i=1}^d |\phi_i\rangle\langle\phi_i| = \mathbb{I}$ and $\mathcal{M}$ is a valid POVM. Measuring ρ via the POVM $\mathcal{M}$ is equivalent to sampling from the distribution $\{\mathrm{Tr}(|\phi_i\rangle\langle\phi_i| \rho\}_i = \{\sum_j \lambda_j \mathrm{Tr}(|\phi_i\rangle\langle\phi_i| |\phi_j\rangle\langle\phi_j|\}_i = \{\lambda_i\}_i$ hence

$$A_1, \ldots, A_N \overset{i.i.d.}{\sim} \{\lambda_i\}_i. \tag{3.72}$$

On the other hand ρ and $\tilde{\rho}$ have the same basis of diagonalization so the 1 norm between them is simply

$$\|\rho - \tilde{\rho}\|_{\mathrm{Tr}} = \left\| \sum_{i=1}^d \lambda_i |\phi_i\rangle\langle\phi_i| - \sum_{i=1}^d \tilde{\lambda}_i |\phi_i\rangle\langle\phi_i| \right\|_{\mathrm{Tr}} = \left\| \sum_{i=1}^d (\lambda_i - \tilde{\lambda}_i) |\phi_i\rangle\langle\phi_i| \right\|_{\mathrm{Tr}} \tag{3.73}$$

$$= \sum_{i=1}^d |\lambda_i - \tilde{\lambda}_i| = 2\,\mathrm{TV}(\lambda, \tilde{\lambda}), \tag{3.74}$$

where $\{\tilde{\lambda}_i\}_i = \left\{ \frac{\sum_{j=1}^N \mathbf{1}_{A_j=i}}{N} \right\}_i$. It is well known that the TV distance can be written as:

$$\mathrm{TV}(\lambda, \tilde{\lambda}) = \max_{B \subset [d]} (\tilde{\lambda}(B) - \lambda(B)). \tag{3.75}$$

Chernoff-Hoeffding([Hoe63]) inequality implies for all $B \subset [d]$:

$$\mathbb{P}\left(\left| \frac{\sum_{j=1}^N \mathbf{1}_{A_j \in B}}{N} - \lambda(B) \right| > \frac{\varepsilon}{20} \right) \leq 2 \exp\left(-2N \left(\frac{\varepsilon}{20} \right)^2 \right). \tag{3.76}$$

Therefore by union bound we obtain

$$\mathbb{P}\left(\|\rho-\tilde{\rho}\|_{\mathrm{Tr}}>\varepsilon/10\right)=\mathbb{P}\left(2\,\mathrm{TV}(\lambda,\tilde{\lambda})>\varepsilon/10\right) \tag{3.77}$$

$$=\mathbb{P}\left(\exists B\subset[d]:\left|\frac{\sum_{j=1}^{N}\mathbf{1}_{A_j\in B}}{N}-\lambda(B)\right|>\frac{\varepsilon}{20}\right) \tag{3.78}$$

$$\leq 2^{d+1}\exp\left(-2N\left(\frac{\varepsilon}{20}\right)^2\right). \tag{3.79}$$

Finally for $N=200\log(2^{d+1}/\delta)/\varepsilon^2$, we have with at least a probability $1-\delta$: $\|\rho-\tilde{\rho}\|_{\mathrm{Tr}}\leq\varepsilon/10$. □

Grouping the two previous Lemmas, Algorithm 6 finds the closest quantum state $\sigma_{i^\star}$ and returns an $\varepsilon/10$-approximation of ρ with a probability at least $1-(\delta/2+\delta/2)=1-\delta$.

Algorithm 6 can be split in two parts for which we independently upper bound the copy complexity. The first part relies on the shadow tomography algorithm of [HKP20] and needs a number $N_1=\mathcal{O}\left(\frac{d\log(m(m-1)/\delta)}{(\varepsilon/10)^2}\right)=\mathcal{O}\left(\frac{d\log(m/\delta)}{\varepsilon^2}\right)$ of copies of ρ. The second part requires a number $N_2=200\frac{\log(2^{d+1}/\delta)}{\varepsilon^2}$ of copies of ρ. Finally, the total copy complexity of Algorithm 6 is $N=N_1+N_2=\mathcal{O}\left(\frac{d\log(m/\delta)}{\varepsilon^2}\right)$.

Non-adaptive strategies.

We can slightly modify Algorithm 6 to have a non-adaptive algorithm for the problem (P) with independent measurements. It amounts to first measuring ρ in all the basis corresponding to the known quantum states $(\sigma_i)_i$ and preparing m approximated quantum states $(\tilde{\rho}_i)_i$. Then the tester can look for the closest quantum state $\sigma_{i^\star}$ and finally returns the approximated quantum state $\tilde{\rho}_{i^\star}$. This non-adaptive algorithm has a copy complexity $mN_2+N_1=\mathcal{O}\left(\frac{md+m\log(1/\delta)+d\log(m/\delta)}{\varepsilon^2}\right)$. This complexity is almost optimal for $m\leq d$ (see Theorem 3.4.2). However, it is no longer optimal for $m\geq d$ since $md/\varepsilon^2\geq d^2/\varepsilon^2$. In that case, we can still design an almost optimal non-adaptive algorithm as follows: for each $k\in[m]$, let $\{|\phi_i^k\rangle\}_i$ an orthonormal basis of diagonalization for σ_k. For each $k\in[m]$ and $B\subset[m]$, let $O_B^k=\sum_{i\in B}|\phi_i^k\rangle\langle\phi_i^k|$. We use the classical shadow tomography of [HKP20] to predict $(\mathrm{Tr}(\rho O_{i,j}))_{i,j\in[m]}\cup(\mathrm{Tr}(\rho O_B^k)_{k\in[m],B\subset[m]}$ to within $\varepsilon/40$ simultaneously using $\mathcal{O}(d\log(m^2+m2^d)/\varepsilon^2)=\mathcal{O}((d^2+\log(m))/\varepsilon^2)$ copies of ρ. We find the closest quantum state $\sigma_{i^\star}$ to ρ the same way as the Algorithm 6 does. Next, we look for a probability distribution $\tilde{\lambda}$ satisfying for all $B\subset[m]$: $\left|\tilde{\lambda}(B)-\mu_B^{i^\star}\right|\leq\varepsilon/40$, where $\mu_B^{i^\star}$ is the prediction of shadow tomography algorithm for $\mathrm{Tr}(\rho O_B^{i^\star})$. Such $\tilde{\lambda}$ exists since the vector λ of eigenvalues of ρ satisfies the following property:

$$\mathrm{Tr}(\rho O_B^{i^\star})=\mathrm{Tr}\left(\sum_{i\in[d]}\lambda_i|\phi_i^{i^\star}\rangle\langle\phi_i^{i^\star}|\sum_{i\in B}|\phi_i^{i^\star}\rangle\langle\phi_i^{i^\star}|\right) \tag{3.80}$$

$$=\sum_{i\in[d],j\in B}\lambda_i|\langle\phi_i^{i^\star}|\phi_j^{i^\star}\rangle|^2=\sum_{i\in B}\lambda_i=\lambda(B), \tag{3.81}$$

and $\left|\lambda(B)-\mu_B^{i^\star}\right|=\left|\mathrm{Tr}(\rho O_B^{i^\star})-\mu_B^{i^\star}\right|\leq\varepsilon/40$. We can thus return the quantum state $\tilde{\rho}=\sum_{i\in[d]}\tilde{\lambda}_i|\phi_i^{i^\star}\rangle\langle\phi_i^{i^\star}|$ as an approximation of ρ. We can verify that it is indeed an $\varepsilon/10$

approximation of ρ:

$$\|\rho - \tilde{\rho}\|_{\mathrm{Tr}} \leq \sum_{i=1}^{d} |\lambda_i - \tilde{\lambda}_i| = 2 \max_{B \subset [d]} \lambda(B) - \tilde{\lambda}(B) \tag{3.82}$$

$$\leq 2 \max_{B \subset [d]} \lambda(B) - \mu_B^{i^\star} + 2 \max_{B \subset [d]} \mu_B^{i^\star} - \tilde{\lambda}(B) \tag{3.83}$$

$$\leq 2\varepsilon/40 + 2\varepsilon/40 \leq \varepsilon/10. \tag{3.84}$$

The copy complexity of this algorithm is $\mathcal{O}((d^2 + \log(m))/\varepsilon^2$ which matches (up to logarithmic factors) the lower bound for $m \geq d$.

3.4.2 Lower bound

In this section, we derive lower bounds for the problem (P) both with adaptive and non-adaptive independent measurements.

We start with a lower bound for non-adaptive algorithms that matches the copy complexity of the algorithm presented in Section 3.4.1. For this section, we fix the error probability to $\delta = 1/3$.

Theorem 3.4.2. *There is a tuple of quantum states $(\sigma_1, \ldots, \sigma_m)$ such that any learning algorithm with non-adaptive independent measurements requires*

$$N = \Omega\left(\min\left\{\frac{md}{\log(m)\varepsilon^2}, \frac{d^2}{\varepsilon^2}\right\}\right)$$

copies of ρ to approximate ρ to at most $\varepsilon/10$ with at least a probability $2/3$.

This result with $m = d$, together with the analysis of the adaptive Algorithm 6 gives a nearly quadratic advantage for adaptive algorithms over non-adaptive ones.

Proof. We start by constructing the quantum states $(\sigma_1, \ldots, \sigma_m)$. We choose m unitary matrices $\{U_y\}_y$ chosen randomly from the Haar(d) distribution, then we choose for each unitary (orthonormal basis) random eigenvalues:

Lemma 3.4.3. *Let $m \leq \exp(d^2/3000)$. Let $\{U_y\}_{y \in [m/2]}$ be $m/2$ unitaries distributed according to the* Haar(d) *distribution. For $y \in [m/2]$, let $\sigma_y = 2\mathbb{I}/d - \sigma_{m+1-y} = U_y \Lambda U_y^\dagger$ where $\Lambda = \frac{\mathbb{I}}{d} + \mathrm{diag}\left(\{\lambda_i\}_{i \in [d]}\right) = \mathrm{diag}\left(\left\{\frac{1+(-1)^i 10\varepsilon}{d}\right\}_{i \in [d]}\right)$. We have with a probability at least $9/10$, for all $y \neq z \in [m]$:*

$$\|\sigma_y - \sigma_z\|_{\mathrm{Tr}} \geq \varepsilon. \tag{3.85}$$

Proof. Let $y \neq z \in [m/2]$ and $0 \preccurlyeq O \preccurlyeq \mathbb{I}$ satisfying $\mathrm{Tr}\,\mathrm{diag}\left(\{\lambda_i\}_{i \in [d]}\right) O = -5\varepsilon$. Let $f(U) = \mathrm{Tr}\left(U\,\mathrm{diag}\left(\{\lambda_i\}_{i \in [d]}\right) U^\dagger - \mathrm{diag}\left(\{\lambda_i\}_{i \in [d]}\right)\right) O$ where $U \sim$ Haar(d), we have $\mathbb{E}\left(f(U)\right) = -\mathrm{Tr}\,\mathrm{diag}\left(\{\lambda_i\}_{i \in [d]}\right) O = 5\varepsilon$ (see Weingarten calculus [Gu13]). The function

f is $\frac{20\varepsilon}{\sqrt{d}}$-Lipschitz. Indeed, recall that $\Lambda = \mathbb{I}/d + \operatorname{diag}\left(\{\lambda_i\}_{i\in[d]}\right)$ we have:

$$
\begin{aligned}
&|f(U) - f(V)| && (3.86)\\
&= \left|\operatorname{Tr}\left(U(\Lambda - \mathbb{I}/d)U^\dagger - (\Lambda - \mathbb{I}/d)\right) O - \operatorname{Tr}\left(V(\Lambda - \mathbb{I}/d)V^\dagger \ - (\Lambda - \mathbb{I}/d)\right) O\right| && (3.87)\\
&\leq |\operatorname{Tr}(U \operatorname{diag}\left(\{\lambda_i\}_{i\in[d]}\right) U^\dagger - V \operatorname{diag}\left(\{\lambda_i\}_{i\in[d]}\right) V^\dagger)O| && (3.88)\\
&\leq \|(U-V)\operatorname{diag}\left(\{\lambda_i\}_{i\in[d]}\right) U^\dagger\|_{\mathrm{Tr}} + \|V \operatorname{diag}\left(\{\lambda_i\}_{i\in[d]}\right)(U-V)^\dagger\|_{\mathrm{Tr}} && (3.89)\\
&\leq \|U-V\|_2 \|\operatorname{diag}\left(\{\lambda_i\}_{i\in[d]}\right) U^\dagger\|_2 + \|V \operatorname{diag}\left(\{\lambda_i\}_{i\in[d]}\right)\|_2 \|(U-V)^\dagger\|_2 && (3.90)\\
&\leq \frac{10\varepsilon}{d}(\|U-V\|_2 \|\operatorname{diag}\left(\left\{(-1)^i\right\}_{i\in[d]}\right) U^\dagger\|_2 + \|V \operatorname{diag}\left(\left\{(-1)^d\right\}_{i\in[d]}\right)\|_2 \|(U-V)^\dagger\|_2) && (3.91)\\
&\leq \frac{20\varepsilon}{\sqrt{d}}\|U-V\|_2, && (3.92)
\end{aligned}
$$

where we have used Cauchy Schwarz inequality.

Using the fact that the Haar distribution is invariant under the multiplication by a unitary and the concentration inequality of Lipschitz functions of Haar unitary matrices [MM13], the probability that the states $\{\sigma_y\}_{y\in[m/2]}$ are not ε-separated is upper bounded by

$$
\begin{aligned}
&\mathbb{P}\left(\exists y, z \in [m/2] : \|\sigma_y - \sigma_z\|_{\mathrm{Tr}} \leq \varepsilon\right) \leq \frac{m^2}{4}\mathbb{P}\left(\|\sigma_y - \sigma_z\|_{\mathrm{Tr}} \leq \varepsilon\right) && (3.93)\\
&\leq \frac{m^2}{4}\mathbb{P}\left(\|U_y \operatorname{diag}\left(\{\lambda_i\}_{i\in[d]}\right) U_y^\dagger - U_z \operatorname{diag}\left(\{\lambda_i\}_{i\in[d]}\right) U_z^\dagger\|_{\mathrm{Tr}} \leq \varepsilon\right) && (3.94)\\
&\leq \frac{m^2}{4}\mathbb{P}\left(\|U_z^\dagger U_y \operatorname{diag}\left(\{\lambda_i\}_{i\in[d]}\right) U_y^\dagger U_z - \operatorname{diag}\left(\{\lambda_i\}_{i\in[d]}\right)\|_{\mathrm{Tr}} \leq \varepsilon\right) && (3.95)\\
&\leq \frac{m^2}{4}\mathbb{P}\left(\|U \operatorname{diag}\left(\{\lambda_i\}_{i\in[d]}\right) U^\dagger - \operatorname{diag}\left(\{\lambda_i\}_{i\in[d]}\right)\|_{\mathrm{Tr}} \leq \varepsilon\right) && (3.96)\\
&\leq \frac{m^2}{4}\mathbb{P}\left(\operatorname{Tr}\left(U \operatorname{diag}\left(\{\lambda_i\}_{i\in[d]}\right) U^\dagger - \operatorname{diag}\left(\{\lambda_i\}_{i\in[d]}\right)\right) O \leq \varepsilon\right) && (3.97)\\
&\leq \frac{m^2}{4}\mathbb{P}\left(f(U) - \mathbb{E}\left(f(U)\right) \leq \varepsilon - 5\varepsilon\right) = \frac{m^2}{4}\mathbb{P}\left(\mathbb{E}\left(f(U)\right) - f(U) \geq 4\varepsilon\right) && (3.98)\\
&\leq \frac{m^2}{4}\exp\left(-\frac{-16d^2\varepsilon^2}{12\times 400\varepsilon^2}\right) \leq \frac{m^2}{4}\exp\left(-\frac{d^2}{1000}\right), && (3.99)
\end{aligned}
$$

which is smaller than $1/10$ if $m^2 \leq 2\exp(d^2/1000)/5$.

For the case when $y \in [m/2]$ and $z \in [m]\setminus[m/2]$, let $x = m+1-z \in [m/2]$ we have

$$
\begin{aligned}
\|\sigma_y - \sigma_z\|_{\mathrm{Tr}} &= \|\sigma_y - 2\mathbb{I}/d + \sigma_x\|_{\mathrm{Tr}} && (3.100)\\
&\geq \|\sigma_x - 2\mathbb{I}/d + \sigma_x\|_{\mathrm{Tr}} - \|\sigma_y - \sigma_x\|_{\mathrm{Tr}} && (3.101)\\
&= 2\|\sigma_x - \mathbb{I}/d\|_{\mathrm{Tr}} - \|\sigma_y - \sigma_x\|_{\mathrm{Tr}} \geq \varepsilon. && (3.102)
\end{aligned}
$$

Finally, for the case when $y \in [m]\setminus[m/2]$ and $z \in [m]\setminus[m/2]$, let $y' = m+1-y \in [m/2]$ and $z' = m+1-z \in [m/2]$ we have

$$
\|\sigma_y - \sigma_z\|_{\mathrm{Tr}} = \|2\mathbb{I}/d - \sigma_{y'} - 2\mathbb{I}/d + \sigma_{z'}\|_{\mathrm{Tr}} \geq \|\sigma_{y'} - \sigma_{z'}\|_{\mathrm{Tr}} \geq \varepsilon. \tag{3.103}
$$

□

We have shown how to construct the unitaries, we move to prove the existence of the eigenvalues:

Lemma 3.4.4. *There exists family of quantum states $\{\rho_{x,y}\}_{|x|\in[e^{cd}],y\in[m]}$ (where c is a universal constant) such that for each $y \in [m]$, $\{\rho_{x,y}\}_{|x|\in[e^{cd}]}$ is $\varepsilon/5$-separated and commute.*

Proof. We start by writing the eigen-decomposition of the known quantum states σ_y as

$$\sigma_y = U_y \left(\sum_{i=1}^{d} \lambda_i^y |i\rangle\langle i| \right) U_y^\dagger. \tag{3.104}$$

We claim that we can choose α_i^x to construct an $\varepsilon/5$-separated family of me^{cd} quantum states (c is a constant to be chosen later) of the form

$$\rho_{x,y} = U_y \left(\sum_{i=1}^{d} \left(\lambda_i^y + \frac{\alpha_i^x (2\varepsilon/3)}{d} \right) |i\rangle\langle i| \right) U_y^\dagger \ , \tag{3.105}$$

for $|x| \in [e^{cd}]$ and $y \in [m]$. Note that for convenience of notation, the labels x can be positive and negative. Moreover the distance between $\rho_{x,y}$ and σ_y is exactly:

$$\|\rho_{x,y} - \sigma_y\|_{\mathrm{Tr}} = \frac{\varepsilon}{3}. \tag{3.106}$$

Concretely, we look for $\{\alpha_i^x\}_{1\le i\le d, 1\le |x|\le e^{cd}/2}$ such that

1. $\alpha_i^x = \pm 1$,
2. $\alpha_i^{-x} = -\alpha_i^x$,
3. $\alpha_i^x + \alpha_{i+d/2}^x = 0$ (we suppose d is even) and
4. $\forall x \neq x' : \sum_{i=1}^{d/2} |\alpha_i^x - \alpha_i^{x'}| > d(1/2 - 1/200)$.

The third point ensures that ρ has trace 1 while the fourth one implies $\|\rho_{x,y} - \rho_{x',y}\|_{\mathrm{Tr}} > \varepsilon/3 - \varepsilon/100 > \varepsilon/5$. Starting by the simple quantum states $\rho_{1,y} = \sigma_y + \sum_{i=1}^{d/2} \frac{(2\varepsilon/3)}{d} U_y |i\rangle\langle i| U_y^\dagger - \sum_{i=d/2+1}^{d} \frac{(2\varepsilon/3)}{d} U_y |i\rangle\langle i| U_y^\dagger$ and $\rho_{-1,y} = 2\mathbb{I}/d - \rho_{1,y} = \sigma_{m+1-y} - \sum_{i=1}^{d/2} \frac{(2\varepsilon/3)}{d} U_y |i\rangle\langle i| U_y^\dagger + \sum_{i=d/2+1}^{d} \frac{(2\varepsilon/3)}{d} U_y |i\rangle\langle i| U_y^\dagger$ and we suppose that we have constructed $\mathcal{Q}$ an ε-separated family of the form described above of cardinality $M < e^{cd}$. Let $\alpha_1, \ldots, \alpha_{d/2}$ i.i.d. random variables taken values in $\{\pm 1\}$ with probability $1/2$ each. We have by Hoeffding's inequality

$$\mathbb{P}\left(\exists \rho_x \in \mathcal{Q} : \sum_{i=1}^{d/2} |\alpha_i^x - \alpha_i| \le d(1/2 - 1/200) \vee \sum_{i=1}^{d/2} |\alpha_i^{-x} - \alpha_i| \le d(1/2 - 1/200) \right) \tag{3.107}$$

$$= \mathbb{P}\left(\exists \rho_x \in \mathcal{Q} : \sum_{i=1}^{d/2} |\alpha_i^x - \alpha_i| \le d(1/2 - 1/200) \vee \sum_{i=1}^{d/2} |\alpha_i^x + \alpha_i| \le d(1/2 - 1/200) \right) \tag{3.108}$$

$$\le \frac{M}{2} \mathbb{P}\left(\sum_{i=1}^{d/2} \mathbf{1}_{\alpha_i = \alpha_i^x} > d/4 + d/400 \right) + \frac{M}{2} \mathbb{P}\left(\sum_{i=1}^{d/2} \mathbf{1}_{\alpha_i = \alpha_i^x} \le d/4 - d/400 \right) \tag{3.109}$$

$$\le M e^{-d/2000}, \tag{3.110}$$

which is strictly less than 1 if $M < e^{d/2000}$. So let's take $c = 1/2000$, we deduce that

$$\mathbb{P}\left(\forall \rho_x \in \mathcal{Q} : \sum_{i=1}^{d/2} |\alpha_i^x - \alpha_i| > d(1/2 - 1/200)\right) > 0. \tag{3.111}$$

therefore there exists some $\alpha \in \{\pm 1\}^d$ verifying the desired conditions. We can repeat this construction until $Card(\mathcal{Q}) \geq e^{cd}$.

□

We have constructed the ε-separated family of quantum states $\{\sigma_y\}_y$ an the corresponding $\varepsilon/5$-separated $\{\rho_{x,y}\}_x$ for all y, we can use tools from communication theory to deduce the lower bound (see e.g., [FGLE12; HHJWY16]). Alice encodes a message $(X, Y) \in \{1, \ldots, e^{cd}\} \times [m]$ in $\rho_{x,y}$ and sends it to Bob. To read the message, Bob tries to approximate the quantum state that he received from Alice. We suppose that Bob can approximate (up to $\varepsilon/10$ in trace norm) a state $\varepsilon/3$ close to one of $\{\sigma_y\}$ and diagonalized in the same basis of this quantum state with a probability at least $2/3$. Bob uses N copies to decode Alice's message and returns $(X', Y') \in \{1, \ldots, e^{cd}\} \times [m]$, therefore by Fano's inequality ([Fan61]) we have the following lower bound on the mutual information:

Lemma 3.4.5 (Fano)**.** *The mutual information between the encoder and the decoder can be lower bounded:*

$$\mathcal{I}(X, Y : X', Y') \geq 2/3 \log(me^{cd}) - \log(2) \geq \Omega(\log(m) + d). \tag{3.112}$$

On the other hand we can upper bound the mutual information between (X, Y) and (X', Y'). Let $I_1, \ldots, I_N$ be the outcomes of a non adaptive algorithm solving the problem (P). By using the data-processing inequality for the Kullback-Leibler divergence and the fact that every non adaptive algorithm for the problem (P) can be used as a $2/3$-correct decoder we can upper bound the mutual information as follows:

Lemma 3.4.6 (Data-processing)**.** *The mutual information between (X, Y) and (X', Y') is smaller than the mutual information between (X, Y) and $(I_1, \ldots, I_N)$:*

$$\mathcal{I}(X, Y : X', Y') \leq \mathcal{I}(X, Y : I_1, \ldots, I_N). \tag{3.113}$$

The next step is to upper bound the mutual information between (X, Y) and $(I_1, \ldots, I_N)$. This latter depends on the quantum states $\{\sigma_y\}_y$, therefore it is a random variable. We will show that with at least a probability $9/10$, it is upper bounded by an expression involving the parameters of the problem. First we start by proving the following upper bound relating the mutual information with the unitaries $\{U_y\}_y$ defining the quantum states $\{\sigma_y\}_y$.

Lemma 3.4.7. *For all unitaries $\{U_y\}_y$, we have:*

$$\mathcal{I}(X, Y : I_1, \ldots, I_N) \leq 4N \sup_{\phi, \|\phi\|_2 \leq 1} \frac{1}{M} \sum_{|x|, y \leq m/2} \langle \phi | U_y O_{x,y} U_y^\dagger | \phi \rangle^2 \varepsilon^2, \tag{3.114}$$

where for (x, y), $O_{x,y} = U_y^\dagger (d\rho_{x,y} - \mathbb{I}) U_y$.

Proof. We suppose that the eigenvalues of σ_y have the form

$$\lambda_i^y = \frac{1 + 10\beta_i^y \varepsilon}{d}, \tag{3.115}$$

where $\beta_i^y = \pm 1$ satisfying $\sum_i \beta_i^y = 0$ (exactly half are equal to $+1$) and $\beta^y = -\beta^{m+1-y}$ (we suppose m even). The diagonalizing matrices $\{U_y\}_y$ are chosen randomly so as they satisfy $U_{m+1-y} = U_y$ for all $y \leq m/2$ and other conditions to be specified later.

Let us denote by $\mathcal{M}^t$ the POVM used at step t. Without loss of generality, we can suppose that the non-adaptive algorithm performs only measurements of the following form:

$$\mathcal{M}^t = \{|\phi_i^t\rangle\langle\phi_i^t|\}_i. \tag{3.116}$$

where we have the condition $\sum_i |\phi_i^t\rangle\langle\phi_i^t| = \mathbb{I}$ implying for all i and t: $\|\phi_i^t\|_2 \leq 1$.

Let $M = 2me^{cd}$, we can write the mutual information as follows:

$$\mathcal{I}(X, Y : I_1, \ldots, I_N) = \frac{1}{M} \sum_{x,y} \sum_{i_1,\ldots,i_N} \prod_{t=1}^N \langle \phi_{i_t}^t | \rho_{x,y} | \phi_{i_t}^t \rangle \log \left(\frac{\prod_{t=1}^N \langle \phi_{i_t}^t | \rho_{x,y} | \phi_{i_t}^t \rangle}{\frac{M}{\sum_{x,y}} \prod_{t=1}^N \langle \phi_{i_t}^t | \rho_{x,y} | \phi_{i_t}^t \rangle} \right) \tag{3.117}$$

$$= \Sigma_1 + \Sigma_2, \tag{3.118}$$

where Σ_1 and Σ_2 are defined as follows:

$$\Sigma_1 = \frac{1}{M} \sum_{x,y} \sum_{i_1,\ldots,i_N} \prod_{t=1}^N \langle \phi_{i_t}^t | \rho_{x,y} | \phi_{i_t}^t \rangle \log \left(\prod_{t=1}^N \langle \phi_{i_t}^t | d\rho_{x,y} | \phi_{i_t}^t \rangle \right), \tag{3.119}$$

$$\Sigma_2 = -\frac{1}{M} \sum_{x,y} \sum_{i_1,\ldots,i_N} \prod_{t=1}^N \langle \phi_{i_t}^t | \rho_{x,y} | \phi_{i_t}^t \rangle \log \left(\frac{1}{M} \sum_{x,y} \prod_{t=1}^N \langle \phi_{i_t}^t | d\rho_{x,y} | \phi_{i_t}^t \rangle \right). \tag{3.120}$$

Since

$$\rho_{x,y} = U_y \operatorname{diag} \left(\left\{ \frac{1 + (10\beta_i^y + 2\alpha_i^x/3)\varepsilon}{d} \right\}_{i \in [d]} \right) U_y^\dagger$$

we can write

$$\langle \phi_{i_t}^t | \rho_{x,y} | \phi_{i_t}^t \rangle = \frac{1 + u_{i_t}^{t,x,y} \varepsilon}{d}, \tag{3.121}$$

where $u_{i_t}^{t,x,y} = \langle \phi_{i_t}^t | U_y \operatorname{diag} \left(\{10\beta_i^y + 2\alpha_i^x/3\}_{i \in [d]} \right) U_y | \phi_{i_t}^t \rangle \in (-11, 11)$. Denote by $O_{x,y} = \operatorname{diag} \left(\{10\beta_i^y + 2\alpha_i^x/3\}_{i \in [d]} \right)$, we remark that

$$\sum_{i_t=1}^d u_{i_t}^{t,x,y} = \sum_{i_t=1}^d \langle \phi_{i_t}^t | U_y \operatorname{diag} \left(\{10\beta_i^y + 2\alpha_i^x/3\}_{i \in [d]} \right) U_y | \phi_{i_t}^t \rangle \tag{3.122}$$

$$= \operatorname{Tr} U_y \operatorname{diag} \left(\{10\beta_i^y + 2\alpha_i^x/3\}_{i \in [d]} \right) U_y = \operatorname{Tr} \operatorname{diag} \left(\{10\beta_i^y + 2\alpha_i^x/3\}_{i \in [d]} \right) \tag{3.123}$$

$$= \sum_{i=1}^d 10\beta_i^y + 2\alpha_i^x/3 = 0. \tag{3.124}$$

Moreover, the couples of quantum states $(\rho_{x,y}, \rho_{-x,y})$ and $(\rho_{x,y}, \rho_{x,m+1-y})$ are symmetric with respect to $\mathbb{I}/d$ by the construction of $(\alpha_i^x)_{i,x}$ and $(\beta_i^x)_{i,x}$ hence

$$u_{i_t}^{t,-x,m+1-y} = \left\langle \phi_{i_t}^t \right| U_{m+1-y} \operatorname{diag}\left(\left\{ 10\beta_i^{m+1-y} + 2\alpha_i^{-x}/3 \right\}_{i\in[d]} \right) U_{m+1-y} \left| \phi_{i_t}^t \right\rangle \tag{3.125}$$

$$= \left\langle \phi_{i_t}^t \right| U_y \operatorname{diag}\left(\left\{ -10\beta_i^{y} - 2\alpha_i^{x}/3 \right\}_{i\in[d]} \right) U_y \left| \phi_{i_t}^t \right\rangle \tag{3.126}$$

$$= -\left\langle \phi_{i_t}^t \right| U_y \operatorname{diag}\left(\left\{ 10\beta_i^{y} + 2\alpha_i^{x}/3 \right\}_{i\in[d]} \right) U_y \left| \phi_{i_t}^t \right\rangle \tag{3.127}$$

$$= -u_{i_t}^{t,x,y}. \tag{3.128}$$

Suppose that $\varepsilon \leq 0.05$. We can start by controlling Σ_2 using Jensen's inequality:

$$\Sigma_2 = -\frac{1}{M} \sum_{x,y} \sum_{i_1,\ldots,i_N} \prod_{t=1}^N \left(\frac{1 + u_{i_t}^{t,x,y}\varepsilon}{d} \right) \log\left(\frac{1}{M} \sum_{x,y} \prod_{t=1}^N (1 + u_{i_t}^{t,x,y}\varepsilon) \right) \tag{3.129}$$

$$\leq -\frac{1}{M} \sum_{x,y,i} \prod_{t=1}^N \left(\frac{1 + u_{i_t}^{t,x,y}\varepsilon}{d} \right) \left(\frac{1}{M} \sum_{x,y} \log\left(\prod_{t=1}^N (1 + u_{i_t}^{t,x,y}\varepsilon) \right) \right) \tag{3.130}$$

$$= -\frac{1}{M} \sum_{x,y,i} \prod_{t=1}^N \left(\frac{1 + u_{i_t}^{t,x,y}\varepsilon}{d} \right) \left(\frac{1}{M} \sum_{x,y,t} \log\left(1 + u_{i_t}^{t,x,y}\varepsilon \right) \right) \tag{3.131}$$

$$= -\frac{1}{M} \sum_{x,y,i} \prod_{t=1}^N \left(\frac{1 + u_{i_t}^{t,x,y}\varepsilon}{d} \right) \left(\frac{1}{M} \sum_{|x|,y\leq m/2,t} \log\left(1 + u_{i_t}^{t,x,y}\varepsilon \right) + \log\left(1 - u_{i_t}^{t,x,y}\varepsilon \right) \right) \tag{3.132}$$

$$= -\frac{1}{M} \sum_{x,y,i} \prod_{t=1}^N \left(\frac{1 + u_{i_t}^{t,x,y}\varepsilon}{d} \right) \left(\frac{1}{M} \sum_{|x|,y\leq m/2,t} \log\left(1 - (u_{i_t}^{t,x,y})^2\varepsilon^2 \right) \right). \tag{3.133}$$

Now, we can use the inequality $-\log(1 - x^2) \leq 2x^2$ for $|x| \leq 1/\sqrt{2}$:

$$\Sigma_2 \leq -\frac{1}{M} \sum_{x,y,i} \prod_{t=1}^N \left(\frac{1 + u_{i_t}^{t,x,y}\varepsilon}{d} \right) \left(\frac{1}{M} \sum_{|x|,y\leq m/2,t} \log\left(1 - (u_{i_t}^{t,x,y})^2\varepsilon^2 \right) \right) \tag{3.134}$$

$$\leq \frac{1}{M} \sum_{x,y,i} \prod_{t=1}^N \left(\frac{1 + u_{i_t}^{t,x,y}\varepsilon}{d} \right) \left(\frac{1}{M} \sum_{|x|,y\leq m/2,t} 2(u_{i_t}^{t,x,y})^2\varepsilon^2 \right) \tag{3.135}$$

$$\leq \frac{1}{M} \sum_{x,y,i} \prod_{t=1}^N \left(\frac{1 + u_{i_t}^{t,x,y}\varepsilon}{d} \right) \left(\sum_t \sup_{\phi,\|\phi\|_2\leq 1} \frac{1}{M} \sum_{|x|,y\leq m/2} 2 \left\langle \phi \right| U_y O_{x,y} U_y^\dagger \left| \phi \right\rangle^2 \varepsilon^2 \right) \tag{3.136}$$

$$\leq N \sup_{\phi,\|\phi\|_2\leq 1} \frac{1}{M} \sum_{|x|,y\leq m/2} 2 \left\langle \phi \right| U_y O_{x,y} U_y^\dagger \left| \phi \right\rangle^2 \varepsilon^2. \tag{3.137}$$

Using the fact that $\sum_{i_t} u_{i_t}^{t,x,y} = 0$ for all t, x, y along with the inequality $(1 + x)\log(1 + x) + (1 - x)\log(1 - x) \leq 2x^2$ for $|x| \leq 1/\sqrt{2}$ we can upper bound the first sum Σ_1:

$$\Sigma_1 = \frac{1}{M}\sum_{x,y,i}\prod_{t=1}^{N}\left(\frac{1+u_{i_t}^{t,x,y}\varepsilon}{d}\right)\log\left(\prod_{t=1}^{N}\left(1+u_{i_t}^{t,x,y}\varepsilon\right)\right) \tag{3.138}$$

$$\leq \frac{1}{M}\sum_{x,y,i}\prod_{t=1}^{N}\left(\frac{1+u_{i_t}^{t,x,y}\varepsilon}{d}\right)\sum_{k}\log\left(1+u_{i_k}^{k,x,y}\varepsilon\right) \tag{3.139}$$

$$\leq \frac{1}{M}\sum_{x,y,k}\sum_{i_k}\sum_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_N}\prod_{t=1}^{N}\left(\frac{1+u_{i_t}^{t,x,y}\varepsilon}{d}\right)\log\left(1+u_{i_k}^{k,x,y}\varepsilon\right) \tag{3.140}$$

$$\leq \frac{1}{M}\sum_{x,y,k}\sum_{i_k}\left(\frac{1+u_{i_k}^{k,x,y}\varepsilon}{d}\right)\log\left(1+u_{i_k}^{k,x,y}\varepsilon\right) \tag{3.141}$$

$$\leq \frac{1}{Md}\sum_{|x|,y\leq m/2,k}\sum_{i_k}\left(1+u_{i_k}^{k,x,y}\varepsilon\right)\log\left(1+u_{i_k}^{k,x,y}\varepsilon\right)+\left(1-u_{i_k}^{k,x,y}\varepsilon\right)\log\left(1-u_{i_k}^{k,x,y}\varepsilon\right) \tag{3.142}$$

$$\leq \frac{1}{Md}\sum_{|x|,y\leq m/2,k,i_k} 2(u_{i_k}^{k,x,y}\varepsilon)^2 \tag{3.143}$$

$$\leq \frac{1}{d}\sum_{k,i_k}\sup_{\phi,\|\phi\|_2\leq 1}\frac{1}{M}\sum_{|x|,y\leq m/2} 2\left\langle\phi\right| U_y O_{x,y} U_y^\dagger \left|\phi\right\rangle^2 \varepsilon^2 \tag{3.144}$$

$$\leq 2N\sup_{\phi,\|\phi\|_2\leq 1}\frac{1}{M}\sum_{|x|,y\leq m/2}\left\langle\phi\right| U_y O_{x,y} U_y^\dagger \left|\phi\right\rangle^2 \varepsilon^2. \tag{3.145}$$

Finally the upper bounds on Σ_1 and Σ_2 imply the required upper bound on their sum $\Sigma_1 + \Sigma_2 = \mathcal{I}(X, Y : I_1, \ldots, I_N)$. □

Note that we need to take a supremum over all possible vectors ϕ because the learner knows the quantum states $\{\sigma_y\}_y$ and so it can choose measurements dependent on the unitaries $\{U_y\}_y$. We can now show that with high probability on the choice of the unitaries $\{U_y\}_y$, the latter supremum can bounded and so the mutual information too.

Lemma 3.4.8. *Let $\{U_y\}_y$ be m unitary matrices distributed according to the* Haar(d) *distribution. We have with a probability at least* $9/10$*:*

$$4N\sup_{\phi,\|\phi\|_2\leq 1}\frac{1}{M}\sum_{|x|,y\leq m/2}\left\langle\phi\right| U_y O_{x,y} U_y^\dagger \left|\phi\right\rangle^2 \varepsilon^2 = \mathcal{O}\left(\frac{N\varepsilon^2\log(m)}{m}+\frac{N\varepsilon^2}{d}\right). \tag{3.146}$$

Proof. To upper bound the previous supremum, we use a similar approach to [CCHL22]: For $U \sim$ Haar(d), ϕ such that $\|\phi\|_2 \leq 1$ and a trace-less Hermitian matrix O, let $f(\phi, U) = \left\langle\phi\right| UOU^\dagger \left|\phi\right\rangle$, we have $\mathbb{E}\left(f(\phi,U)\right) = \frac{1}{d}\mathrm{Tr}(O)\mathrm{Tr}(|\phi\rangle\langle\phi|) = 0$ (see Weingarten calculus [Gu13]) and f is $2\|O\|$-Lipschitz:

$$|f(U) - f(V)| \leq 2\left|\left\langle\phi\right| (U-V)OU^\dagger \left|\phi\right\rangle\right| \leq 2\|O\|\|U-V\|_2. \tag{3.147}$$

Therefore by the concentration inequality of Lipschitz functions of Haar unitary matrices [MM13]:

$$\mathbb{P}\left(|f(U)| > t\right) \leq \exp(-dt^2/48). \tag{3.148}$$

Hence

$$\mathbb{P}\left(|f(U)|^2 > t\right) \leq \exp(-dt/48). \tag{3.149}$$

For $m/2$ unitaries $U_1, \ldots, U_{m/2}$ and $\lambda = 2d/C$ for sufficiently large C. Denote by $X = |f(U)|^2$, by Markov's inequality:

$$\mathbb{P}\left(\frac{2}{m}\sum_{1\leq y\leq m/2}|f(U_y)|^2 > t\right) \leq \exp(-\lambda mt/2)\mathbb{E}\left(e^{\lambda X}\right)^{m/2} \tag{3.150}$$

$$\leq \exp(-\lambda mt/2)\left(1+\int_0^\infty dx\lambda e^{\lambda x}e^{-dx/48}\right)^{m/2} \tag{3.151}$$

$$\leq \exp(-dmt/2C)\left(C'\right)^{m/2} \leq \exp(-dmt/C + m\log(C')), \tag{3.152}$$

with C' another constant. In order to prove an inequality valid for all ϕ in the unit ball. Let's take an η-net $\{\phi_i\}_i$ of the unit ball of size at most $(1+2/\eta)^{2d}$. For ϕ such that $\|\phi\|_2 \leq 1$, there is ϕ_i in the net such that $\|\phi - \phi_i\|_2 \leq \eta$. Moreover $|f(\phi, U)| \leq \|O\|$ so

$$\left|\frac{2}{m}\sum_{1\leq y\leq m/2} f(\phi, U_y)^2 - f(\phi_i, U_y)^2\right| \leq \frac{2}{m}\sum_{1\leq y\leq m/2}|f(\phi, U_y)^2 - f(\phi_i, U_y)^2| \tag{3.153}$$

$$\leq \frac{2}{m}\sum_{1\leq y\leq m/2} 2\|O\|\|(\langle\phi| - \langle\phi_i|)U_y O U_y^\dagger|\phi\rangle\,| \leq 2\eta\|O\|^2. \tag{3.154}$$

Therefore

$$\mathbb{P}\left(\exists\phi : \frac{2}{m}\sum_{1\leq y\leq m/2}|f(\phi, U_y)|^2 > t + 2\eta\|O\|^2\right) \leq \mathbb{P}\left(\exists\phi_i : \frac{1}{m}\sum_{k=1}^m |f(\phi_i, U_k)|^2 > t\right) \tag{3.155}$$

$$\leq (1+2/\eta)^{2d}\exp(-dmt/C + m\log(C')). \tag{3.156}$$

Taking $\eta = 1/m$ yields:

$$\mathbb{P}\left(\exists\phi : \frac{2}{m}\sum_{1\leq y\leq m/2}|f(\phi, U_y)|^2 > t + 2\|O\|^2/m\right) \leq (1+2m)^{2d}\exp(-dmt/C + m\log(C')). \tag{3.157}$$

Applying the union bound, we can obtain:

$$\mathbb{P}\left(\exists\phi, \exists x, \frac{2}{m}\sum_{y\leq m/2}\langle\phi|\, U_y O_{x,y} U_y^\dagger\,|\phi\rangle^2 \geq t + \frac{2\|O_{x,y}\|^2}{m}\right) \tag{3.158}$$

$$\leq 4e^{cd}(1+2m)^{2d}\exp(-dmt/C + m\log(C')). \tag{3.159}$$

Let's take $t = C\frac{\log(40)+cd+2d\log(1+2m)+m\log(C')}{dm}$ in order to have

$$\mathbb{P}\left(\forall\phi, \frac{1}{M}\sum_{|x|,y\leq m/2}\langle\phi|\, U_y O_{x,y} U_y^\dagger\,|\phi\rangle^2 \leq t + \frac{2\|O_{x,y}\|^2}{m}\right) \tag{3.160}$$

$$\geq \mathbb{P}\left(\forall\phi,\forall x, \frac{1}{m}\sum_{y\leq m/2}\langle\phi|\, U_y O_{x,y} U_y^\dagger\,|\phi\rangle^2 \leq t + \frac{2\|O_{x,y}\|^2}{m}\right) \geq 9/10. \tag{3.161}$$

Therefore we have the existence of $\{U_y\}_y$ such that for all $y \neq z$, $\|\sigma_y - \sigma_z\|_{\mathrm{Tr}} > \varepsilon$, and

$$\sup_{\phi,\|\phi\|_2\leq 1}\frac{1}{M}\sum_{|x|,y\leq m/2}\langle\phi|\, U_y O_{x,y} U_y^\dagger\,|\phi\rangle^2 \leq \frac{\mathrm{Tr}(O_{x,y}^2)}{d(d+1)} + t + \frac{2\|O_{x,y}\|^2}{m} \tag{3.162}$$

$$\leq \frac{201}{d+1} + C\frac{\log(40)+cd+2d\log(1+2m)+m\log(C')}{dm} + \frac{242}{m}. \tag{3.163}$$

Finally, we have shown the existence of quantum states $\{\sigma_{x,y}\}_{x,y}$ such that:

$$\mathcal{I}(X,Y:I_1,\ldots,I_N) = \mathcal{O}\left(\frac{1}{d}+\frac{\log(m)}{m}\right)N\varepsilon^2. \tag{3.164}$$

□

To sum up, we have shown the existence of quantum states $\{\sigma_{x,y}\}_{x,y}$ such that:

$$\Omega(\log(m)+d) \leq \mathcal{I}(X,Y:X',Y') \leq \mathcal{I}(X,Y:I_1,\ldots,I_N) \leq \mathcal{O}\left(\frac{1}{d}+\frac{\log(m)}{m}\right)N\varepsilon^2. \tag{3.165}$$

We conclude that $N = \Omega\left(\min\left\{\frac{md}{\log(m)\varepsilon^2}, \frac{d^2}{\varepsilon^2}\right\}\right)$.

□

A similar proof strategy allows to derive a lower bound on the copy complexity of adaptive strategies. The result is stated in the next proposition.

Proposition 3.4.1. *There is a tuple of quantum states $(\sigma_1,\ldots,\sigma_m)$ such that any learning algorithm with possibly adaptive independent measurements requires*

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

copies of ρ to approximate ρ to at most $\varepsilon/10$ with at least a probability $2/3$.

Proof. Recall that the mutual information between (X,Y) and $(I_1,\ldots,I_N)$ can be expressed as:

$$\mathcal{I}(X,Y:I_1,\ldots,I_N) = \Sigma_1 + \Sigma_2. \tag{3.166}$$

The second sum can be upper bounded by the same technique as before (using for example Jensen's inequality and the inequality $-\log(1-x^2) \leq 2x^2$) and yields the same upper bound. The first sum is more involved because the product cannot be simplified due to

the dependence between the POVMs and the previous outcomes. To see this, we can try to simplify the first sum as far as possible:

$$\Sigma_1 = \frac{1}{M}\sum_{x,y}\sum_{u_1,\ldots,u_N}\prod_{t=1}^{N}\left\langle\phi_{u_t}^{u_{<t}}\right|\rho_{x,y}\left|\phi_{u_t}^{u_{<t}}\right\rangle\log\left(\prod_{t=1}^{N}\left\langle\phi_{u_t}^{u_{<t}}\right|d\rho_{x,y}\left|\phi_{u_t}^{u_{<t}}\right\rangle\right) \tag{3.167}$$

$$= \frac{1}{M}\sum_{x,y}\sum_{u_1,\ldots,u_N}\prod_{t=1}^{N}\left\langle\phi_{u_t}^{u_{<t}}\right|\rho_{x,y}\left|\phi_{u_t}^{u_{<t}}\right\rangle\sum_{t=1}^{N}\log\left(\left\langle\phi_{u_t}^{u_{<t}}\right|d\rho_{x,y}\left|\phi_{u_t}^{u_{<t}}\right\rangle\right) \tag{3.168}$$

$$= \frac{1}{M}\sum_{x,y,k}\sum_{u_1,\ldots,u_N}\prod_{t=1}^{N}\left\langle\phi_{u_t}^{u_{<t}}\right|\rho_{x,y}\left|\phi_{u_t}^{u_{<t}}\right\rangle\log\left(\left\langle\phi_{u_k}^{u_{<k}}\right|d\rho_{x,y}\left|\phi_{u_k}^{u_{<k}}\right\rangle\right) \tag{3.169}$$

$$= \frac{1}{M}\sum_{x,y,k}\sum_{u_1,\ldots,u_k}\prod_{t=1}^{k}\left\langle\phi_{u_t}^{u_{<t}}\right|\rho_{x,y}\left|\phi_{u_t}^{u_{<t}}\right\rangle\log\left(\left\langle\phi_{u_k}^{u_{<k}}\right|d\rho_{x,y}\left|\phi_{u_k}^{u_{<k}}\right\rangle\right), \tag{3.170}$$

where the last equality follows from the fact that

$$\sum_{u_t}\left\langle\phi_{u_t}^{u_{<t}}\right|\rho_{x,y}\left|\phi_{u_t}^{u_{<t}}\right\rangle = \mathrm{Tr}(\rho_{x,y}) = 1, \tag{3.171}$$

for $t > k$ and $\log\left(\left\langle\phi_{u_k}^{u_{<k}}\right|d\rho_{x,y}\left|\phi_{u_k}^{u_{<k}}\right\rangle\right)$ is independent from u_t. But we are stuck at k, we cannot simplify the sums on u_s for $s < k$ since $\left\langle\phi_{u_s}^{u_{<s}}\right|\rho_{x,y}\left|\phi_{u_s}^{u_{<s}}\right\rangle$ has common terms with $\left\langle\phi_{u_k}^{u_{<k}}\right|\rho_{x,y}\left|\phi_{u_k}^{u_{<k}}\right\rangle$ which is inside the log function.

In order to circumvent this difficulty, we can upper bound the k^{th} term which poses the obstacle of simplification. Using the inequality $\log(x) \leq x - 1$ for all $x > -1$ we

obtain:

$$\Sigma_1 = \frac{1}{M}\sum_{x,y,k}\sum_{u_1,\ldots,u_k}\prod_{t=1}^{k}\left\langle\phi_{u_t}^{u_{<t}}\right|\rho_{x,y}\left|\phi_{u_t}^{u_{<t}}\right\rangle\log\left(\left\langle\phi_{u_k}^{u_{<k}}\right|d\rho_{x,y}\left|\phi_{u_k}^{u_{<k}}\right\rangle\right) \tag{3.172}$$

$$= \frac{1}{M}\sum_{x,y,k}\sum_{u_1,\ldots,u_k}\prod_{t=1}^{k}\left\langle\phi_{u_t}^{u_{<t}}\right|\rho_{x,y}\left|\phi_{u_t}^{u_{<t}}\right\rangle\left(\left\langle\phi_{u_k}^{u_{<k}}\right|d\rho_{x,y}\left|\phi_{u_k}^{u_{<k}}\right\rangle - 1\right) \tag{3.173}$$

$$= \frac{1}{M}\sum_{x,y,k}\sum_{u_1,\ldots,u_k}\prod_{t=1}^{k}\left\langle\phi_{u_t}^{u_{<t}}\right|\left(\frac{\mathbb{I}}{d}+\varepsilon\frac{O_{x,y}}{d}\right)\left|\phi_{u_t}^{u_{<t}}\right\rangle\left\langle\phi_{u_k}^{u_{<k}}\right|\varepsilon O_{x,y}\left|\phi_{u_k}^{u_{<k}}\right\rangle \tag{3.174}$$

$$= \frac{1}{M}\sum_{x,y,k}\sum_{u_1,\ldots,u_{k-1}}\prod_{t=1}^{k-1}\left\langle\phi_{u_t}^{u_{<t}}\right|\left(\frac{\mathbb{I}}{d}+\varepsilon\frac{O_{x,y}}{d}\right)\left|\phi_{u_t}^{u_{<t}}\right\rangle\sum_{u_k}\frac{1}{d}\left\langle\phi_{u_k}^{u_{<k}}\right|\varepsilon O_{x,y}\left|\phi_{u_k}^{u_{<k}}\right\rangle \tag{3.175}$$

$$+ \frac{1}{M}\sum_{x,y,k}\sum_{u_1,\ldots,u_{k-1}}\prod_{t=1}^{k-1}\left\langle\phi_{u_t}^{u_{<t}}\right|\left(\frac{\mathbb{I}}{d}+\varepsilon\frac{O_{x,y}}{d}\right)\left|\phi_{u_t}^{u_{<t}}\right\rangle\sum_{u_k}\frac{1}{d}\left\langle\phi_{u_k}^{u_{<k}}\right|\varepsilon O_{x,y}\left|\phi_{u_k}^{u_{<k}}\right\rangle^2 \tag{3.176}$$

$$\leq \frac{1}{M}\sum_{x,y,k}\sum_{u_1,\ldots,u_{k-1}}\prod_{t=1}^{k-1}\left\langle\phi_{u_t}^{u_{<t}}\right|\left(\frac{\mathbb{I}}{d}+\varepsilon\frac{O_{x,y}}{d}\right)\left|\phi_{u_t}^{u_{<t}}\right\rangle\times\frac{1}{d}\times\mathrm{Tr}(\varepsilon O_{x,y}) \tag{3.177}$$

$$+ \frac{1}{M}\sum_{x,y,k}\sum_{u_1,\ldots,u_{k-1}}\prod_{t=1}^{k-1}\left\langle\phi_{u_t}^{u_{<t}}\right|\left(\frac{\mathbb{I}}{d}+\varepsilon\frac{O_{x,y}}{d}\right)\left|\phi_{u_t}^{u_{<t}}\right\rangle\times\frac{1}{d}\times\mathrm{Tr}(\varepsilon^2 O_{x,y}^2) \tag{3.178}$$

$$\leq \frac{1}{M}\sum_{x,y,k}\sum_{u_1,\ldots,u_{k-1}}\prod_{t=1}^{k-1}\left\langle\phi_{u_t}^{u_{<t}}\right|\left(\frac{\mathbb{I}}{d}+\varepsilon\frac{O_{x,y}}{d}\right)\left|\phi_{u_t}^{u_{<t}}\right\rangle\times\frac{1}{d}\times 11^2 d\varepsilon^2 \tag{3.179}$$

$$\leq \frac{1}{M}\sum_{x,y,k}1\times 11^2\varepsilon^2 \leq 11^2 N\varepsilon^2, \tag{3.180}$$

where we use again $\sum_{u_t}\left\langle\phi_{u_t}^{u_{<t}}\right|O_{x,y}\left|\phi_{u_t}^{u_{<t}}\right\rangle = \mathrm{Tr}(O_{x,y}) = 0$ for all t and

$$\sum_{u_k}\left\langle\phi_{u_k}^{u_{<k}}\right|O_{x,y}\left|\phi_{u_k}^{u_{<k}}\right\rangle^2 = \sum_{u_k}\mathrm{Tr}(O_{x,y}\left|\phi_{u_k}^{u_{<k}}\right\rangle\left\langle\phi_{u_k}^{u_{<k}}\right|O_{x,y}\left|\phi_{u_k}^{u_{<k}}\right\rangle\left\langle\phi_{u_k}^{u_{<k}}\right|) \tag{3.181}$$

$$\leq \sum_{u_k}\mathrm{Tr}(O_{x,y}^2\left|\phi_{u_k}^{u_{<k}}\right\rangle\left\langle\phi_{u_k}^{u_{<k}}\right|) = \mathrm{Tr}(O_{x,y}^2) \leq 11^2 d. \tag{3.182}$$

Therefore the mutual information can be upper bounded by

$$\mathcal{I}(X,Y:X',Y') \leq 121N\varepsilon^2 + 2N\sup_{\phi,\|\phi\|_2\leq 1}\frac{1}{M}\sum_{|x|,y\leq m/2}\langle\phi|U_y O_{x,y}U_y^\dagger|\phi\rangle^2\varepsilon^2 \tag{3.183}$$

$$\leq 123N\varepsilon^2. \tag{3.184}$$

Since the mutual information is always lower bounded by $\Omega(\log(m)+d)$ we conclude that $N = \Omega((d+\log(m)/\varepsilon^2)$. Finally, we have proven the following lower bound on adaptive strategies for hypothesis selection problem: □

This proposition along with the analysis of Algorithm 6 show that the near optimal copy complexity of the problem (P) using adaptive independent measurements is $\tilde{\Theta}\left(\frac{d}{\varepsilon^2}\right)$. This latter along with Theorem 3.4.2 imply the separation between adaptive and non-adaptive strategies for the problem (P) for $m \gg 1$. In other words, knowing that the

eigenbasis of the quantum state belongs to some family of bases gives an advantage to adaptive strategies since they can find the eigenbasis, and then focus on measuring the quantum state with the corresponding POVM. Up to our knowledge, this is the first example for which adaptive independent strategies outperform non-adaptive ones for quantum testing problems.

3.5 Conclusion

We have constructed hypothesis selection problems for which sequential strategies are more efficient than non-adaptive ones. The problem for which the advantage is the most significant is the one presented in Section 3.4. However, this problem is quite contrived and it would be interesting to see if there is another more natural problem for which such a separation exists. We conjecture the separation would be polynomial in m for the composite hypothesis selection problem: distinguishing between $\rho \in \{\sigma_1, \ldots, \sigma_m\}$ and $\rho \in \{\sigma_{m+1}, \ldots, \sigma_{2m}\}$ with high probability.

Chapter 4

Quantum Channel Certification with Incoherent Measurements

4.1 Introduction

We consider the problem of quantum channel certification which consists in verifying whether a quantum process to which we have black box access behaves as intended. A valid process in quantum theory is modelled by a quantum channel. Given a complete description of a known quantum channel $\mathcal{N}_0$ and N copies of an unknown quantum channel $\mathcal{N}$ that can be either $\mathcal{N}_0$ or ε-far from it, at each step $1 \leq t \leq N$, we can choose an input quantum state, send it through the unknown process $\mathcal{N}$ then measure the output quantum state. After collecting a sufficient amount of classical observations, our goal is to decide in which case is the quantum channel $\mathcal{N}$ with high probability and while minimizing N. We also call this problem testing identity to the quantum channel $\mathcal{N}_0$.

This testing task is important for many reasons. Firstly, the building blocks of a quantum computation are unitary gates. It is thus important to understand the complexity of checking that an unknown channel implements a given gate as specified. Secondly, quantum channel certification is the natural generalization of the quantum state certification. Indeed, if the channels are constant quantum states, then testing them becomes equivalent to testing those states. Besides, using the Choi–Jamiołkowski isomorphism, quantum channel certification can be obtained by applying quantum state certification protocols on the Choi states. However, since we do not allow the use of an auxiliary system and we use a notion of distance adapted for channels that does not correspond to the trace distance between Choi states, channel certification is strictly more general than state certification. Finally, quantum process tomography, the problem of learning completely a channel in the diamond norm is costly (Chapter 5) and our hope is that certification can be done with fewer copies than full tomography.

In this chapter, we focus on two extreme cases, $\mathcal{N}_0(\rho) = \mathcal{N}_U(\rho) = U\rho U^\dagger$ is a unitary channel where U is a unitary matrix and $\mathcal{N}_0(\rho) = \mathcal{D}(\rho) = \mathrm{Tr}(\rho)\frac{\mathbb{I}}{d_{\mathrm{out}}}$ is the completely depolarizing channel.

Contribution. We propose an ancilla-free testing algorithm for testing identity to a fixed unitary channel $\mathcal{N}_U$ in the trace distance using $\mathcal{O}(d/\varepsilon^2)$ independent measurements (here $d_{\mathrm{in}} = d_{\mathrm{out}} = d$). The tester chooses a random input state and measures with the corresponding POVM conjugated by the unitary U. This result is stated in Theorem 4.3.1. The standard inequality relating the 1-norm of a Choi state and the diamond norm of the channel only implies an upper bound $\mathcal{O}(d^2/\varepsilon^2)$. We obtain the quadratic improvement in

the dimension dependency by proving a new inequality between the entanglement fidelity and the trace distance to the identity channel (Lemma 4.3.3). Moreover, we establish a matching lower bound of $\Omega(d/\varepsilon^2)$ for testing identity to a fixed unitary channel in the trace distance. For this, we construct a well-chosen distribution of channels ε-far from the identity channel. After a sufficient number of measurements, the observations under the two hypotheses should be distinguishable, i.e., the (Kullback-Leibler) KL divergence is $\Omega(1)$. However, we can show that for this particular choice of distribution over channels, any ancilla-free adaptive tester can only increase the KL divergence by at most $\mathcal{O}(\varepsilon^2/d)$ after a measurement no matter the dependence on the previous observations. The lower bound is stated and proved in Theorem 4.3.2.
Concerning the certification of the completely depolarizing channel $\mathcal{D}(\rho) = \mathrm{Tr}(\rho)\frac{\mathbb{I}}{d_{\mathrm{out}}}$, we propose an ancilla-free strategy to distinguish between $\mathcal{N} = \mathcal{D}$ and $\mathcal{N}$ is ε-far from it in the diamond distance using $\mathcal{O}(d_{\mathrm{in}}^2 d_{\mathrm{out}}^{1.5}/\varepsilon^2)$ independent measurements (see Theorem 4.4.2). For this we show how to reduce this certification problem to the certification of the maximally mixed state (testing mixedness) $\frac{\mathbb{I}}{d_{\mathrm{out}}}$. We choose the input state $|\phi\rangle\langle\phi|$ randomly and we compare the 2-norm between the output state $\mathcal{N}(|\phi\rangle\langle\phi|)$ and the maximally mixed state $\frac{\mathbb{I}}{d_{\mathrm{out}}}$. We show that with at least a probability $\Omega(1)$, we have $Y = \|\mathcal{N}(|\phi\rangle\langle\phi|) - \mathbb{I}/d_{\mathrm{out}}\|_2^2 \geq \varepsilon^2/(4d_{\mathrm{out}}d_{\mathrm{in}}^2)$. This inequality is sufficient to obtain the required complexity, however, it requires some work to be proved. First, we show a similar inequality in expectation using Weingarten calculus. Then we control the variance of the random variable Y carefully in a way that this upper bound depends on the actual difficulty of the problem, mainly the expectation of Y and the diamond distance between $\mathcal{N}$ and $\mathcal{D}$. Next, we obtain the anti-concentration inequality using the Paley-Zygmund inequality.

On the other hand, we establish a lower bound of $\Omega\left(d_{\mathrm{in}}^2 d_{\mathrm{out}}^{1.5}/(\log(d_{\mathrm{in}}d_{\mathrm{out}}/\varepsilon)^2\varepsilon^2)\right)$ for testing identity to the depolarizing channel with ancilla-free non-adaptive strategies (Theorem 4.4.3). For this, we construct a random quantum channel whose output states are almost $\mathcal{O}(\varepsilon/d_{\mathrm{in}})$-close (in the 1-norm) to the maximally mixed state $\frac{\mathbb{I}}{d_{\mathrm{out}}}$ except for a neighborhood of an input state chosen randomly whose output is ε-far from $\frac{\mathbb{I}}{d_{\mathrm{out}}}$ in the 1-norm. Then we use LeCam's method [LeC73] as in [BCL20] with some differences. First, we need to condition on the event that the input states chosen by the testing algorithm have very small overlaps with the best input state. This conditioning is the main reason for the additional logarithmic factor we obtain in the lower bound. Next, with a construction using random matrices with Gaussian entries rather than Haar distributed unitaries, we can invoke hypercontractivity [AS17, Proposition 5.48] which allows us to control all the moments once we upper bound the second moment. In the special case $d_{\mathrm{in}} = 1$, this recovers a result of [BCL20] while significantly simplifying their analysis. Furthermore, a lower bound of $\Omega(d_{\mathrm{in}}^2 d_{\mathrm{out}}/\varepsilon^2)$ is proved for ancilla-free adaptive strategies (Theorem 4.4.4) using the same construction. In this proof, we use Kullback-Leibler divergence instead of the Total-Variation distance. We refer to Table 4.1 for a summary of these results.

Related work. Testing identity to a unitary channel can be seen as a generalization to the usual testing identity problem for discrete distributions [VV16] and quantum states [CLO22]. However, in the worst-case setting, testing identity to the identity channel requires $\Omega(d/\varepsilon^2)$ measurements in the contrast to testing identity to a rank 1 quantum state or a Dirac distribution which can be done with only $\mathcal{O}(1/\varepsilon^2)$ measurements/samples. Also, in the definition of testing identity to a unitary channel problem, we do not require the *unknown* tested channel to be unitary. In this latter setting, efficient tests can be designed easily if an auxiliary system is allowed. This can be found along with other tests

Testing id. to $\mathcal{N}_0$	Lower bound	Upper bound
$\mathcal{N}_0 = \mathcal{N}_U$ adaptive, d_{Tr}	$\Omega\left(\frac{d}{\varepsilon^2}\right)$, Theorem 4.3.1	$\mathcal{O}\left(\frac{d}{\varepsilon^2}\right)$, Theorem 4.3.1
$\mathcal{N}_0 = \mathcal{N}_U$ adaptive, $d_\diamond$	$\Omega\left(\frac{d}{\varepsilon^2}\right)$, Theorem 4.3.1	$\mathcal{O}\left(\frac{d}{\varepsilon^4}\right)$, Theorem 4.3.1
$\mathcal{N}_0 = \mathcal{D}$ non-adaptive	$\Omega\left(\frac{d_{\mathrm{in}}^2 d_{\mathrm{out}}^{1.5}}{\log(d_{\mathrm{in}} d_{\mathrm{out}}/\varepsilon)^2 \varepsilon^2}\right)$, Theorem 4.4.3	$\mathcal{O}\left(\frac{d_{\mathrm{in}}^2 d_{\mathrm{out}}^{1.5}}{\varepsilon^2}\right)$, Theorem 4.4.2
$\mathcal{N}_0 = \mathcal{D}$ adaptive	$\Omega\left(\frac{d_{\mathrm{in}}^2 d_{\mathrm{out}} + d_{\mathrm{out}}^{1.5}}{\varepsilon^2}\right)$, Theorem 4.4.4	$\mathcal{O}\left(\frac{d_{\mathrm{in}}^2 d_{\mathrm{out}}^{1.5}}{\varepsilon^2}\right)$, Theorem 4.4.2

Table 4.1: Lower and upper bounds for testing identity of quantum channels in the diamond and trace distances using incoherent ancilla-free strategies. $\mathcal{N}_U$ is the unitary quantum channel $\mathcal{N}_U(\rho) = U\rho U^\dagger$ and $\mathcal{D}$ is the depolarizing channel $\mathcal{D}(\rho) = \mathrm{Tr}(\rho)\frac{\mathbb{I}}{d_{\mathrm{out}}}$.

on properties of unitary channels in [Wan11]. We also refer to the survey [MW13] for other examples of tests on unitary channels. Since a unitary channel has a Choi rank equal to 1 and the depolarizing channel has a Choi rank equal to $d_{\mathrm{in}}d_{\mathrm{out}}$, the results of this chapter can be seen as a first step to obtain instance-optimal quantum channel certification as for the classical case [VV16] or quantum states [CLO22]. On the other hand, testing identity to the completely depolarizing channel is a generalization of the testing uniform of distributions [DGPP17] and testing mixedness of states [BCL20; CHLL22]. In particular, if the input dimension is $d_{\mathrm{in}} = 1$, the channels are constant and the problem reduces to a testing mixedness of states of dimension d_{out}. In this case, we recover the optimal complexity of [BCL20; CHLL22]. Another noteworthy work is unitarity estimation of [CWLY22]. In this work, it is shown that ancilla-free non-adaptive strategies could estimate $\mathrm{Tr}(\mathcal{J}_{\mathcal{N}}^2)$ to within ε using $\mathcal{O}(d^{0.5}/\varepsilon^2)$ independent measurements where $\mathcal{J}_{\mathcal{N}}$ is the Choi state of the channel $\mathcal{N}$. In particular, this estimation can be used to distinguish between $\mathcal{N} = \mathcal{N}_U$ for which $\mathrm{Tr}(\mathcal{J}_{\mathcal{N}}^2) = 1$ and $\mathcal{N} = \mathcal{D}$ for which $\mathrm{Tr}(\mathcal{J}_{\mathcal{N}}^2) = 1/d^2$. A matching lower bound (in d) is given for adaptive strategies in [CWLY22] improving the previous lower bound of [CCHL22]. This complexity may seem to contradict our results which is not the case. Indeed, such a test cannot be used for testing identity to a fixed unitary channel for instance since we can have two unitary channels (same unitarity) that are ε-far in the diamond distance.

4.2 Preliminaries

We consider quantum channels of input dimension d_{in} and output dimension d_{out}. The fidelity between two quantum states is defined $\mathrm{F}(\rho, \sigma) = \left(\mathrm{Tr}\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right)^2$. It is symmetric and admits the simpler expression if one of the quantum states ρ or σ has rank 1: $\mathrm{F}(\rho, |\phi\rangle\langle\phi|) = \langle\phi|\,\rho\,|\phi\rangle$. Recall that we denote by $\mathrm{Haar}(d)$ the Haar probability measure over the compact group of unitary $d \times d$ matrices. A Haar random vector is then any column vector of a Haar distributed unitary.

We define the trace distance between two quantum channels $\mathcal{N}$ and $\mathcal{M}$ as the trace norm of their difference: $d_{\mathrm{Tr}}(\mathcal{N}, \mathcal{M}) := \max_\rho \|(\mathcal{N} - \mathcal{M})(\rho)\|_1$ where the maximization is over quantum states. In some situations, it can be helpful to allow an auxiliary system and apply the identity on it. In this case, we obtain the diamond distance which is defined formally as $d_\diamond(\mathcal{N}, \mathcal{M}) := \max_\rho \|\mathrm{id}_d \otimes (\mathcal{N} - \mathcal{M})(\rho)\|_1$ where the maximization is

over quantum states $\rho \in \mathbb{C}^{d\times d} \otimes \mathbb{C}^{d_{\text{in}}\times d_{\text{in}}}$. Note that because of the Schmidt decomposition we can always suppose that $d = d_{\text{in}}$. The trace/diamond distance can be thought of as a worst-case distance, while we can define an average case distance by the Schatten 2-norm between the corresponding Choi states.

We consider the problem of testing identity to a fixed channel. Given a fixed quantum channel $\mathcal{N}_0$ and a precision parameter $\varepsilon > 0$, the goal is to test whether an unknown quantum channel $\mathcal{N}$ is exactly $\mathcal{N}_0$ or ε-far from it with at least a probability 2/3:

$$H_0 : \mathcal{N} = \mathcal{N}_0 \qquad \text{vs.} \qquad H_1 : \text{dist}(\mathcal{N}, \mathcal{N}_0) \geq \varepsilon \tag{4.1}$$

where dist $\in \{d_\diamond, d_{\text{Tr}}\}$. H_0 is called the null hypothesis while H_1 is called the alternate hypothesis. An algorithm $\mathcal{A}$ is 1/3-correct for this problem if it outputs H_1 while H_0 is true with a probability at most 1/3 and outputs H_0 while H_1 is true with a probability at most 1/3. If $0 < \text{dist}(\mathcal{N}, \mathcal{N}_0) < \varepsilon$, the algorithm $\mathcal{A}$ can output any hypothesis. The natural figure of merit for this test is the diamond (resp. trace) distance because it characterizes the minimal error probability to distinguish between two quantum channels when auxiliary systems are allowed (resp. not allowed) [Wat18]. A testing algorithm can only extract classical information from the unknown quantum channel $\mathcal{N}$ by performing a measurement on the output state. Moreover, we only consider ancilla-free incoherent/independent strategies. That is, the tester can only use d-dimensional input states and measure with d-dimensional measurement devices. In fact, for ancilla-free strategies we do not need to assume that we have access to a perfect channel and we do not need to be able to keep entanglement between this ancilla system and the channel we are testing. We refer to Section 1.4.3 for the definition of different types of strategies.

4.3 Testing identity to a unitary channel

In this section, we focus on the problem of testing identity to a fixed unitary channel in the diamond and trace distances. Given a fixed unitary U and a precision parameter $\varepsilon > 0$, the goal is to distinguish between the hypotheses:

$$H_0 : \mathcal{N} = \mathcal{N}_U = U \cdot U^\dagger \qquad \text{vs.} \qquad H_1 : \text{dist}(\mathcal{N}, \mathcal{N}_U) \geq \varepsilon \tag{4.2}$$

with at least a probability 2/3 where dist $\in \{d_\diamond, d_{\text{Tr}}\}$. Since we consider a unitary channel, the input and output dimensions should be equal $d_{\text{in}} = d_{\text{out}} = d$.

Reduction to the case $U = \mathbb{I}$. Knowing the unitary channel $\mathcal{N}_U$ is equivalent to knowing U. We can thus reduce every testing identity to $\mathcal{N}_U$ to testing identity to $\mathcal{N}_{\mathbb{I}} = \text{id}_d$ by conjugating the measurement device by U. This is possible because the trace/diamond distance is unitary invariant: $\text{dist}(\mathcal{N}, \mathcal{N}_U) = \text{dist}(U^\dagger \mathcal{N} U, \text{id})$ and $\text{Tr}(U^\dagger \mathcal{N}(\rho) U M) = \text{Tr}(\mathcal{N}(\rho) U M U^\dagger)$ for all ρ and M. From now on, we only consider the case $U = \mathbb{I}$ and we call this particular testing problem to $\mathcal{N}_{\mathbb{I}} = \text{id}_d$ simply "*testing identity to identity*".

Given the nature of the diamond and trace distances, under the alternate hypothesis, a channel $\mathcal{N}$ could be equal to the identity channel except on a neighbourhood of some state. In addition, this state is unknown to the learner. When the algorithm is allowed to use an auxiliary system, it can prepare the Choi state $\mathcal{J}_{\mathcal{N}}$ of the channel $\mathcal{N}$ (which essentially captures everything about the channel) by taking as input the maximally entangled state $|\Psi\rangle\langle\Psi|$. Under the null hypothesis H_0, the Choi state is exactly

$\mathcal{J}_{\mathrm{id}} = \mathrm{id} \otimes \mathrm{id}(|\Psi\rangle\langle\Psi|) = |\Psi\rangle\langle\Psi|$ while under the alternate hypothesis H_1, the Choi state $\mathcal{J}_{\mathcal{N}}$ has a fidelity with $|\Psi\rangle\langle\Psi|$ satisfying:

$$\mathrm{Tr}(\mathrm{id} \otimes \mathcal{N}(|\Psi\rangle\langle\Psi|)\,|\Psi\rangle\langle\Psi|) = \mathrm{F}(\mathcal{J}_{\mathcal{N}}, |\Psi\rangle\langle\Psi|) \leq 1 - \frac{1}{4}\left\|\mathcal{J}_{\mathcal{N}} - \mathcal{J}_{\mathrm{id}}\right\|_1^2 \leq 1 - \frac{d_\diamond(\mathcal{N}, \mathrm{id})^2}{4d^2} \tag{4.3}$$

where we use a Fuchs–van de Graaf inequality [FVDG99] and the standard inequality relating the diamond norm between two channels and the trace norm between their corresponding Choi states: $\|\mathcal{J}_{\mathcal{N}} - \mathcal{J}_{\mathcal{M}}\|_1 \geq \frac{d_\diamond(\mathcal{N},\mathcal{M})}{d}$ (e.g., [JP16]). Thus, a measurement using the POVM $\mathcal{M}_\Psi = \{|\Psi\rangle\langle\Psi|, \mathbb{I} - |\Psi\rangle\langle\Psi|\}$ can distinguish between the two situations. However, if the tester is not allowed to use an auxiliary system it can neither prepare the Choi state $\mathcal{J}_{\mathcal{N}}$ nor measure using the POVM $\mathcal{M}_\Psi$. Instead, we use a random d-dimensional rank-1 input state. Indeed, this choice is natural because the expected fidelity between the input state $|\phi\rangle\langle\phi|$ and the output state $\mathcal{N}(|\phi\rangle\langle\phi|)$ can be easily related to the fidelity between Choi states:

Lemma 4.3.1. *Let $|\phi\rangle$ be a random* Haar *vector of dimension d. We have*

$$\mathbb{E}_\phi\left[\mathrm{F}(\mathcal{N}(|\phi\rangle\langle\phi|), |\phi\rangle\langle\phi|)\right] = \frac{1 + d\ \mathrm{F}(\mathcal{J}_{\mathcal{N}}, |\Psi\rangle\langle\Psi|)}{1 + d}. \tag{4.4}$$

Proof. Using Weingarten calculus [Gu13; CMS12], we have

$$\mathbb{E}_\phi\left[\mathrm{F}(\mathcal{N}(|\phi\rangle\langle\phi|), |\phi\rangle\langle\phi|)\right] = \sum_k \mathbb{E}_U\left[\langle 0|\, U^\dagger A_k U\, |0\rangle\langle 0|\, U^\dagger A_k^\dagger U\, |0\rangle\right] \tag{4.5}$$

$$= \sum_k \frac{(\mathrm{Tr}(A_k A_k^\dagger) + \mathrm{Tr}(A_k)\mathrm{Tr}(A_k^\dagger))}{d(d+1)} \tag{4.6}$$

$$= \frac{d + \sum_k |\mathrm{Tr}(A_k)|^2}{d(d+1)} = \frac{1 + d\ \mathrm{F}(\mathcal{J}_{\mathcal{N}}, |\Psi\rangle\langle\Psi|)}{1 + d} \tag{4.7}$$

where we use Lemma 4.3.2.

Lemma 4.3.2. *Let $\mathcal{N}$ be a quantum channel of Kraus operators $\{A_k\}_k$. Let $S = \sum_k |\mathrm{Tr}(A_k)|^2$. We can relate the average fidelity and S as follows:*

$$\mathrm{F}(\mathcal{J}_{\mathcal{N}}, |\Psi\rangle\langle\Psi|) = \frac{S}{d^2}. \tag{4.8}$$

Proof. We have:

$$\begin{aligned}\mathrm{F}(\mathcal{J}_{\mathcal{N}}, |\Psi\rangle\langle\Psi|) &= \frac{1}{d^2}\sum_{i,j,k,l} \langle ii|\, \mathrm{id} \otimes \mathcal{N}(|kk\rangle\langle ll|)\, |jj\rangle = \frac{1}{d^2}\sum_{i,j,k,l} \langle i|\, I\, |k\rangle\langle l|\, |j\rangle\langle i|\, \mathcal{N}(|k\rangle\langle l|)\, |j\rangle \\ &= \frac{1}{d^2}\sum_{i,j} \langle i|\, \mathcal{N}(|i\rangle\langle j|)\, |j\rangle = \frac{1}{d^2}\sum_{i,j,k} \langle i|\, A_k\, |i\rangle\langle j|\, A_k^\dagger\, |j\rangle = \frac{1}{d^2}\sum_k |\mathrm{Tr}(A_k)|^2 = \frac{S}{d^2}.\end{aligned} \tag{4.9}$$

□

□

Algorithm 7 Testing identity to identity in the diamond/trace distance

$N = \mathcal{O}(d/\varepsilon^4)$ (replace with $N = \mathcal{O}(d/\varepsilon^2)$ for testing in the trace distance).
for $k = 1 : N$ **do**
  Sample ϕ_k a Haar random vector in $\mathbf{S}^d$.
  Measure the output state $\mathcal{N}(|\phi_k\rangle\langle\phi_k|)$ using the POVM $\{|\phi_k\rangle\langle\phi_k|, \mathbb{I} - |\phi_k\rangle\langle\phi_k|\}$.
  Observe $X_k \sim \mathrm{Bern}(1 - \langle\phi_k|\mathcal{N}(|\phi_k\rangle\langle\phi_k|)|\phi_k\rangle)$.
end for
if $\exists k : X_k = 1$ **then return** $\mathcal{N}$ is ε-far from id **else return** $\mathcal{N} = \mathrm{id}$.

This Lemma is well known because it relates the average fidelity $\mathbb{E}_{|\phi\rangle\sim\mathrm{Haar}}[\mathrm{F}(\mathcal{N}(|\phi\rangle\langle\phi|), |\phi\rangle\langle\phi|))]$ and the entanglement fidelity $\mathrm{F}(\mathcal{J}_{\mathcal{N}}, |\Psi\rangle\langle\Psi|)$. If we measure using the measurement device $\mathcal{M}_\phi = \{|\phi\rangle\langle\phi|, \mathbb{I} - |\phi\rangle\langle\phi|\}$, the error probability under H_0 is 0, and under H_1 is $\langle\phi|\mathcal{N}(|\phi\rangle\langle\phi|)|\phi\rangle = \mathrm{F}(\mathcal{N}(|\phi\rangle\langle\phi|), |\phi\rangle\langle\phi|)$. The algorithm is detailed in Algorithm 7. The following Lemma relates the entanglement fidelity and the diamond/trace distance which is crucial for the correctness of Algorithm 7.

Lemma 4.3.3. *We have for all quantum channels $\mathcal{N}$:*

$$\mathrm{F}(\mathcal{J}_{\mathcal{N}}, |\Psi\rangle\langle\Psi|) \leq 1 - \frac{d_{\mathrm{Tr}}(\mathcal{N}, \mathrm{id})^2}{4d} \leq 1 - \frac{d_\diamond(\mathcal{N}, \mathrm{id})^4}{16d}. \tag{4.10}$$

Proof. The following inequality permits to prove the second inequality [Wat18, Theorem 3.56, rephrased]

$$d_{\mathrm{Tr}}(\mathcal{N}, \mathrm{id}) \geq \frac{d_\diamond(\mathcal{N}, \mathrm{id})^2}{2}. \tag{4.11}$$

It remains to prove the first inequality. For this, let $\varepsilon = d_{\mathrm{Tr}}(\mathcal{N}, \mathrm{id}) = \max_{|\phi\rangle\in\mathbf{S}^d} \|\mathcal{N}(|\phi\rangle\langle\phi|) - |\phi\rangle\langle\phi|\|_1$. Let $|\phi\rangle$ be a unit vector satisfying the previous maximization, we show that using Fuchs–van de Graaf inequality [FVDG99]:

$$\langle\phi|\mathcal{N}(|\phi\rangle\langle\phi|)|\phi\rangle = \mathrm{F}(\mathcal{N}(|\phi\rangle\langle\phi|), |\phi\rangle\langle\phi|) \leq 1 - \frac{1}{4}\|\mathcal{N}(|\phi\rangle\langle\phi|) - |\phi\rangle\langle\phi|\|_1^2 \leq 1 - \frac{\varepsilon^2}{4}. \tag{4.12}$$

On the other hand, we use the Kraus decomposition to describe the quantum channel $\mathcal{N}(\rho) = \sum_k A_k \rho A_k^\dagger$. We can write the previous fidelity in terms of the Kraus operators:

$$\langle\phi|\mathcal{N}(|\phi\rangle\langle\phi|)|\phi\rangle = \sum_k \langle\phi|A_k|\phi\rangle\langle\phi|A_k^\dagger|\phi\rangle = \sum_k |\langle\phi|A_k|\phi\rangle|^2. \tag{4.13}$$

Hence:

$$\sum_k |\langle\phi|A_k|\phi\rangle|^2 \leq 1 - \frac{\varepsilon^2}{4}. \tag{4.14}$$

Let $|\phi_1\rangle = |\phi\rangle$ and we can complete it to have an ortho-normal basis $\{|\phi_i\rangle\}_{i=1}^d$. Moreover, we have $\mathrm{F}(\mathcal{J}_{\mathcal{N}}, |\Psi\rangle\langle\Psi|) = \frac{1}{d^2}\sum_k |\mathrm{Tr}(A_k)|^2$ (Lemma 4.3.2). By applying the Cauchy-

Schwarz inequality and using Equation (4.14):

$$\sum_k |\mathrm{Tr}(A_k)|^2 = \sum_k \left| \sum_{i=1}^d \langle\phi_i| A_k |\phi_i\rangle \right|^2 \leq \sum_{k,i} d |\langle\phi_i| A_k |\phi_i\rangle|^2 \tag{4.15}$$

$$= d \sum_k |\langle\phi_1| A_k |\phi_1\rangle|^2 + d \sum_{i=2}^d \sum_k |\langle\phi_i| A_k |\phi_i\rangle|^2 \tag{4.16}$$

$$\leq d(1 - \varepsilon^2/4) + d(d-1) = d(d - \varepsilon^2/4) \tag{4.17}$$

because for all $i \geq 2$:

$$\sum_k |\langle\phi_i| A_k |\phi_i\rangle|^2 \leq \sum_k \langle\phi_i| A_k^\dagger A_k |\phi_i\rangle = \langle\phi_i| \sum_k A_k^\dagger A_k |\phi_i\rangle = 1. \tag{4.18}$$

Finally, $\mathrm{F}(\mathcal{J}_\mathcal{N}, |\Psi\rangle\langle\Psi|) = \frac{1}{d^2} \sum_k |\mathrm{Tr}(A_k)|^2 \leq 1 - \frac{\varepsilon^2}{4d}$. □

We can upper bound the error probability under H_1 using the well-known Lemma 4.3.1 and our Lemma 4.3.3:

$$\mathbb{E}_\phi \left[\mathrm{F}(\mathcal{N}(|\phi\rangle\langle\phi|), |\phi\rangle\langle\phi|)\right] = \frac{1 + d\ \mathrm{F}(\mathcal{J}_\mathcal{N}, |\Psi\rangle\langle\Psi|)}{1+d} \leq 1 - \frac{d_{\mathrm{Tr}}(\mathcal{N}, \mathrm{id})^2}{4(d+1)} \leq 1 - \frac{d_\diamond(\mathcal{N}, \mathrm{id})^4}{16(d+1)}. \tag{4.19}$$

Observe that the standard inequality $\|\mathcal{J}_\mathcal{N} - \mathcal{J}_\mathcal{M}\|_1 \geq \frac{d_\diamond(\mathcal{N},\mathcal{M})}{d}$ implies: $\mathbb{E}_\phi \left[\mathrm{F}(\mathcal{N}(|\phi\rangle\langle\phi|), |\phi\rangle\langle\phi|)\right] \leq 1 - \frac{d_\diamond(\mathcal{N},\mathrm{id})^2}{4d(d+1)}$ which has a better dependency in the diamond distance but a worst dependency in the dimension d. This simple lemma (Lemma 4.3.3), which relates the entanglement fidelity and the trace/diamond distance when one of the channels is unitary might be of independent interest. To obtain a 1/3 correct algorithm it suffices to repeat the described procedure using $N = \mathcal{O}(d/\varepsilon^2)$ independent copies of $|\phi\rangle\langle\phi|$ in the case of trace distance and $N = \mathcal{O}(d/\varepsilon^4)$ independent copies of $|\phi\rangle\langle\phi|$ in the case of diamond distance. Indeed, for instance for the trace distance, the probability of error under H_1 can be controlled as follows:

$$\mathbb{P}_{H_1}(\text{error}) = \mathbb{P}_{H_1}(\forall k \in [N] : X_k = 0) = \prod_{k=1}^N \mathbb{P}_{H_1}(X_k = 0) \tag{4.20}$$

$$\leq \left(1 - \frac{\varepsilon^2}{4(d+1)}\right)^N \leq \exp\left(-\frac{\varepsilon^2 N}{4(d+1)}\right) \leq \frac{1}{3} \tag{4.21}$$

for $N = 4\log(3)(d+1)/\varepsilon^2 = \mathcal{O}(d/\varepsilon^2)$. A similar proof shows that $\mathcal{O}(d/\varepsilon^4)$ copies are sufficient to test in the diamond distance. This concludes the correctness of Algorithm 7.

We summarize the main upper bound of this section in the following theorem.

Theorem 4.3.1. *There is an ancilla-free algorithm for testing identity to identity in the trace distance using only $N = \mathcal{O}\left(\frac{d}{\varepsilon^2}\right)$ incoherent measurements. Moreover, this algorithm can also solve the testing identity to identity problem in the diamond distance using only $N = \mathcal{O}\left(\frac{d}{\varepsilon^4}\right)$ incoherent measurements.*

A matching lower bound of $N = \Omega\left(\frac{d}{\varepsilon^2}\right)$ can be proved in the worst case setting.

Theorem 4.3.2. *Any adaptive ancilla-free strategy using incoherent measurements requires a number of steps satisfying:*

$$N = \Omega\left(\frac{d}{\varepsilon^2}\right) \tag{4.22}$$

to distinguish between $\mathcal{N} = \mathrm{id}$ *and* $d_\diamond(\mathcal{N}, \mathrm{id}) > \varepsilon$ *with a probability at least* $2/3$.

This theorem shows that Algorithm 7 has an optimal complexity. Interestingly, the analogous classical problem, testing identity to identity has a complexity $\Theta(d/\varepsilon)$. Thus, for this task, when going to the quantum case, the dependence on the dimension d remains the same whereas the dependency in the precision parameter ε changes from ε to ε^2. In fact, obtaining the correct ε^2 dependence is the main difficulty in this lower bound. It requires a carefully chosen construction inspired by the skew divergence [Aud14] and a fine analysis using Weingarten calculus. Even though here we are interested in ancilla-free strategies, this lower bound applies also for ancilla assisted strategies.

Proof. Under the null hypothesis H_0, the quantum channel $\mathcal{N} = \mathrm{id}$. Under the alternate hypothesis H_1, we can choose $\mathcal{N}$ so that $d_\diamond(\mathcal{N}, \mathrm{id}) \geq d_{\mathrm{Tr}}(\mathcal{N}, \mathrm{id}) \geq \varepsilon$. A difficult to test channel is a channel sending almost every vector of a basis to itself. With this intuition, we choose $V \in \mathrm{Haar}(d)$, and construct the channel $\mathcal{N}_V(\rho) = \frac{1}{2}\rho + \frac{1}{2}U_V \rho U_V^\dagger$ where U_V satisfies:

$$U_V V |l\rangle = \begin{cases} \sqrt{1-\varepsilon^2} V |0\rangle + \varepsilon V |1\rangle & \text{if } l = 0 \\ \sqrt{1-\varepsilon^2} V |1\rangle - \varepsilon V |0\rangle & \text{if } l = 1 \\ V |l\rangle & \text{otherwise.} \end{cases}$$

Taking a mixture of the identity channel and the unitary channel $U_V \cdot U_V^\dagger$ in the definition of $\mathcal{N}_V$ is crucial in this proof and is inspired by the quantum skew divergence [Aud14]. We need to show first that such a channel is ε-far from the identity channel. Indeed, let $|\phi\rangle = V |0\rangle$, we have:

$$\begin{aligned} d_\diamond(\mathcal{N}_V, \mathrm{id}) \geq d_{\mathrm{Tr}}(\mathcal{N}_V, \mathrm{id}) &\geq \| \mathcal{N}_V(|\phi\rangle\langle\phi|) - \mathrm{id}(|\phi\rangle\langle\phi|) \|_1 \\ &= \left\| \frac{1}{2} |\phi\rangle\langle\phi| + \frac{1}{2} U_V |\phi\rangle\langle\phi| U_V^\dagger - |\phi\rangle\langle\phi| \right\|_1 && (4.23) \\ &= \frac{1}{2} \left\| V |0\rangle\langle 0| V^\dagger - U_V V |0\rangle\langle 0| V^\dagger U_V^\dagger \right\|_1 && (4.24) \\ &= \frac{1}{2} \left\| |0\rangle\langle 0| - \left(\sqrt{1-\varepsilon^2} |0\rangle + \varepsilon |1\rangle\right)\left(\sqrt{1-\varepsilon^2} \langle 0| + \varepsilon \langle 1|\right) \right\|_1 && (4.25) \\ &= \frac{1}{2} \left\| \varepsilon^2 |0\rangle\langle 0| - \varepsilon\sqrt{1-\varepsilon^2}(|0\rangle\langle 1| + |1\rangle\langle 0|) - \varepsilon^2 |1\rangle\langle 1| \right\|_1 = \varepsilon. \end{aligned}$$

Hence a $1/3$-correct algorithm should distinguish between the identity channel and $\mathcal{N}_V$ with at least a probability $2/3$ of success. This algorithm can only choose an input ρ_t at each step t and perform a measurement using the POVM $\mathcal{M}_t = \{\lambda_i |\phi_i\rangle\langle\phi_i|\}_{i \in \mathcal{I}_t}$ on the output quantum state $\mathcal{N}(\rho_t)$. These choices can depend on the previous observations, that is, the algorithm can be adaptive. Let $I_{\leq N} = (I_1, \ldots, I_N)$ be the observations of this algorithm where N is a sufficient number of steps to decide correctly with a probability at least $2/3$. We can compare the distributions of the observations under the two hypotheses using the Kullback-Leibler divergence. Let P (resp. Q) be the distribution of $(I_1, \ldots, I_N)$

under H_0 (resp. H_1). The distribution of $(I_1, \ldots, I_N)$ under H_0 is:

$$P := \left\{ \prod_{t=1}^{N} \lambda_{i_t} \langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle \right\}_{i_1,\ldots,i_N}. \tag{4.26}$$

Moreover, the distribution of $(I_1, \ldots, I_N)$ under H_1 conditioned on V is:

$$Q_V := \left\{ \prod_{t=1}^{N} \lambda_{i_t} \langle \phi_{i_t} | \mathcal{N}_V(\rho_t) | \phi_{i_t} \rangle \right\}_{i_1,\ldots,i_N}. \tag{4.27}$$

The KL divergence between P and Q_V can be expressed as follows:

$$\mathrm{KL}(P\|Q_V) = \mathbb{E}_{i\sim P}(-\log)\left(\frac{Q_{V,i}}{P_i}\right) \tag{4.28}$$

$$= \sum_{t=1}^{N} \mathbb{E}_{i\leq N}(-\log)\left(\frac{\langle \phi_{i_t} | \mathcal{N}_V(\rho_t) | \phi_{i_t} \rangle}{\langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle}\right) = \sum_{t=1}^{N} \mathbb{E}_{i\leq t}(-\log)\left(\frac{\langle \phi_{i_t} | \mathcal{N}_V(\rho_t) | \phi_{i_t} \rangle}{\langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle}\right) \tag{4.29}$$

where we use the notation for $t \in [N]$, $\mathbb{E}_{i\leq t}(X(i_1, \ldots, i_t)) = \sum_{i_1,\ldots,i_t} \prod_{k=1}^{t} \lambda_{i_k} \langle \phi_{i_k} | \rho_k | \phi_{i_k} \rangle X(i_1, \ldots, i_t)$ and the fact that the term $\frac{\langle \phi_{i_t} | \mathcal{N}_V(\rho_t) | \phi_{i_t} \rangle}{\langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle}$ depends only on $(i_1, \ldots, i_t)$.
Let $\mathcal{E}$ be the event that the algorithm accepts H_0, we apply the Data-Processing inequality on the KL divergence:

$$\mathrm{KL}(P\|Q_V) \geq \mathrm{KL}(P(\mathcal{E})\|Q_V(\mathcal{E})) \tag{4.30}$$

$$\geq \mathrm{KL}(2/3\|1/3) = \frac{2}{3}\log(2) - \frac{1}{3}\log(2) = \frac{1}{3}\log(2) \tag{4.31}$$

where $\mathrm{KL}(p||q) = \mathrm{KL}(\mathrm{Bern}(p)\|\,\mathrm{Bern}(q))$. Hence

$$\mathbb{E}_{V\sim \mathrm{Haar}(d)}\, \mathrm{KL}(P\|Q_V) \geq \frac{1}{3}\log(2). \tag{4.32}$$

Let $M_V = \mathbb{I} - U_V$ and $S_V = \mathbb{I} - \frac{1}{2}M_V$. We can write the logarithmic term in the expression of $\mathrm{KL}(P\|Q_V)$ as follows:

$$(-\log)\left(\frac{\langle \phi_{i_t} | \mathcal{N}_V(\rho_t) | \phi_{i_t} \rangle}{\langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle}\right) = (-\log)\left(\frac{\langle \phi_{i_t} | (\frac{1}{2}\rho_t + \frac{1}{2}U_V\rho_t U_V^\dagger) | \phi_{i_t} \rangle}{\langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle}\right) \tag{4.33}$$

$$= (-\log)\left(1 - \frac{\mathrm{Re}(\langle \phi_{i_t} | M_V\rho_t | \phi_{i_t} \rangle)}{\langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle} + \frac{1}{2}\frac{\langle \phi_{i_t} | M_V\rho_t M_V^\dagger | \phi_{i_t} \rangle}{\langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle}\right) \tag{4.34}$$

$$= (-\log)\left(1 - \frac{1}{2}\frac{\langle \phi_{i_t} | M_V\rho_t S_V^\dagger | \phi_{i_t} \rangle}{\langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle} - \frac{1}{2}\frac{\langle \phi_{i_t} | S_V\rho_t M_V^\dagger | \phi_{i_t} \rangle}{\langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle}\right). \tag{4.35}$$

For $t \in [N]$ and $i_{\leq t} = (i_1, \ldots, i_t)$, define the event $\mathcal{G}(t, i_{\leq t}) = \left\{ \langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle \leq \frac{\varepsilon^2}{d^2} \right\}$. We can distinguish whether the event $\mathcal{G}$ is satisfied or not:

$$\mathbb{E}_{V \sim \mathrm{Haar}(d)} \mathrm{KL}(P \| Q_V) \tag{4.36}$$

$$= \sum_{t=1}^{N} \mathbb{E}_{V \sim \mathrm{Haar}(d)} \mathbb{E}_{i \leq t} (\mathbf{1}\{\mathcal{G}(t, i_{\leq t})\} + \mathbf{1}\{\mathcal{G}^c(t, i_{\leq t})\})(-\log) \left(\frac{\langle \phi_{i_t} | \mathcal{N}_V(\rho_t) | \phi_{i_t} \rangle}{\langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle} \right). \tag{4.37}$$

Let us first analyze the setting when the event $\mathcal{G}$ holds. Fix $t \in [N]$, observe that we have the inequality:

$$\begin{aligned} (-\log) \left(\frac{\langle \phi_{i_t} | \mathcal{N}_V(\rho_t) | \phi_{i_t} \rangle}{\langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle} \right) &= (-\log) \left(\frac{\langle \phi_{i_t} | \left(\frac{1}{2} \rho_t + \frac{1}{2} U_V \rho_t U_V^\dagger \right) | \phi_{i_t} \rangle}{\langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle} \right) \\ &= (-\log) \left(\frac{1}{2} + \frac{\langle \phi_{i_t} | \left(\frac{1}{2} U_V \rho_t U_V^\dagger \right) | \phi_{i_t} \rangle}{\langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle} \right) \leq \log(2) \leq 1. \end{aligned} \tag{4.38}$$

Then we can control the expectation under the event $\mathcal{G}$ as follows:

$$\begin{aligned} &\mathbb{E}_{V \sim \mathrm{Haar}(d)} \mathbb{E}_{i \leq t} \mathbf{1}\{\mathcal{G}(t, i_{\leq t})\} (-\log) \left(\frac{\langle \phi_{i_t} | \mathcal{N}_V(\rho_t) | \phi_{i_t} \rangle}{\langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle} \right) \\ &\leq \mathbb{E}_{V \sim \mathrm{Haar}(d)} \mathbb{E}_{i \leq t-1} \sum_{i_t} \lambda_{i_t} \langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle \mathbf{1}\{\mathcal{G}(t, i_{\leq t})\} \\ &\leq \mathbb{E}_{V \sim \mathrm{Haar}(d)} \mathbb{E}_{i \leq t-1} \sum_{i_t} \lambda_{i_t} \left(\frac{\varepsilon^2}{d^2} \right) \mathbf{1}\{\mathcal{G}(t, i_{\leq t})\} \qquad \left(\text{under } \mathcal{G} : \langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle \leq \frac{\varepsilon^2}{d^2} \right) \\ &\leq \mathbb{E}_{V \sim \mathrm{Haar}(d)} \mathbb{E}_{i \leq t-1} \sum_{i_t} \lambda_{i_t} \left(\frac{\varepsilon^2}{d^2} \right) = \frac{\varepsilon^2}{d} \end{aligned} \tag{4.39}$$

where we use $\sum_{i_t} \lambda_{i_t} = d$ which is an implication of the fact that $\mathcal{M}_t = \{\lambda_i |\phi_i\rangle\langle\phi_i|\}_{i \in \mathcal{I}_t}$ is a POVM.
On the other hand under $\mathcal{G}^c(t, i_{\leq t})$, we will use instead the inequality $(-\log)(x) \leq -(x-1) + (x-1)^2$ valid for all $x \in [\frac{1}{2}, +\infty)$. We apply this inequality for $x = \frac{\langle \phi_{i_t} | \mathcal{N}_V(\rho_t) | \phi_{i_t} \rangle}{\langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle} = \frac{1}{2} + \frac{\langle \phi_{i_t} | (\frac{1}{2} U_V \rho_t U_V^\dagger) | \phi_{i_t} \rangle}{\langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle} \geq \frac{1}{2}$, the first term of the upper bound is:

$$\begin{aligned} -(x-1) &= 1 - \frac{\langle \phi_{i_t} | \mathcal{N}_V(\rho_t) | \phi_{i_t} \rangle}{\langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle} \\ &= \frac{1}{2} \frac{\langle \phi_{i_t} | M_V \rho_t S_V^\dagger | \phi_{i_t} \rangle}{\langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle} + \frac{1}{2} \frac{\langle \phi_{i_t} | S_V \rho_t M_V^\dagger | \phi_{i_t} \rangle}{\langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle} = \mathrm{Re} \frac{\langle \phi_{i_t} | M_V \rho_t S_V^\dagger | \phi_{i_t} \rangle}{\langle \phi_{i_t} | \rho_t | \phi_{i_t} \rangle} \end{aligned} \tag{4.40}$$

and by using first the inequality $(x+y)^2 \leq 2(x^2+y^2)$ and then the Cauchy Schwarz inequality applied for the vectors $\sqrt{\rho_t} |\phi_{i_t}\rangle$ and $\sqrt{\rho_t} M_V^\dagger |\phi_{i_t}\rangle$ we can upper bound the

second term as follows:

$$\begin{aligned}(x-1)^2 &= \left(\frac{\langle\phi_{i_t}|\,\mathcal{N}_V(\rho_t)\,|\phi_{i_t}\rangle}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle} - 1\right)^2 \\ &= \left(\frac{\mathrm{Re}(\langle\phi_{i_t}|\,M_V\rho_t\,|\phi_{i_t}\rangle)}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle} - \frac{1}{2}\frac{\langle\phi_{i_t}|\,M_V\rho_t M_V^\dagger\,|\phi_{i_t}\rangle}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\right)^2 \\ &\leq 2\left(\frac{|\,\langle\phi_{i_t}|\,M_V\rho_t\,|\phi_{i_t}\rangle\,|}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\right)^2 + 2\left(\frac{1}{2}\frac{\langle\phi_{i_t}|\,M_V\rho_t M_V^\dagger\,|\phi_{i_t}\rangle}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\right)^2 \\ &\leq 2\left(\frac{\langle\phi_{i_t}|\,M_V\rho_t M_V^\dagger\,|\phi_{i_t}\rangle}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\right) + 2\left(\frac{1}{2}\frac{\langle\phi_{i_t}|\,M_V\rho_t M_V^\dagger\,|\phi_{i_t}\rangle}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\right)^2. \end{aligned} \tag{4.41}$$

Let us compute the expectation of (4.40). Let M, S such that $M_V = VMV^\dagger$ and $S_V = VSV^\dagger$. Concretely

$$M = \left(\begin{array}{cc|c} 1-\sqrt{1-\varepsilon^2} & -\varepsilon & 0 \\ \varepsilon & 1-\sqrt{1-\varepsilon^2} & \\ \hline 0 & & 0_{d-2} \end{array}\right) \quad \text{and} \quad S = \left(\begin{array}{cc|c} \frac{1}{2}+\frac{\sqrt{1-\varepsilon^2}}{2} & \frac{\varepsilon}{2} & 0 \\ \frac{-\varepsilon}{2} & \frac{1}{2}+\frac{\sqrt{1-\varepsilon^2}}{2} & \\ \hline 0 & & \mathbb{I}_{d-2} \end{array}\right).$$

Note that $\mathrm{Tr}(M) = 2(1-\sqrt{1-\varepsilon^2})$, $\mathrm{Tr}(S) = d-1+\sqrt{1-\varepsilon^2}$, $\mathrm{Tr}(MS^\dagger) = \mathrm{Tr}(M^\dagger S) = 0$ and $MM^\dagger = M + M^\dagger = M^\dagger M$. Let $\varepsilon' = (1-\sqrt{1-\varepsilon^2}) = \Theta(\varepsilon^2)$, we have by Weingarten calculus [Gu13]:

$$\begin{aligned} &\left|\mathbb{E}_{V\sim\mathrm{Haar}(d)}\left(\mathrm{Re}\,\frac{\langle\phi_{i_t}|\,M_V\rho_t S_V^\dagger\,|\phi_{i_t}\rangle}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\right)\right| \\ &= \left|\frac{1}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\,\mathrm{Re}\,\mathbb{E}_{V\sim\mathrm{Haar}(d)}\left(\langle\phi_{i_t}|\,VMV^\dagger\rho_t VS^\dagger V^\dagger\,|\phi_{i_t}\rangle\right)\right| \\ &= \left|\frac{1}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\,\mathrm{Re}\sum_{\alpha,\beta\in\mathfrak{S}_2}\mathrm{Wg}(\alpha\beta)\mathrm{Tr}_\alpha(M,S^\dagger)\mathrm{Tr}_{\beta(12)}(\rho_t,|\phi_{i_t}\rangle\langle\phi_{i_t}|)\right| \\ &= \left|\frac{1}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\,\mathrm{Re}\left(\frac{d\mathrm{Tr}(M)\mathrm{Tr}(S)\,\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle - \mathrm{Tr}(M)\mathrm{Tr}(S^\dagger)}{d(d^2-1)}\right)\right| \\ &= \left|\frac{1}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\,\mathrm{Re}\left(\frac{2d\varepsilon'(d-\varepsilon')\,\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle - 2\varepsilon'(d-\varepsilon')}{d(d^2-1)}\right)\right| \\ &\leq \frac{2\varepsilon^2}{d} + \frac{2\varepsilon^4}{d^3\,\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle} + \frac{2\varepsilon^2}{(d^2-1)\,\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle} \leq \frac{2\varepsilon^2}{d} + \frac{4\varepsilon^2}{d^2\,\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}. \end{aligned} \tag{4.42}$$

Recall the notation $\mathbb{E}_{i\leq t}(X(i_1,\ldots,i_t)) = \sum_{i_1,\ldots,i_t}\prod_{k=1}^t \lambda_{i_k}\,\langle\phi_{i_k}|\,\rho_k\,|\phi_{i_k}\rangle\,X(i_1,\ldots,i_t)$. If we

take the expectation $\mathbb{E}_{i\leq t}$ under the event $\mathcal{G}^c(t, i_{\leq t})$, we obtain

$$\begin{aligned}
&\mathbb{E}_{V\sim\mathrm{Haar}(d)}\mathbb{E}_{i\leq t}\mathbf{1}\{\mathcal{G}^c(t,i_t)\}\left(\mathrm{Re}\,\frac{\langle\phi_{i_t}|\,M_V\rho_t S_V^\dagger\,|\phi_{i_t}\rangle}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\right)\\
&\leq \mathbb{E}_{i\leq t}\mathbf{1}\{\mathcal{G}^c(t,i_t)\}\left|\mathbb{E}_{V\sim\mathrm{Haar}(d)}\left(\mathrm{Re}\,\frac{\langle\phi_{i_t}|\,M_V\rho_t S_V^\dagger\,|\phi_{i_t}\rangle}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\right)\right|\\
&\leq \mathbb{E}_{i\leq t}\mathbf{1}\{\mathcal{G}^c(t,i_t)\}\left(\frac{2\varepsilon^2}{d}+\frac{4\varepsilon^2}{d^2\,\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\right)\\
&\leq \mathbb{E}_{i\leq t}\left(\frac{2\varepsilon^2}{d}+\frac{4\varepsilon^2}{d^2\,\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\right)\\
&= \frac{2\varepsilon^2}{d}+\mathbb{E}_{i_{\leq t-1}}\sum_{i_t}\lambda_{i_t}\,\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle\times\frac{4\varepsilon^2}{d^2\,\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}=\frac{6\varepsilon^2}{d}
\end{aligned}\tag{4.43}$$

where we use $\sum_{i_t}\lambda_{i_t}=d$. We move to the expectation $\mathbb{E}_{i\leq t}$ of the first term of (4.41), it is non negative so we can safely remove the condition $\mathbf{1}\{\mathcal{G}^c(t,i_t)\}$:

$$\begin{aligned}
&\mathbb{E}_V\mathbb{E}_{i\leq t}\mathbf{1}\{\mathcal{G}^c(t,i_t)\}\frac{2\,\langle\phi_{i_t}|\,M_V\rho_t M_V^\dagger\,|\phi_{i_t}\rangle}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\leq\mathbb{E}_V\mathbb{E}_{i\leq t}\frac{2\,\langle\phi_{i_t}|\,M_V\rho_t M_V^\dagger\,|\phi_{i_t}\rangle}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\\
&=\mathbb{E}_V\mathbb{E}_{i\leq t-1}\sum_{i_t}2\lambda_{i_t}\,\langle\phi_{i_t}|\,M_V\rho_t M_V^\dagger\,|\phi_{i_t}\rangle\\
&=\mathbb{E}_{i\leq t-1}\mathbb{E}_V 2\mathrm{Tr}(M_V\rho_t M_V^\dagger)=\mathbb{E}_{i\leq t-1}\mathbb{E}_V 2\mathrm{Tr}((M_V+M_V^\dagger)\rho_t)\\
&=\mathbb{E}_{i\leq t-1}\frac{8\varepsilon'}{d}\leq\frac{8\varepsilon^2}{d}
\end{aligned}\tag{4.44}$$

because $\mathbb{E}_V M_V=\frac{2(1-\sqrt{1-\varepsilon^2})}{d}\mathbb{I}\preccurlyeq\frac{2\varepsilon^2}{d}\mathbb{I}$.
Concerning the expectation of the second term of (4.41), we apply again the Weingarten calculus [Gu13] to have:

$$\mathbb{E}_{V\sim\mathrm{Haar}(d)}\frac{1}{2}\left(\frac{\langle\phi_{i_t}|\,M_V\rho_t M_V^\dagger\,|\phi_{i_t}\rangle}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\right)^2=\frac{1}{2}\frac{\mathbb{E}_{V\sim\mathrm{Haar}(d)}\,\langle\phi_{i_t}|\,M_V\rho_t M_V^\dagger\,|\phi_{i_t}\rangle^2}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle^2}\tag{4.45}$$

$$=\frac{1}{2}\frac{\mathbb{E}_{V\sim\mathrm{Haar}(d)}\mathrm{Tr}(|\phi_{i_t}\rangle\langle\phi_{i_t}|\,VMV^\dagger\rho_t VM^\dagger V^\dagger\,|\phi_{i_t}\rangle\langle\phi_{i_t}|\,VMV^\dagger\rho_t VM^\dagger V^\dagger)}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle^2}\tag{4.46}$$

$$=\frac{1}{2\,\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle^2}\sum_{\alpha,\beta\in\mathfrak{S}_4}\mathrm{Wg}(\alpha\beta)\mathrm{Tr}_\beta(M,M^\dagger,M,M^\dagger)\mathrm{Tr}_{\alpha\gamma}(\rho_t,|\phi_{i_t}\rangle\langle\phi_{i_t}|\,,\rho_t,|\phi_{i_t}\rangle\langle\phi_{i_t}|).\tag{4.47}$$

Note that $\mathrm{Tr}_{\alpha\gamma}(\rho_t,|\phi_{i_t}\rangle\langle\phi_{i_t}|\,,\rho_t,|\phi_{i_t}\rangle\langle\phi_{i_t}|)\in\{1,\mathrm{Tr}(\rho_t^2),\langle\phi_{i_t}|\,\rho_t^2\,|\phi_{i_t}\rangle\,,\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle\,,\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle^2\}$, $\mathrm{Tr}(\rho_t^2)\leq 1$ and $\langle\phi_{i_t}|\,\rho_t^2\,|\phi_{i_t}\rangle\leq\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle\leq 1$. Moreover, it is clear that when β is not a 4-cycle, we have $|\mathrm{Tr}_\beta(M,M^\dagger,M,M^\dagger)|\leq\mathcal{O}(\varepsilon^4)$ since it can be written as a product of at least two elements each of them is $\mathcal{O}(\varepsilon^2)$. In the case β is a 4 cycle we have $\mathrm{Tr}(MM^\dagger MM^\dagger)=\mathrm{Tr}(MMM^\dagger M^\dagger)=\mathrm{Tr}((M+M^\dagger)^2)=2(2-2\sqrt{1-\varepsilon^2})^2\leq 8\varepsilon^4$. On the other hand, we know that for all $(\alpha,\beta)\in\mathfrak{S}_4^2:|\,\mathrm{Wg}(\alpha\beta)|\leq\frac{2}{d^4}$ [CŚ06] so

$$|\,\mathrm{Wg}(\alpha\beta)\mathrm{Tr}_\beta(M,M^\dagger,M,M^\dagger)\mathrm{Tr}_{\alpha\gamma}(|\phi_{i_t}\rangle\langle\phi_{i_t}|\,,\rho_t,|\phi_{i_t}\rangle\langle\phi_{i_t}|\,,\rho_t)|\leq\mathcal{O}\left(\frac{\varepsilon^4}{d^4}\right).$$

Therefore we have:

$$\mathbb{E}_{V\sim\text{Haar}(d)}\frac{1}{2}\left(\frac{\langle\phi_{i_t}|\, M_V\rho_t M_V^\dagger\,|\phi_{i_t}\rangle}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\right)^2 \le \mathcal{O}\left(\frac{\varepsilon^4}{d^4\,\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle^2}\right). \tag{4.48}$$

Now if we take the expectation $\mathbb{E}_{i\le t}$ under the event $\mathcal{G}^c(t,i_{\le t}) = \left\{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle > \frac{\varepsilon^2}{d^2}\right\}$, we obtain:

$$\begin{aligned}
&\mathbb{E}_{i_{\le t}}\mathbb{E}_{V\sim\text{Haar}(d)}\mathbf{1}\{\mathcal{G}^c(t,i_{\le t})\}\frac{1}{2}\left(\frac{\langle\phi_{i_t}|\, M_V\rho_t M_V^\dagger\,|\phi_{i_t}\rangle}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\right)^2\\
&\le \mathbb{E}_{i_{\le t}}\mathbf{1}\{\mathcal{G}^c(t,i_{\le t})\}\mathcal{O}\left(\frac{\varepsilon^4}{d^4\,\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\times\frac{1}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\right)\\
&\le \mathbb{E}_{i_{\le t}}\mathbf{1}\{\mathcal{G}^c(t,i_{\le t})\}\mathcal{O}\left(\frac{\varepsilon^4}{d^4\,\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\times\frac{d^2}{\varepsilon^2}\right) \qquad \left(\text{under } \mathcal{G}^c: \frac{1}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle} < \frac{d^2}{\varepsilon^2}\right)\\
&= \mathbb{E}_{i_{\le t-1}}\sum_{i_t}\lambda_{i_t}\,\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle\,\mathbf{1}\{\mathcal{G}^c(t,i_{\le t})\}\mathcal{O}\left(\frac{\varepsilon^2}{d^2\,\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\right)\\
&\le \mathbb{E}_{i_{\le t-1}}\sum_{i_t}\lambda_{i_t}\mathcal{O}\left(\frac{\varepsilon^2}{d^2}\right) = \mathbb{E}_{i_{\le t-1}}\mathcal{O}\left(\frac{d\varepsilon^2}{d^2}\right) = \mathcal{O}\left(\frac{\varepsilon^2}{d}\right)
\end{aligned} \tag{4.49}$$

where we use $\sum_{i_t}\lambda_{i_t} = d$. By adding up (4.43), (4.44) and (4.49), we obtain:

$$\mathbb{E}_{i_{\le t}}\mathbb{E}_{V\sim\text{Haar}(d)}\mathbf{1}\{\mathcal{G}^c(t,i_{\le t})\}(-\log)\left(\frac{\langle\phi_{i_t}|\,\mathcal{N}_V(\rho_t)\,|\phi_{i_t}\rangle}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\right) \tag{4.50}$$

$$\le \mathbb{E}_{i_{\le t}}\mathbb{E}_{V\sim\text{Haar}(d)}\mathbf{1}\{\mathcal{G}^c(t,i_{\le t})\}\left(\frac{\text{Re}(\langle\phi_{i_t}|\, M_V\rho_t S_V^\dagger\,|\phi_{i_t}\rangle)}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle} + \frac{1}{2}\frac{\langle\phi_{i_t}|\, M_V\rho_t M_V^\dagger\,|\phi_{i_t}\rangle^2}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle^2}\right) \tag{4.51}$$

$$+\,\mathbb{E}_{i_{\le t}}\mathbb{E}_{V\sim\text{Haar}(d)}\mathbf{1}\{\mathcal{G}^c(t,i_{\le t})\}\left(2\frac{\langle\phi_{i_t}|\, M_V\rho_t M_V^\dagger\,|\phi_{i_t}\rangle}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\right) = \mathcal{O}\left(\frac{\varepsilon^2}{d}\right). \tag{4.52}$$

Therefore using this upper bound and the upper bound (4.39) we get an upper bound on the expected KL divergence:

$$\mathbb{E}_{V\sim\text{Haar}(d)}\,\text{KL}(P\|Q_V) \tag{4.53}$$

$$= \sum_{t=1}^{N}\mathbb{E}_{i\le t}\mathbb{E}_{V\sim\text{Haar}(d)}(\mathbf{1}\{\mathcal{G}(t,i_{\le t})\} + \mathbf{1}\{\mathcal{G}^c(t,i_{\le t})\})(-\log)\left(\frac{\langle\phi_{i_t}|\,\mathcal{N}_V(\rho_t)\,|\phi_{i_t}\rangle}{\langle\phi_{i_t}|\,\rho_t\,|\phi_{i_t}\rangle}\right) \tag{4.54}$$

$$\le \sum_{t=1}^{N}\mathcal{O}\left(\frac{\varepsilon^2}{d}\right) + \mathcal{O}\left(\frac{\varepsilon^2}{d}\right) = \mathcal{O}\left(\frac{N\varepsilon^2}{d}\right). \tag{4.55}$$

Finally since $\mathbb{E}_{V\sim\text{Haar}(d)}\,\text{KL}(P\|Q_V) \ge \frac{\log(2)}{3}$ we conclude:

$$N = \Omega\left(\frac{d}{\varepsilon^2}\right). \tag{4.56}$$

□

4.4 Testing identity to the depolarizing channel

In this section, we move to study the problem of testing identity to the completely depolarizing channel $\mathcal{N}_0 = \mathcal{D}$ in the diamond distance. Given a precision parameter $\varepsilon > 0$ and an unknown quantum channel $\mathcal{N}$, we would like to test whether $H_0 : \mathcal{N} = \mathcal{D}$ or $H_1 : d_\diamond(\mathcal{N}, \mathcal{D}) \geq \varepsilon$ with a probability of error at most $1/3$. If $\mathcal{N} = \mathcal{D}$, the tester should answer the null hypothesis H_0 with a probability at least $2/3$ whereas if $d_\diamond(\mathcal{N}, \mathcal{D}) \geq \varepsilon$, the tester should answer the alternate hypothesis H_1 with a probability at least $2/3$. The alternate condition means that

$$d_\diamond(\mathcal{N}, \mathcal{D}) \geq \varepsilon \iff \exists\, |\phi\rangle \in \mathbf{S}^{d_{\text{in}} \times d_{\text{in}}} : \|\text{id}_{d_{\text{in}}} \otimes \mathcal{N}(|\phi\rangle\langle\phi|) - \text{id}_{d_{\text{in}}} \otimes \mathcal{D}(|\phi\rangle\langle\phi|)\|_1 \geq \varepsilon. \quad (4.57)$$

This inequality implies a lower bound of the 1-norm between the Choi states: $\|\mathcal{J}_\mathcal{N} - \mathcal{J}_\mathcal{D}\|_1 \geq \frac{\varepsilon}{d_{\text{in}}}$. The simplest idea would be to use this inequality and reduce the problem of testing channels to testing states. Actually, this kind of reduction from channels to states has been used for quantum process tomography [SSKKG22] and shadow process tomography [KTCT21]. The Choi state of the depolarizing channel $\mathcal{D}$ is $\mathcal{J}_\mathcal{D} = \frac{1}{d_{\text{in}}} \sum_{i,j=1}^{d_{\text{in}}} |i\rangle\langle j| \otimes \text{Tr}(|i\rangle\langle j|) \frac{\mathbb{I}}{d_{\text{out}}} = \frac{\mathbb{I}_{d_{\text{in}}} \otimes \mathbb{I}_{d_{\text{out}}}}{d_{\text{in}} d_{\text{out}}} = \frac{\mathbb{I}_{d_{\text{in}} d_{\text{out}}}}{d_{\text{in}} d_{\text{out}}}$, so by applying the previous inequality, we obtain for a quantum channel $\mathcal{N}$ ε-far from the depolarizing channel $\mathcal{D}$: $\left\| \mathcal{J}_\mathcal{N} - \frac{\mathbb{I}}{d_{\text{in}} d_{\text{out}}} \right\|_1 \geq \frac{\varepsilon}{d_{\text{in}}}$. Then, we can apply a reduction to the testing mixedness of quantum states [BCL20] to design an ancilla-assisted strategy requiring $\mathcal{O}\left(\frac{d_{\text{in}}^{1.5} d_{\text{out}}^{1.5}}{(\varepsilon/d_{\text{in}})^2}\right) = \mathcal{O}\left(\frac{d_{\text{in}}^{3.5} d_{\text{out}}^{1.5}}{\varepsilon^2}\right)$ independent measurements since the dimension of the states $\mathcal{J}_\mathcal{N}$ and $\frac{\mathbb{I}}{d_{\text{in}} d_{\text{out}}}$ is $d_{\text{in}} d_{\text{out}}$ and the precision parameter is $\frac{\varepsilon}{d_{\text{in}}}$. However, this approach has two problems. First, we need to be able to use an auxiliary system to prepare the Choi state $\mathcal{J}_\mathcal{N}$, which cannot be done in the ancilla-free model we consider. Next, the complexity $\mathcal{O}\left(\frac{d_{\text{in}}^{3.5} d_{\text{out}}^{1.5}}{\varepsilon^2}\right)$, as we shall see later, is not optimal. If one tries to reduce to testing identity of states in the 2-norm (Section 3.3.2) one obtains a slightly better bound but still not optimal and still using an auxiliary system.

Inspired by the testing identity to identity problem (Section 4.3), when we do not know one of the optimal input states, we choose it to be random. Let $|\phi\rangle$ be a Haar random vector. If the input state is $|\phi\rangle\langle\phi|$, the output state under H_0 is $\mathcal{D}(|\phi\rangle\langle\phi|) = \frac{\mathbb{I}}{d_{\text{out}}}$ and under H_1 is $\mathcal{N}(|\phi\rangle\langle\phi|)$. So it is natural to ask what would be the distance between $\mathcal{N}(|\phi\rangle\langle\phi|)$ and $\frac{\mathbb{I}}{d_{\text{out}}}$. Note that in general, it is much easier to compute the expectation of the 2-norms than the 1-norms. For this reason, we start by computing the expectation of the 2-norm between $\mathcal{N}(|\phi\rangle\langle\phi|)$ and $\frac{\mathbb{I}}{d_{\text{out}}}$.

Lemma 4.4.1. *Let $\mathcal{M} = \mathcal{N} - \mathcal{D}$ and $\mathcal{J}_\mathcal{M} = \text{id} \otimes \mathcal{M}(|\Psi\rangle\langle\Psi|)$. We have:*

$$\mathbb{E}_{|\phi\rangle \sim \text{Haar}} \left(\|\mathcal{M}(|\phi\rangle\langle\phi|)\|_2^2 \right) = \frac{\|\mathcal{M}(\mathbb{I})\|_2^2 + d_{\text{in}}^2 \|\mathcal{J}_\mathcal{M}\|_2^2}{d_{\text{in}}(d_{\text{in}} + 1)} \geq \frac{d_{\text{in}}}{d_{\text{in}} + 1} \|\mathcal{J}_\mathcal{M}\|_2^2 . \quad (4.58)$$

Proof. We write

$$\mathbb{E}\left(\left\| \mathcal{N}(|\phi\rangle\langle\phi|) - \frac{\mathbb{I}}{d_{\text{out}}} \right\|_2^2 \right) = \mathbb{E}\left(\text{Tr}(\mathcal{N}(|\phi\rangle\langle\phi|)^2) \right) - \frac{1}{d_{\text{out}}}. \quad (4.59)$$

then if we use the Kraus decomposition of the quantum channel $\mathcal{N}(\rho) = \sum_k A_k \rho A_k^\dagger$, we

can compute the following expectation using Weingarten calculus [Gu13; CMS12]:

$$\mathbb{E}\left(\mathrm{Tr}(\mathcal{N}(|\phi\rangle\langle\phi|)^2)\right) = \mathbb{E}\left(\mathrm{Tr}\left(\sum_k A_k |\phi\rangle\langle\phi| A_k^\dagger\right)^2\right) = \sum_{k,l}\mathbb{E}\left(\mathrm{Tr}(A_k |\phi\rangle\langle\phi| A_k^\dagger A_l |\phi\rangle\langle\phi| A_l^\dagger)\right) \tag{4.60}$$

$$= \frac{1}{d_{\mathrm{in}}(d_{\mathrm{in}}+1)}\left(\sum_{k,l}\mathrm{Tr}(A_k^\dagger A_l A_l^\dagger A_k) + \sum_{k,l}|\mathrm{Tr}(A_k A_l^\dagger)|^2\right) \tag{4.61}$$

$$= \frac{1}{d_{\mathrm{in}}(d_{\mathrm{in}}+1)}\left(\mathrm{Tr}\left(\sum_k A_k A_k^\dagger\right)^2 + \sum_{k,l}|\mathrm{Tr}(A_k A_l^\dagger)|^2\right) \tag{4.62}$$

$$= \frac{1}{d_{\mathrm{in}}(d_{\mathrm{in}}+1)}\left(\mathrm{Tr}(\mathcal{N}(\mathbb{I})^2) + \sum_{k,l}|\mathrm{Tr}(A_k A_l^\dagger)|^2\right) \tag{4.63}$$

Observe that $\mathrm{Tr}(\mathcal{N}(\mathbb{I})^2) = \mathrm{Tr}\left(\mathcal{N}(\mathbb{I}) - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}\mathbb{I}\right)^2 + \frac{d_{\mathrm{in}}^2}{d_{\mathrm{out}}}$. Moreover,

$$\mathrm{Tr}(\mathcal{J}^2) = \mathrm{Tr}\left(\frac{1}{d_{\mathrm{in}}}\sum_{i,j}|i\rangle\langle j|\otimes\mathcal{N}(|i\rangle\langle j|)\right)^2 = \frac{1}{d_{\mathrm{in}}^2}\sum_{i,j}\mathrm{Tr}(\mathcal{N}(|i\rangle\langle j|)\mathcal{N}(|j\rangle\langle i|)) \tag{4.64}$$

$$= \frac{1}{d_{\mathrm{in}}^2}\sum_{i,j,k,l}\mathrm{Tr}(A_k |i\rangle\langle j| A_k^\dagger A_l |j\rangle\langle i| A_l^\dagger) = \frac{1}{d_{\mathrm{in}}^2}\sum_{k,l}|\mathrm{Tr}(A_k^\dagger A_l)|^2 \tag{4.65}$$

hence:

$$\mathbb{E}\left(\mathrm{Tr}(\mathcal{N}(|\phi\rangle\langle\phi|)^2)\right) = \mathbb{E}\left(\mathrm{Tr}\left(\sum_k A_k |\phi\rangle\langle\phi| A_k^\dagger\right)^2\right) \tag{4.66}$$

$$= \frac{1}{d_{\mathrm{in}}(d_{\mathrm{in}}+1)}\left(\mathrm{Tr}\left(\mathcal{N}(\mathbb{I}) - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}\mathbb{I}\right)^2 + \frac{d_{\mathrm{in}}^2}{d_{\mathrm{out}}} + d_{\mathrm{in}}^2\mathrm{Tr}(\mathcal{J}^2)\right) \tag{4.67}$$

Finally,

$$\mathbb{E}\left(\left\|\mathcal{N}(|\phi\rangle\langle\phi|) - \frac{\mathbb{I}}{d_{\mathrm{out}}}\right\|_2^2\right) = \mathbb{E}\left(\mathrm{Tr}(\mathcal{N}(|\phi\rangle\langle\phi|)^2)\right) - \frac{1}{d_{\mathrm{out}}} \tag{4.68}$$

$$= \frac{1}{d_{\mathrm{in}}(d_{\mathrm{in}}+1)}\left(\mathrm{Tr}\left(\mathcal{N}(\mathbb{I}) - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}\mathbb{I}\right)^2 + \frac{d_{\mathrm{in}}^2}{d_{\mathrm{out}}} + d_{\mathrm{in}}^2\mathrm{Tr}(\mathcal{J}^2)\right) - \frac{1}{d_{\mathrm{out}}} \tag{4.69}$$

$$= \frac{1}{d_{\mathrm{in}}(d_{\mathrm{in}}+1)}\left(\mathrm{Tr}\left(\mathcal{N}(\mathbb{I}) - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}\mathbb{I}\right)^2 + d_{\mathrm{in}}^2\mathrm{Tr}(\mathcal{J}^2) - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}\right) \tag{4.70}$$

$$= \frac{1}{d_{\mathrm{in}}(d_{\mathrm{in}}+1)}\left(\mathrm{Tr}\left(\mathcal{N}(\mathbb{I}) - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}\mathbb{I}\right)^2 + d_{\mathrm{in}}^2\left\|\mathcal{J} - \frac{\mathbb{I}}{d_{\mathrm{in}}d_{\mathrm{out}}}\right\|_2^2\right). \tag{4.71}$$

□

We now need to relate the 2-norm between the Choi state of the channel $\mathcal{N}$ and the Choi state of the depolarizing channel $\|\mathcal{J}_{\mathcal{N}} - \mathcal{J}_{\mathcal{D}}\|_2$ with the diamond distance between two channels $d_\diamond(\mathcal{N},\mathcal{D})$. This is done in the following Lemma:

Lemma 4.4.2. *Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be two $(d_{\text{in}}, d_{\text{out}})$-quantum channels. We have:*

$$\|\mathcal{J}_{\mathcal{N}_1} - \mathcal{J}_{\mathcal{N}_2}\|_2 \geq \frac{d_\diamond(\mathcal{N}_1, \mathcal{N}_2)}{d_{\text{in}}\sqrt{d_{\text{out}}}}. \tag{4.72}$$

Proof. Denote by $\mathcal{M} = \mathcal{N}_1 - \mathcal{N}_2$ and $\mathcal{J} = \mathcal{J}_{\mathcal{M}} = \mathcal{J}_{\mathcal{N}_1} - \mathcal{J}_{\mathcal{N}_2}$. Let $|\phi\rangle$ be a maximizing unit vector of the diamond norm, i.e., $\|\text{id} \otimes \mathcal{M}(|\phi\rangle\langle\phi|)\|_1 = d_\diamond(\mathcal{N}_1, \mathcal{N}_2)$. We can write $|\phi\rangle = A \otimes \mathbb{I} |\Psi\rangle$ where $|\Psi\rangle = \frac{1}{\sqrt{d_{\text{in}}}} \sum_{i=1}^{d_{\text{in}}} |i\rangle \otimes |i\rangle$ is the maximally entangled state. $|\phi\rangle$ has norm 1 so $\text{Tr}(A^\dagger A) = d_{\text{in}} \langle\Psi| A^\dagger A \otimes \mathbb{I} |\Psi\rangle = d_{\text{in}} \langle\phi|\phi\rangle = d_{\text{in}}$. We can write the diamond distance as follows:

$$\begin{aligned} d_\diamond(\mathcal{N}_1, \mathcal{N}_2) &= \|\text{id} \otimes \mathcal{M}(|\phi\rangle\langle\phi|)\|_1 = \|\text{id} \otimes \mathcal{M}(A \otimes \mathbb{I} |\Psi\rangle\langle\Psi| A^\dagger \otimes \mathbb{I})\|_1 && (4.73) \\ &= \|(A \otimes \mathbb{I})\text{id} \otimes \mathcal{M}(|\Psi\rangle\langle\Psi|)(A^\dagger \otimes \mathbb{I})\|_1 = \|(A \otimes \mathbb{I})\mathcal{J}_{\mathcal{M}}(A^\dagger \otimes \mathbb{I})\|_1. && (4.74) \end{aligned}$$

$\mathcal{J}_{\mathcal{M}}$ is Hermitian so can be written as : $\mathcal{J}_{\mathcal{M}} = \sum_i \lambda_i |\psi_i\rangle\langle\psi_i|$. Using the triangle inequality and the Cauchy Schwarz inequality, we obtain:

$$\begin{aligned} \|(A \otimes \mathbb{I})\mathcal{J}_{\mathcal{M}}(A^\dagger \otimes \mathbb{I})\|_1 &= \left\| (A \otimes \mathbb{I}) \sum_i \lambda_i |\psi_i\rangle\langle\psi_i| (A^\dagger \otimes \mathbb{I}) \right\|_1 && (4.75) \\ \leq \sum_i |\lambda_i| \|(A \otimes \mathbb{I}) |\psi_i\rangle\langle\psi_i| (A^\dagger \otimes \mathbb{I})\|_1 &\leq \sqrt{\sum_i \lambda_i^2} \sqrt{\sum_i \langle\psi_i| (A^\dagger A \otimes \mathbb{I}) |\psi_i\rangle^2} && (4.76) \\ \leq \|\mathcal{J}\|_2 \sqrt{\sum_i \langle\psi_i| (A^\dagger A A^\dagger A \otimes \mathbb{I}) |\psi_i\rangle} &= \|\mathcal{J}_{\mathcal{M}}\|_2 \sqrt{\text{Tr}(A^\dagger A A^\dagger A \otimes \mathbb{I})} && (4.77) \\ = \|\mathcal{J}_{\mathcal{M}}\|_2 \sqrt{d_{\text{out}} \text{Tr}(A^\dagger A A^\dagger A)} \leq \|\mathcal{J}_{\mathcal{M}}\|_2 \sqrt{d_{\text{out}} (\text{Tr}(A^\dagger A))^2} &= d_{\text{in}} \sqrt{d_{\text{out}}} \|\mathcal{J}_{\mathcal{M}}\|_2. && (4.78) \end{aligned}$$

□

Let $\mathcal{N}$ be a channel satisfying $d_\diamond(\mathcal{N}, \mathcal{D}) \geq \varepsilon$ and $X = \left\| \mathcal{N}(|\phi\rangle\langle\phi|) - \frac{\mathbb{I}}{d_{\text{out}}} \right\|_2^2$, we obtain from Lemmas 4.4.1 and 4.4.2:

$$\mathbb{E}(X) = \mathbb{E}_{|\phi\rangle \sim \text{Haar}} \left(\left\| \mathcal{N}(|\phi\rangle\langle\phi|) - \frac{\mathbb{I}}{d_{\text{out}}} \right\|_2^2 \right) \geq \frac{d_{\text{in}}}{d_{\text{in}} + 1} \left\| \mathcal{J}_{\mathcal{N}} - \frac{\mathbb{I}}{d_{\text{in}} d_{\text{out}}} \right\|_2^2 \geq \frac{\varepsilon^2}{2 d_{\text{in}}^2 d_{\text{out}}}. \tag{4.79}$$

If we could show that X is larger than $\Omega(\mathbb{E}(X))$ with constant probability, then we can reduce our problem to the usual testing identity of quantum states in the 2-norm (quantum state certification in Section 3.3.2) and obtain an *ancilla-free* algorithm using $\mathcal{O}\left(\frac{\sqrt{d_{\text{out}}}}{(\varepsilon/d_{\text{in}}\sqrt{d_{\text{out}}})^2}\right) = \mathcal{O}\left(\frac{d_{\text{in}}^2 d_{\text{out}}^{1.5}}{\varepsilon^2}\right)$ independent measurements. Establishing this turns out to be the most technical part of the proof as is summarized in the following theorem.

Theorem 4.4.1. *Let $|\phi\rangle$ be a* Haar *distributed vector in $\mathbf{S}^d$ (or any 4-design). Let $X = \left\| \mathcal{N}(|\phi\rangle\langle\phi|) - \frac{\mathbb{I}}{d_{\text{out}}} \right\|_2^2$. We have:*

$$\text{Var}(X) = \mathcal{O}\left([\mathbb{E}(X)]^2\right). \tag{4.80}$$

This theorem is the most technical part of this chapter and we believe that it can be generalized for any difference of channels with a similar approach. Moreover, applying this inequality along with the Paley-Zygmund inequality are sufficient for our reduction: we only need to repeat our test $\mathcal{O}(1)$ times to reduce the error probability to $1/3$ for testing identity to the depolarizing channel.

Outline of the proof. We use the Kraus decomposition for the quantum channel $\mathcal{N}$:

$$\mathcal{N}(\rho) = \sum_k A_k \rho A_k^\dagger. \tag{4.81}$$

We observe first that $\mathrm{Var}(X) = \mathrm{Var}\left(\mathrm{Tr}\left[\left(\mathcal{N}(|\phi\rangle\langle\phi|)\right)^2\right]\right)$. Then using Weingarten calculus [Gu13; CMS12], we can compute the expectation

$$\mathbb{E}\left(\left(\mathrm{Tr}\left[\left(\mathcal{N}(|\phi\rangle\langle\phi|)\right)^2\right]\right)^2\right) = \frac{1}{d_{\mathrm{in}}(d_{\mathrm{in}}+1)(d_{\mathrm{in}}+2)(d_{\mathrm{in}}+3)} \sum_{\alpha\in\mathfrak{S}_4} F(\alpha) \tag{4.82}$$

where for $\alpha \in \mathfrak{S}_4$,

$$F(\alpha) = \sum_{i,j,k,l,k',l'} \mathrm{Tr}_\alpha(A_{l'}^\dagger |j\rangle \langle i| A_k, A_k^\dagger A_l, A_l^\dagger |i\rangle \langle j| A_{k'}, A_{k'}^\dagger A_{l'}) \tag{4.83}$$

and $\mathrm{Tr}_\alpha(M_1, \ldots, M_n) = \Pi_j \mathrm{Tr}(\Pi_{i\in C_j} M_i)$ for $\alpha = \Pi_j C_j$ and C_j are cycles. Let $m = \left\|\mathcal{N}(\mathbb{I}) - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}\mathbb{I}\right\|_2$ and $\eta = d_{\mathrm{in}}\left\|\mathcal{J}_\mathcal{N} - \frac{\mathbb{I}}{d_{\mathrm{in}}d_{\mathrm{out}}}\right\|_2$. Next, we upper bound the function F as shown in Table 4.2. This is the hardest step of the proof and requires a fine analysis

Permutation α	**Upper bound on** $F(\alpha)$	**Reference**
(13)	$\left(\frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} + \eta^2\right)^2$	
$\mathrm{id}, (132), (314), (24)(13)$	$\frac{d_{\mathrm{in}}/d_{\mathrm{out}}+\eta^2}{d_{\mathrm{out}}} + \frac{\eta^2}{d_{\mathrm{out}}} + 5\eta^4$	(Lemma 4.4.7)
$(312), (134)$	$\left(\frac{d_{\mathrm{in}}^2}{d_{\mathrm{out}}} + m^2\right)\left(\frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} + \eta^2\right)$	
(1234)	$\left(\frac{d_{\mathrm{in}}^2}{d_{\mathrm{out}}} + m^2\right)^2$	
$(24), (1432)$	$\frac{d_{\mathrm{in}}^2}{d_{\mathrm{out}}^2} + \frac{2m^2}{d_{\mathrm{out}}} + 25\eta^4$	(Lemma 4.4.8)
$(142), (243)$	$\frac{d_{\mathrm{in}}/d_{\mathrm{out}}+\eta^2}{d_{\mathrm{out}}} + \frac{\eta^2}{d_{\mathrm{out}}} + 5\eta^4$	(Lemma 4.4.6)
$(14), (12), (23), (34), (1324), (1423), (1243), (1342)$	$\frac{d_{\mathrm{in}}^2}{d_{\mathrm{out}}^2} + \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}\eta^2 + \frac{m^2}{d_{\mathrm{out}}} + 5m\eta^3$	(Lemma 4.4.10)
$(12)(34), (14)(23), (234), (124)$	$\frac{d_{\mathrm{in}}^3}{d_{\mathrm{out}}^2} + 2\frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}m^2 + m^2\eta^2$	(Lemma 4.4.9)

Table 4.2: Upper bounds on the function F (defined in (4.83)) for different input permutations.

for many of the $F(\alpha)'s$. A particularly useful trick we use repeatedly is known as the *replica trick* and says that if $\mathbb{F} = \sum_{i,j=1}^d (|i\rangle \otimes |j\rangle)(\langle j| \otimes \langle i|)$ is the flip operator then we have for all $A, B \in \mathbb{C}^{d\times d}$: $\mathrm{Tr}((A\otimes B)\mathbb{F}) = \mathrm{Tr}(AB)$ and similarly $\mathrm{Tr}(A\otimes B) = \mathrm{Tr}(A)\mathrm{Tr}(B)$. Moreover, another trick we need frequently is to use the partial transpose to make appear the positive semi-definite matrices $M^\dagger M = \sum_{k,l} A_k^\dagger A_l \otimes A_k^\top \bar{A}_l$ and $MM^\dagger = \sum_{k,l} A_k A_l^\dagger \otimes \bar{A}_k A_l^\top$ where $M = \sum_k A_k \otimes \bar{A}_k$ is defined using the Kraus operators $\{A_k\}_k$ of

the channel $\mathcal{N}$. Furthermore we can prove the approximation $\left\|M^\dagger M - \frac{d_{\text{in}}}{d_{\text{out}}} |\Psi\rangle\langle\Psi|\right\|_1 \leq 5\eta^2$ where $|\Psi\rangle = \frac{1}{\sqrt{d_{\text{in}}}} \sum_{i=1}^{d_{\text{in}}} |i\rangle \otimes |i\rangle$ is the maximally entangled state (see Lemma 4.4.5). This is used for $\alpha =$ (142) and (24). An application of the data processing inequality on the previous approximation gives: $\left\|\text{Tr}_2(M^\dagger M) - \frac{1}{d_{\text{out}}}\mathbb{I}\right\|_1 \leq 5\eta^2$ which is used for the permutations $\alpha =$ id and (14) . It turns out that applying such approximation will not give a sufficiently good upper bound of $F((12)(34))$ because it can be written as $F((12)(34)) = \text{Tr}\left(\left(MM^\dagger\right)^{\top_2} \cdot \mathcal{M}(\mathbb{I})^{\otimes 2}\mathbb{F}\right) + \frac{d_{\text{in}}^3}{d_{\text{out}}^2} + 2\frac{d_{\text{in}}}{d_{\text{out}}}\text{Tr}(\mathcal{M}(\mathbb{I})^2)$ which depends instead on the matrix $MM^\dagger$. In this case we proceed by projecting $MM^\dagger$ onto the hyperplane orthogonal to $|\Psi\rangle$. This can be interpreted using representation theory. In fact, the space spanned by $|\Psi\rangle$ and its complementary are irreducible representations that decompose the space $\left(\mathbb{C}^d\right)^{\otimes 2}$ for the action of $U \otimes \bar{U}$ where U is a unitary matrix. On the other hand, changing the Kraus operators $A_k \leftrightarrow UA_k$ or $A_k \leftrightarrow A_k U$ in the expression of the channel $\mathcal{N}$ does not affect the variance of X because Haar measure is invariant under left and right multiplication with a unitary matrix. Now, we give the detailed proof of this theorem.

Proof. Recall that $X = \left\|\mathcal{N}(|\phi\rangle\langle\phi|) - \frac{\mathbb{I}}{d_{\text{out}}}\right\|_2^2$. We can observe that:

$$\text{Var}(X) = \text{Var}\left(\text{Tr}\left(\mathcal{N}(|\phi\rangle\langle\phi|)\right)^2 - \frac{1}{d_{\text{out}}}\right) = \text{Var}\left(\text{Tr}\left(\mathcal{N}(|\phi\rangle\langle\phi|)\right)^2\right). \tag{4.84}$$

Let $d = d_{\text{in}}$. We use the Kraus decomposition for the quantum channel $\mathcal{N}$:

$$\mathcal{N}(\rho) = \sum_k A_k \rho A_k^\dagger. \tag{4.85}$$

By Weingarten calculus [Gu13; CMS12], we can compute the expectation:

$$\mathbb{E}\left(\left(\text{Tr}\left(\mathcal{N}(|\phi\rangle\langle\phi|)\right)^2\right)^2\right) = \mathbb{E}\left(\left(\text{Tr}\sum_{k,l} A_k |\phi\rangle\langle\phi| A_k^\dagger A_l |\phi\rangle\langle\phi| A_l^\dagger\right)^2\right) \tag{4.86}$$

$$= \sum_{i,j}\sum_{k,l,k',l'} \mathbb{E}\left(\text{Tr}A_{l'}^\dagger |j\rangle \langle i| A_k |\phi\rangle\langle\phi| A_k^\dagger A_l |\phi\rangle\langle\phi| A_l^\dagger |i\rangle \langle j| A_{k'} |\phi\rangle\langle\phi| A_{k'}^\dagger A_{l'} |\phi\rangle\langle\phi|\right) \tag{4.87}$$

$$= \sum_{i,j}\sum_{k,l,k',l'} \frac{1}{d(d+1)(d+2)(d+3)} \sum_{\alpha\in\mathfrak{S}_4} \text{Tr}_\alpha(A_{l'}^\dagger |j\rangle \langle i| A_k, A_k^\dagger A_l, A_l^\dagger |i\rangle \langle j| A_{k'}, A_{k'}^\dagger A_{l'}) \tag{4.88}$$

$$= \frac{1}{d(d+1)(d+2)(d+3)} \sum_{\alpha\in\mathfrak{S}_4} \sum_{i,j,k,l,k',l'} \text{Tr}_\alpha(A_{l'}^\dagger |j\rangle \langle i| A_k, A_k^\dagger A_l, A_l^\dagger |i\rangle \langle j| A_{k'}, A_{k'}^\dagger A_{l'}) \tag{4.89}$$

$$= \frac{1}{d(d+1)(d+2)(d+3)} \sum_{\alpha\in\mathfrak{S}_4} F(\alpha) \tag{4.90}$$

where for $\alpha \in \mathfrak{S}_4$ we adopt the notation $F(\alpha) = \sum_{i,j,k,l,k',l'} \text{Tr}_\alpha(A_{l'}^\dagger |j\rangle \langle i| A_k, A_k^\dagger A_l, A_l^\dagger |i\rangle \langle j| A_{k'}, A_{k'}^\dagger A_{l'})$ where $\text{Tr}_\alpha(M_1, \ldots, M_n) = \Pi_j \text{Tr}(\Pi_{i\in C_j} M_i)$ for $\alpha = \Pi_j C_j$ and C_j are cycles. It is thus necessary to control

each of these 24 terms in order to upper bound the variance. Furthermore, we need to be careful so that our upper bounds on $\{F(\alpha)\}_{\alpha\in\mathfrak{S}_4}$ depend on the actual parameters of the testing problem. Recall that the expected value of X can be expressed as follows:

$$\mathbb{E}(X) = \frac{1}{d_{\text{in}}(d_{\text{in}}+1)}\left(\text{Tr}\left(\mathcal{N}(\mathbb{I}) - \frac{d_{\text{in}}}{d_{\text{out}}}\mathbb{I}\right)^2 + d_{\text{in}}^2\left\|\mathcal{J} - \frac{\mathbb{I}}{d_{\text{in}}d_{\text{out}}}\right\|_2^2\right). \tag{4.91}$$

Let us define $\mathcal{M} = \mathcal{N} - \mathcal{D}$, $m = \|\mathcal{M}(\mathbb{I}_{d_{\text{in}}})\|_2 = \left\|\mathcal{N}(\mathbb{I}) - \frac{d_{\text{in}}}{d_{\text{out}}}\mathbb{I}\right\|_2$ and $\eta = d_{\text{in}}\left\|\mathcal{J} - \frac{\mathbb{I}}{d_{\text{in}}d_{\text{out}}}\right\|_2$. We state a useful Lemma relating η, $\mathcal{M}$ and the Kraus operators $\{A_k\}_k$ (defined in (4.85)):

Lemma 4.4.3. *Let $\eta = d_{\text{in}}\|\mathcal{J} - \mathbb{I}/(d_{\text{in}}d_{\text{out}})\|_2$ and $\mathcal{M} = \mathcal{N} - \mathcal{D}$, we have:*

- $\eta^2 = \sum_{k,l}|\text{Tr}(A_k^\dagger A_l)|^2 - \frac{d_{\text{in}}}{d_{\text{out}}}$,
- $\eta^2 = \sum_{x,y}\|\mathcal{M}(|x\rangle\langle y|)\|_2^2$.

Proof. Recall that we use the Kraus representation of the channel $\mathcal{N}(\rho) = \sum_k A_k\rho A_k^\dagger$. We can express η^2:

$$\eta^2 = \left\|d_{\text{in}}\mathcal{J} - \frac{\mathbb{I}}{d_{\text{out}}}\right\|_2^2 = d_{\text{in}}^2\,\text{Tr}(\mathcal{J}^2) - \frac{d_{\text{in}}}{d_{\text{out}}} = \text{Tr}\left(\sum_{i,j,k}|i\rangle\langle j|\otimes A_k|i\rangle\langle j|A_k^\dagger\right)^2 - \frac{d_{\text{in}}}{d_{\text{out}}} \tag{4.92}$$

$$= \sum_{i,j,k,l}\text{Tr}(A_k|i\rangle\langle j|A_k^\dagger A_l|j\rangle\langle i|A_l^\dagger) - \frac{d_{\text{in}}}{d_{\text{out}}} = \sum_{k,l}|\text{Tr}(A_k^\dagger A_l)|^2 - \frac{d_{\text{in}}}{d_{\text{out}}}. \tag{4.93}$$

We move to the second point, we have $\mathcal{J} - \mathbb{I}/(d_{\text{in}}d_{\text{out}}) = \mathcal{J}_\mathcal{N} - \mathcal{J}_\mathcal{D} = \mathcal{J}_\mathcal{M} = \text{id}\otimes\mathcal{M}(|\Psi\rangle\langle\Psi|)$ so

$$\eta^2 = d_{\text{in}}^2\text{Tr}(\text{id}\otimes\mathcal{M}(|\Psi\rangle\langle\Psi|))^2 = \text{Tr}(\text{id}\otimes\mathcal{M}(d_{\text{in}}|\Psi\rangle\langle\Psi|))^2 = \text{Tr}\left(\sum_{x,y}|x\rangle\langle y|\otimes\mathcal{M}(|x\rangle\langle y|)\right)^2 \tag{4.94}$$

$$= \sum_{x,y}\text{Tr}(\mathcal{M}(|x\rangle\langle y|)\mathcal{M}(|y\rangle\langle x|)) = \sum_{x,y}\|\mathcal{M}(|x\rangle\langle y|)\|_2^2. \tag{4.95}$$

□

On the other hand, when dealing with some $F(\alpha)$'s, we will need to have some properties of the matrix $\sum_k A_k\otimes\bar{A}_k$ where $\{A_k\}_k$ are defined in (4.85).

Lemma 4.4.4. *Let $M = \sum_k A_k\otimes\bar{A}_k$. Let $\{\lambda_i\}_i$ be the set of the eigenvalues of $M^\dagger M$ (in a decreasing order) corresponding to the eigenstates $\{|\phi_i\rangle\}_i$. We have:*

- $\sum_i\lambda_i = \frac{d_{\text{in}}}{d_{\text{out}}} + \eta^2$,
- $\lambda_1 \geq \frac{d_{\text{in}}}{d_{\text{out}}}, \sum_{i>1}\lambda_i \leq \eta^2$,
- $\frac{d_{\text{in}}^2}{d_{\text{out}}^2}(1 - |\langle\phi_1|\Psi\rangle|^2) \leq \frac{2m^2\eta^2}{d} + 2\eta^4$,
- $\langle\Psi|MM^\dagger|\Psi\rangle = \frac{d_{\text{in}}}{d_{\text{out}}}$.

Proof. We have $\sum_i \lambda_i = \mathrm{Tr}(M^\dagger M) = \sum_{k,l} \mathrm{Tr}(A_k^\dagger A_l)\mathrm{Tr}(A_k^\top \bar{A}_l) = \sum_{k,l} |\mathrm{Tr}(A_k^\dagger A_l)|^2 = \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} + \eta^2$ by Lemma 4.4.3. Recall the definition of the maximally entangled state $|\Psi\rangle = \frac{1}{\sqrt{d_{\mathrm{in}}}} \sum_i |i\rangle \otimes |i\rangle$, we can compute its image by the matrix $M^\dagger M$:

$$M^\dagger M(\sqrt{d_{\mathrm{in}}}\,|\Psi\rangle) = \sum_{i,k,l} A_k^\dagger A_l\,|i\rangle \otimes A_k^\top \bar{A}_l\,|i\rangle = \sum_{x,y,i,k,l} \langle x|\, A_k^\dagger A_l\,|i\rangle\, \langle y|\, A_k^\top \bar{A}_l\,|i\rangle\,|xy\rangle \tag{4.96}$$

$$= \sum_{x,y,i,k,l} \langle x|\, A_k^\dagger A_l\,|i\rangle\, \langle i|\, A_l^\dagger A_k\,|y\rangle\,|xy\rangle = \sum_{x,y,k} \langle x|\, A_k^\dagger \mathcal{N}(\mathbb{I}) A_k\,|y\rangle\,|xy\rangle \tag{4.97}$$

therefore

$$\langle\Psi|\, M^\dagger M\,|\Psi\rangle = \frac{1}{d_{\mathrm{in}}} \sum_{i,k} \langle i|\, A_k^\dagger \mathcal{N}(\mathbb{I}) A_k\,|i\rangle = \frac{\mathrm{Tr}(\mathcal{N}(\mathbb{I})^2)}{d_{\mathrm{in}}} \geq \frac{d_{\mathrm{in}}^2}{d_{\mathrm{in}} d_{\mathrm{out}}} = \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} \tag{4.98}$$

where we used the Cauchy-Schwarz inequality. This implies that the largest eigenvalue verifies $\lambda_1 \geq \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}$ thus $\sum_{i>1} \lambda_i = \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} + \eta^2 - \lambda_1 \leq \eta^2$.

We move to prove the third point. Recall the notation $\mathcal{M}(\rho) = (\mathcal{N} - \mathcal{D})(\rho) = \sum_k A_k \rho A_k^\dagger - \mathrm{Tr}(\rho)\frac{\mathbb{I}}{d_{\mathrm{out}}}$. We have on the one hand:

$$M^\dagger M(\sqrt{d_{\mathrm{in}}}\,|\Psi\rangle) = \sum_{x,y,k} \langle x|\, A_k^\dagger \mathcal{N}(\mathbb{I}) A_k\,|y\rangle\,|xy\rangle = \sum_{x,y} \mathrm{Tr}(\mathcal{N}(\mathbb{I})\mathcal{N}(|y\rangle\langle x|))\,|xy\rangle \tag{4.99}$$

$$= \sum_{x,y} \mathrm{Tr}(\mathcal{N}(\mathbb{I})\mathcal{M}(|y\rangle\langle x|))\,|xy\rangle + \sum_{x,y} \mathrm{Tr}(\mathcal{N}(\mathbb{I})\mathcal{D}(|y\rangle\langle x|))\,|xy\rangle \tag{4.100}$$

$$= \sum_{x,y} \mathrm{Tr}(\mathcal{M}(\mathbb{I})\mathcal{M}(|y\rangle\langle x|))\,|xy\rangle + \sum_{x} \frac{1}{d_{\mathrm{out}}} \mathrm{Tr}(\mathcal{N}(\mathbb{I})\mathbb{I})\,|xx\rangle \tag{4.101}$$

$$= \sum_{x,y} \mathrm{Tr}(\mathcal{M}(\mathbb{I})\mathcal{M}(|y\rangle\langle x|))\,|xy\rangle + \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} \sqrt{d_{\mathrm{in}}}\,|\Psi\rangle\,. \tag{4.102}$$

On the other hand, using the spectral decomposition of $M^\dagger M$, we can write:

$$\sum_{x,y} \mathrm{Tr}(\mathcal{M}(\mathbb{I})\mathcal{M}(|y\rangle\langle x|))\,|xy\rangle + \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} \sqrt{d_{\mathrm{in}}}\,|\Psi\rangle = M^\dagger M(\sqrt{d_{\mathrm{in}}}\,|\Psi\rangle) = \sum_i \lambda_i \sqrt{d_{\mathrm{in}}} \langle\phi_i|\Psi\rangle\,|\phi_i\rangle\,. \tag{4.103}$$

Therefore

$$\lambda_1 \langle\phi_1|\Psi\rangle\,|\phi_1\rangle - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}\,|\Psi\rangle = \frac{1}{\sqrt{d_{\mathrm{in}}}} \sum_{x,y} \mathrm{Tr}(\mathcal{M}(\mathbb{I})\mathcal{M}(|y\rangle\langle x|))\,|xy\rangle - \sum_{i>1} \lambda_i \langle\phi_i|\Psi\rangle\,|\phi_i\rangle\,. \tag{4.104}$$

Taking the 2-norm squared on both sides, we obtain by the Cauchy-Schwarz inequality:

$$\lambda_1^2|\langle\phi_1|\Psi\rangle|^2 + \frac{d_{\text{in}}^2}{d_{\text{out}}^2} - 2\frac{d_{\text{in}}}{d_{\text{out}}}\lambda_1|\langle\phi_1|\Psi\rangle|^2 \tag{4.105}$$

$$= \left\| \frac{1}{\sqrt{d}} \sum_{x,y} \text{Tr}(\mathcal{M}(\mathbb{I})\mathcal{M}(|y\rangle\langle x|))\,|xy\rangle - \sum_{i>1} \lambda_i \langle\phi_i|\Psi\rangle\,|\phi_i\rangle \right\|_2^2 \tag{4.106}$$

$$\leq \left(\left\| \frac{1}{\sqrt{d}} \sum_{x,y} \text{Tr}(\mathcal{M}(\mathbb{I})\mathcal{M}(|y\rangle\langle x|))\,|xy\rangle \right\|_2 + \left\| \sum_{i>1} \lambda_i \langle\phi_i|\Psi\rangle\,|\phi_i\rangle \right\|_2 \right)^2 \tag{4.107}$$

$$\leq 2\left\| \frac{1}{\sqrt{d}} \sum_{x,y} \text{Tr}(\mathcal{M}(\mathbb{I})\mathcal{M}(|y\rangle\langle x|))\,|xy\rangle \right\|_2^2 + 2\left\| \sum_{i>1} \lambda_i \langle\phi_i|\Psi\rangle\,|\phi_i\rangle \right\|_2^2 \tag{4.108}$$

$$\leq \frac{2}{d} \sum_{x,y} |\text{Tr}(\mathcal{M}(\mathbb{I})\mathcal{M}(|y\rangle\langle x|))|^2 + 2\sum_{i>1} \lambda_i^2 \tag{4.109}$$

$$\leq \frac{2}{d} \sum_{x,y} \|\mathcal{M}(\mathbb{I})\|_2^2 \|\mathcal{M}(|y\rangle\langle x|)\|_2^2 + 2\left(\sum_{i>1} \lambda_i \right)^2 = \frac{2m^2\eta^2}{d_{\text{in}}} + 2\eta^4. \tag{4.110}$$

Observe that the LHS can be lower bounded as follows:

$$\lambda_1^2|\langle\phi_1|\Psi\rangle|^2 + \frac{d_{\text{in}}^2}{d_{\text{out}}^2} - 2\frac{d_{\text{in}}}{d_{\text{out}}}\lambda_1|\langle\phi_1|\Psi\rangle|^2 = \left(\lambda_1 - \frac{d_{\text{in}}}{d_{\text{out}}}\right)^2 |\langle\phi_1|\Psi\rangle|^2 + \frac{d_{\text{in}}^2}{d_{\text{out}}^2} - \frac{d_{\text{in}}^2}{d_{\text{out}}^2}|\langle\phi_1|\Psi\rangle|^2 \tag{4.111}$$

$$\geq \frac{d_{\text{in}}^2}{d_{\text{out}}^2}(1 - |\langle\phi_1|\Psi\rangle|^2). \tag{4.112}$$

Finally, we deduce from the two previous inequalities:

$$\frac{d_{\text{in}}^2}{d_{\text{out}}^2}(1 - |\langle\phi_1|\Psi\rangle|^2) \leq \frac{2m^2\eta^2}{d_{\text{in}}} + 2\eta^4. \tag{4.113}$$

We move to the fourth point. We have:

$$MM^\dagger(\sqrt{d_{\text{in}}}\,|\Psi\rangle) = \sum_{i,k,l} A_k A_l^\dagger\,|i\rangle \otimes \bar{A}_k A_l^\top\,|i\rangle = \sum_{x,y,i,k,l} \langle x|\,A_k A_l^\dagger\,|i\rangle\,\langle y|\,\bar{A}_k A_l^\top\,|i\rangle\,|xy\rangle \tag{4.114}$$

$$= \sum_{x,y,i,k,l} \langle x|\,A_k A_l^\dagger\,|i\rangle\,\langle i|\,A_l A_k^\dagger\,|y\rangle\,|xy\rangle = \sum_{x,y} \langle x|\,\mathcal{N}(\mathbb{I})\,|y\rangle\,|xy\rangle \tag{4.115}$$

$$= \sum_{x,y} \langle x|\,\mathcal{M}(\mathbb{I})\,|y\rangle\,|xy\rangle + \frac{d_{\text{in}}}{d_{\text{out}}}\sqrt{d_{\text{in}}}\,|\Psi\rangle\,. \tag{4.116}$$

Hence:

$$\langle\Psi|\,MM^\dagger\,|\Psi\rangle = \frac{1}{d_{\text{in}}} \sum_x \langle x|\,\mathcal{M}(\mathbb{I})\,|x\rangle + \frac{d_{\text{in}}}{d_{\text{out}}}\langle\Psi|\Psi\rangle = \frac{1}{d}\text{Tr}(\mathcal{M}(\mathbb{I})) + \frac{d_{\text{in}}}{d_{\text{out}}} = \frac{d_{\text{in}}}{d_{\text{out}}}. \tag{4.117}$$

□

The first and fourth points will be used in the proof of Lemma 4.4.9. The first three points imply that the matrix $M^\dagger M$ is close to the maximally entangled state in the 1-norm.

Lemma 4.4.5. *Let* $M = \sum_k A_k \otimes \bar{A}_k$ *and* $|\Psi\rangle = \frac{1}{\sqrt{d_{\text{in}}}} \sum_{i=1}^{d_{\text{in}}} |i\rangle \otimes |i\rangle$. *We have:*

$$\left\| M^\dagger M - \frac{d_{\text{in}}}{d_{\text{out}}} |\Psi\rangle\langle\Psi| \right\|_1 \leq 5\eta^2. \tag{4.118}$$

Proof. Let $|\phi_1\rangle$ the eigenvector of $M^\dagger M$ corresponding to the largest eigenvalue. Using the Fuchs–van de Graaf inequality [FVDG99] and Lemma 4.4.4:

$$\| |\phi_1\rangle\langle\phi_1| - |\Psi\rangle\langle\Psi| \|_1 \leq 2\sqrt{1 - |\langle\phi_1|\Psi\rangle|^2} \leq 2\frac{d_{\text{out}}}{d_{\text{in}}}\sqrt{\frac{2m^2\eta^2}{d_{\text{in}}} + 2\eta^4} \leq 4\frac{d_{\text{out}}}{d_{\text{in}}}\eta^2 \tag{4.119}$$

where we use the Cauchy Schwarz inequality and Lemma 4.4.3:

$$m^2 = \text{Tr}(\mathcal{M}(\mathbb{I})^2) = \sum_{i,j} \text{Tr}(\mathcal{M}(|i\rangle\langle i|)\mathcal{M}(|j\rangle\langle j|)) \tag{4.120}$$

$$\leq \sum_{i,j} \text{Tr}(\mathcal{M}(|i\rangle\langle i|)^2) = d_{\text{in}} \sum_i \text{Tr}(\mathcal{M}(|i\rangle\langle i|)^2) \leq d_{\text{in}}\eta^2. \tag{4.121}$$

By the triangle inequality and Lemma 4.4.4 we deduce:

$$\left\| M^\dagger M - \frac{d_{\text{in}}}{d_{\text{out}}} |\Psi\rangle\langle\Psi| \right\|_1 \leq \left\| M^\dagger M - \frac{d_{\text{in}}}{d_{\text{out}}} |\phi_1\rangle\langle\phi_1| \right\|_1 + \frac{d_{\text{in}}}{d_{\text{out}}} \| |\phi_1\rangle\langle\phi_1| - |\Psi\rangle\langle\Psi| \|_1 \tag{4.122}$$

$$\leq \lambda_1 - \frac{d_{\text{in}}}{d_{\text{out}}} + \sum_{i>1} \lambda_i + 4\eta^2 \leq 5\eta^2. \tag{4.123}$$

□

This Lemma will be used in the proofs of Lemmas 4.4.6 to 4.4.8 and 4.4.10. We move now to upper bound different values of the function F.

Lemma 4.4.6. *We can upper bound* $F((142))$ *and* $F((243))$ *as follows:*

$$F((142)) = F((243)) \leq \frac{d_{\text{in}}/d_{\text{out}} + \eta^2}{d_{\text{out}}} + \frac{\eta^2}{d_{\text{out}}} + 5\eta^4. \tag{4.124}$$

Proof. Recall the notation $\mathcal{M}(\rho) = (\mathcal{N} - \mathcal{D})(\rho) = \sum_k A_k \rho A_k^\dagger - \text{Tr}(\rho)\frac{\mathbb{I}}{d_{\text{out}}}$. We will first write $F((142))$ as a sum of an ideal term reflecting the null hypothesis ($\mathcal{N} = \mathcal{D}$) and an error term reflecting the difference between $\mathcal{N}$ and $\mathcal{D}$. The ideal term is computed exactly and depends on dimensions d_{in} and d_{out}. The error term can also be splited to a simple error depending on η and d_{out} and a more involved term that depends on the Kraus operators $\{A_k\}_k$ and the difference of channels $\mathcal{M}$. To control this latter error, we first write it in a closed form in terms of $\mathcal{M}$, $M = \sum_k A_k \otimes A_k^\dagger$ and the flip operator $\mathbb{F}$. Then we can use the spectral decomposition of the matrix $M^\dagger M - \frac{d_{\text{in}}}{d_{\text{out}}} |\Psi\rangle\langle\Psi|$ in order to decompose this error term into a combination of negligible elements. The final step requires to control the ℓ_1 norm of the coefficients of this combination which is done using Lemma 4.4.5.

We have:

$$F((142)) = \sum_{k,l,k',l'} \mathrm{Tr}(A_k A_{k'}^\dagger A_{l'} A_k^\dagger A_l A_{l'}^\dagger A_{k'} A_l^\dagger) = \sum_{k,l} \mathrm{Tr}(\mathcal{N}(A_l^\dagger A_k)\mathcal{N}(A_k^\dagger A_l)) \tag{4.125}$$

$$= \sum_{k,l} \mathrm{Tr}(\mathcal{M}(A_l^\dagger A_k)\mathcal{N}(A_k^\dagger A_l)) + \mathrm{Tr}((\mathcal{D}(A_l^\dagger A_k)\mathcal{N}(A_k^\dagger A_l))) \tag{4.126}$$

$$= \sum_{k,l} \mathrm{Tr}(\mathcal{M}(A_l^\dagger A_k)\mathcal{M}(A_k^\dagger A_l)) + \mathrm{Tr}((\mathcal{D}(A_l^\dagger A_k)\mathcal{N}(A_k^\dagger A_l))) + \mathrm{Tr}(\mathcal{M}(A_l^\dagger A_k)\mathcal{D}(A_k^\dagger A_l)) \tag{4.127}$$

$$= \sum_{k,l} \mathrm{Tr}(\mathcal{M}(A_l^\dagger A_k)\mathcal{M}(A_k^\dagger A_l)) + \frac{\mathrm{Tr}(A_l^\dagger A_k)}{d_{\mathrm{out}}}\mathrm{Tr}((\mathcal{N}(A_k^\dagger A_l))) + \frac{\mathrm{Tr}(A_k^\dagger A_l)}{d_{\mathrm{out}}}\mathrm{Tr}(\mathcal{M}(A_l^\dagger A_k)) \tag{4.128}$$

$$= \sum_{k,l} \mathrm{Tr}(\mathcal{M}(A_l^\dagger A_k)\mathcal{M}(A_k^\dagger A_l)) + \frac{\mathrm{Tr}(A_l^\dagger A_k)}{d_{\mathrm{out}}}\mathrm{Tr}(A_k^\dagger A_l) \tag{4.129}$$

$$= \sum_{k,l} \mathrm{Tr}(\mathcal{M}(A_l^\dagger A_k)\mathcal{M}(A_k^\dagger A_l)) + \frac{\eta^2 + d_{\mathrm{in}}/d_{\mathrm{out}}}{d_{\mathrm{out}}}. \tag{4.130}$$

It remains to control the sum $\sum_{k,l} \mathrm{Tr}(\mathcal{M}(A_l^\dagger A_k)\mathcal{M}(A_k^\dagger A_l))$. Let $\mathrm{T}_2 : X \otimes Y \mapsto X \otimes Y^\top$ be the partial transpose operator and $M = \sum_l A_l \otimes \bar{A}_l$. Let $\mathbb{F} = \sum_{i,j=1}^d (|i\rangle \otimes |j\rangle)(\langle j| \otimes \langle i|)$ be the flip operator, we have $\mathrm{Tr}(A \otimes B\mathbb{F}) = \sum_{i,j,k,l} A_{i,j}B_{k,l}\mathrm{Tr}(|i\rangle\langle j| \otimes |k\rangle\langle l|\,\mathbb{F}) = \sum_{i,j,k,l} A_{i,j}B_{k,l}\mathrm{Tr}(|i\rangle\langle l| \otimes |k\rangle\langle j|) = \sum_{i,j} A_{i,j}B_{j,i} = \mathrm{Tr}(AB)$ which is known as the replica trick. We have using the replica trick:

$$\sum_{k,l} \mathrm{Tr}(\mathcal{M}(A_l^\dagger A_k)\mathcal{M}(A_k^\dagger A_l)) = \sum_{k,l} \mathrm{Tr}(\mathcal{M}(A_l^\dagger A_k) \otimes \mathcal{M}(A_k^\dagger A_l)\mathbb{F}) \tag{4.131}$$

$$= \mathrm{Tr}\left(\mathcal{M} \otimes \mathcal{M}\left(\sum_{k,l} A_l^\dagger A_k \otimes A_k^\dagger A_l\right)\mathbb{F}\right) \tag{4.132}$$

$$= \mathrm{Tr}\left(\mathcal{M} \otimes \mathcal{M} \circ \mathrm{T}_2\left(\sum_{k,l} A_l^\dagger A_k \otimes A_l^\top \bar{A}_k\right)\mathbb{F}\right) \tag{4.133}$$

$$= \mathrm{Tr}\left(\mathcal{M} \otimes \mathcal{M} \circ \mathrm{T}_2(M^\dagger M)\mathbb{F}\right) \tag{4.134}$$

Let $|\phi\rangle$ be a unit vector, we can write $|\phi\rangle = \sum_{x,y} \phi_{x,y} |x\rangle \otimes |y\rangle$ then we can express:

$$|\mathrm{Tr}\left(\mathcal{M}\otimes\mathcal{M}\circ \mathrm{T}_2(|\phi\rangle\langle\phi|)\mathbb{F}\right)| = \left|\sum_{x,y,z,t} \phi_{x,y}\bar{\phi}_{z,t}\mathrm{Tr}\left(\mathcal{M}\otimes\mathcal{M}\circ \mathrm{T}_2(|x\rangle\langle z|\otimes|y\rangle\langle t|)\mathbb{F}\right)\right| \tag{4.135}$$

$$= \left|\sum_{x,y,z,t} \phi_{x,y}\bar{\phi}_{z,t}\mathrm{Tr}\left(\mathcal{M}\otimes\mathcal{M}(|x\rangle\langle z|\otimes|t\rangle\langle y|)\mathbb{F}\right)\right| \tag{4.136}$$

$$= \left|\sum_{x,y,z,t} \phi_{x,y}\bar{\phi}_{z,t}\mathrm{Tr}\left(\mathcal{M}(|x\rangle\langle z|)\otimes\mathcal{M}(|t\rangle\langle y|)\mathbb{F}\right)\right| \tag{4.137}$$

$$= \left|\sum_{x,y,z,t} \phi_{x,y}\bar{\phi}_{z,t}\mathrm{Tr}\left(\mathcal{M}(|x\rangle\langle z|)\mathcal{M}(|t\rangle\langle y|)\right)\right| \tag{4.138}$$

$$\leq \sqrt{\sum_{x,y,z,t} |\phi_{x,y}\bar{\phi}_{z,t}|^2 \sum_{x,y,z,t} |\mathrm{Tr}\left(\mathcal{M}(|x\rangle\langle z|)\mathcal{M}(|t\rangle\langle y|)\right)|^2} \tag{4.139}$$

$$= \sqrt{\sum_{x,y,z,t} \|\mathcal{M}(|x\rangle\langle z|)\|_2^2 \|\mathcal{M}(|t\rangle\langle y|)\|_2^2} = \eta^2 \tag{4.140}$$

where we use the Cauchy Schwarz inequality and Lemma 4.4.3. On the other hand, we can compute:

$$\mathrm{Tr}(\mathcal{M}\otimes\mathcal{M}\circ \mathrm{T}_2(|\Psi\rangle\langle\Psi|)\mathbb{F}) = \frac{1}{d_{\mathrm{in}}}\sum_{i,j}\mathrm{Tr}(\mathcal{M}\otimes\mathcal{M}\circ \mathrm{T}_2(|i\rangle\otimes|i\rangle\langle jj|)\mathbb{F}) \tag{4.141}$$

$$= \frac{1}{d_{\mathrm{in}}}\sum_{i,j}\mathrm{Tr}(\mathcal{M}(|i\rangle\langle j|)\mathcal{M}(|j\rangle\langle i|)) = \frac{\eta^2}{d_{\mathrm{in}}}. \tag{4.142}$$

Then, we can decompose the Hermitian matrix $M^\dagger M - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}|\Psi\rangle\langle\Psi| = \sum_i \mu_i |\psi_i\rangle\langle\psi_i|$. Hence Lemma 4.4.5 implies:

$$\sum_{k,l}\mathrm{Tr}(\mathcal{M}(A_l^\dagger A_k)\mathcal{M}(A_k^\dagger A_l)) = \mathrm{Tr}\left(\mathcal{M}\otimes\mathcal{M}\circ \mathrm{T}_2(M^\dagger M)\mathbb{F}\right) \tag{4.143}$$

$$= \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}\mathrm{Tr}\left(\mathcal{M}\otimes\mathcal{M}\circ \mathrm{T}_2(|\Psi\rangle\langle\Psi|)\mathbb{F}\right) + \sum_i \mu_i \mathrm{Tr}\left(\mathcal{M}\otimes\mathcal{M}\circ \mathrm{T}_2(|\psi_i\rangle\langle\psi_i|)\mathbb{F}\right) \tag{4.144}$$

$$\leq \frac{\eta^2}{d_{\mathrm{out}}} + \sum_i |\mu_i|\eta^2 = \frac{\eta^2}{d_{\mathrm{out}}} + \mathrm{Tr}\left|M^\dagger M - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}|\Psi\rangle\langle\Psi|\right|\eta^2 \leq \frac{\eta^2}{d_{\mathrm{out}}} + 5\eta^4. \tag{4.145}$$

Finally,

$$F((142)) = \sum_{k,l}\mathrm{Tr}(\mathcal{M}(A_l^\dagger A_k)\mathcal{M}(A_k^\dagger A_l)) + \frac{\frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}+\eta^2}{d_{\mathrm{out}}} \leq \frac{\frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}+\eta^2}{d_{\mathrm{out}}} + \frac{\eta^2}{d_{\mathrm{out}}} + 5\eta^4. \tag{4.146}$$

□

Lemma 4.4.7. *We can upper bound $F(\mathrm{id}), F((132)), F((314))$ and $F((24)(13))$ as follows:*

$$F(\mathrm{id}) = F((132)) = F((314)) = F((24)(13)) \leq \frac{d_{\mathrm{in}}/d_{\mathrm{out}}+\eta^2}{d_{\mathrm{out}}} + \frac{\eta^2}{d_{\mathrm{out}}} + 5\eta^4. \tag{4.147}$$

Proof. For the identity permutation, the function F can be expressed as follows:

$$F(\mathrm{id}) = \sum_{k,l,k',l'} \mathrm{Tr}(A_k A_{l'}^\dagger A_{k'} A_l^\dagger)\mathrm{Tr}(A_{k'}^\dagger A_{l'})\mathrm{Tr}(A_k^\dagger A_l) \tag{4.148}$$

$$= \sum_{k,l,k',l'} \sum_{i,j} \mathrm{Tr}(|j\rangle\langle i| A_{l'}^\dagger A_{k'} A_l^\dagger A_k |i\rangle\langle j| A_k^\dagger A_l)\mathrm{Tr}(A_{k'}^\dagger A_{l'}) \tag{4.149}$$

$$= \sum_{k',l'} \sum_{i,j} \mathrm{Tr}(\mathcal{N}(|j\rangle\langle i| A_{l'}^\dagger A_{k'})\mathcal{N}(|i\rangle\langle j|))\mathrm{Tr}(A_{k'}^\dagger A_{l'}) \tag{4.150}$$

$$= \sum_{k',l'} \sum_{i,j} \mathrm{Tr}(\mathcal{N}(|j\rangle\langle i| A_{l'}^\dagger A_{k'})\mathcal{M}(|i\rangle\langle j|))\mathrm{Tr}(A_{k'}^\dagger A_{l'}) + \sum_{k',l'} \sum_{i,j} \mathrm{Tr}(\mathcal{N}(|j\rangle\langle i| A_{l'}^\dagger A_{k'})\mathcal{D}(|i\rangle\langle j|))\mathrm{Tr}(A_{k'}^\dagger A_{l'}) \tag{4.151}$$

$$= \sum_{k',l'} \sum_{i,j} \mathrm{Tr}(\mathcal{N}(|j\rangle\langle i| A_{l'}^\dagger A_{k'})\mathcal{M}(|i\rangle\langle j|))\mathrm{Tr}(A_{k'}^\dagger A_{l'}) + \sum_{k',l'} \sum_{i} \frac{1}{d_{\mathrm{out}}} \mathrm{Tr}(\mathcal{N}(|i\rangle\langle i| A_{l'}^\dagger A_{k'}))\mathrm{Tr}(A_{k'}^\dagger A_{l'}) \tag{4.152}$$

$$= \sum_{k',l'} \sum_{i,j} \mathrm{Tr}(\mathcal{N}(|j\rangle\langle i| A_{l'}^\dagger A_{k'})\mathcal{M}(|i\rangle\langle j|))\mathrm{Tr}(A_{k'}^\dagger A_{l'}) + \sum_{k',l'} \frac{1}{d_{\mathrm{out}}} \mathrm{Tr}(A_{l'}^\dagger A_{k'})\mathrm{Tr}(A_{k'}^\dagger A_{l'}) \tag{4.153}$$

$$= \sum_{k',l'} \sum_{i,j} \mathrm{Tr}(\mathcal{M}(|j\rangle\langle i| A_{l'}^\dagger A_{k'})\mathcal{M}(|i\rangle\langle j|))\mathrm{Tr}(A_{k'}^\dagger A_{l'}) + \frac{d_{\mathrm{in}}/d_{\mathrm{out}} + \eta^2}{d_{\mathrm{out}}} \tag{4.154}$$

$$= \sum_{i,j} \mathrm{Tr}(\mathcal{M}(|j\rangle\langle i| N)\mathcal{M}(|i\rangle\langle j|)) + \frac{d_{\mathrm{in}}/d_{\mathrm{out}} + \eta^2}{d_{\mathrm{out}}} \tag{4.155}$$

where $N = \sum_{k,l} \mathrm{Tr}(A_k^\dagger A_l) A_l^\dagger A_k$. Let us introduce $\tilde{N} = N - \frac{\mathbb{I}}{d_{\mathrm{out}}} = \sum_i \mu_x |\psi_x\rangle\langle\psi_x|$ (this is possible because N is Hermitian) so that we can write using Lemma 4.4.3:

$$\sum_{i,j} \mathrm{Tr}(\mathcal{M}(|j\rangle\langle i| N)\mathcal{M}(|i\rangle\langle j|)) \tag{4.156}$$

$$= \sum_{i,j} \mathrm{Tr}(\mathcal{M}(|j\rangle\langle i| \tilde{N})\mathcal{M}(|i\rangle\langle j|)) + \frac{1}{d_{\mathrm{out}}} \sum_{i,j} \mathrm{Tr}(\mathcal{M}(|j\rangle\langle i|)\mathcal{M}(|i\rangle\langle j|)) \tag{4.157}$$

$$= \sum_{i,j,x} \mu_x \mathrm{Tr}(\mathcal{M}(|j\rangle\langle i| |\psi_x\rangle\langle\psi_x|)\mathcal{M}(|i\rangle\langle j|)) + \frac{\eta^2}{d_{\mathrm{out}}} \tag{4.158}$$

$$= \sum_{i,j,x} \mu_x \langle i|\psi_x\rangle \mathrm{Tr}(\mathcal{M}(|j\rangle\langle\psi_x|)\mathcal{M}(|i\rangle\langle j|)) + \frac{\eta^2}{d_{\mathrm{out}}} \tag{4.159}$$

$$= \sum_{i,j,x,k} \mu_x \mathrm{Tr}(\langle i|\psi_x\rangle \mathcal{M}(|j\rangle\langle k|)\langle\psi_x|k\rangle \mathcal{M}(|i\rangle\langle j|)) + \frac{\eta^2}{d_{\mathrm{out}}} \tag{4.160}$$

$$\leq \frac{1}{2} \sum_{i,j,x,k} |\mu_x| (\|\langle i|\psi_x\rangle \mathcal{M}(|j\rangle\langle k|)\|_2^2 + \|\langle\psi_x|k\rangle \mathcal{M}(|i\rangle\langle j|)\|_2^2) + \frac{\eta^2}{d_{\mathrm{out}}} \tag{4.161}$$

$$= \mathrm{Tr}|\tilde{N}|\eta^2 + \frac{\eta^2}{d_{\mathrm{out}}}. \tag{4.162}$$

We can see the matrix $N = \sum_{k,l} \mathrm{Tr}(A_k^\dagger A_l) A_l^\dagger A_k$ as a partial trace of $M^\dagger M$:

$$\mathrm{Tr}_2(M^\dagger M) = \mathrm{Tr}_2\left(\sum_{k,l} A_l^\dagger A_k \otimes A_l^\top \bar{A}_k\right) = \sum_{k,l} \mathrm{Tr}(A_l^\top \bar{A}_k) A_l^\dagger A_k \tag{4.163}$$

$$= \sum_{k,l} \mathrm{Tr}(A_l A_k^*) A_l^\dagger A_k = N. \tag{4.164}$$

Moreover $\frac{\mathbb{I}}{d_{\mathrm{in}}} = \mathrm{Tr}_2(|\Psi\rangle\langle\Psi|)$ so by the data processing inequality (the partial trace is a valid quantum channel) and Lemma 4.4.5, we deduce:

$$\mathrm{Tr}|\tilde{N}| = \left\| \mathrm{Tr}_2(M^\dagger M) - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} \mathrm{Tr}_2(|\Psi\rangle\langle\Psi|) \right\|_1 \leq \left\| M^\dagger M - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} |\Psi\rangle\langle\Psi| \right\|_1 \leq 5\eta^2. \tag{4.165}$$

Finally,

$$F(\mathrm{id}) = \sum_{i,j} \mathrm{Tr}(\mathcal{M}(|j\rangle\langle i|\, N) \mathcal{M}(|i\rangle\langle j|)) + \frac{d_{\mathrm{in}}/d_{\mathrm{out}} + \eta^2}{d_{\mathrm{out}}} \leq \frac{d_{\mathrm{in}}/d_{\mathrm{out}} + \eta^2}{d_{\mathrm{out}}} + \frac{\eta^2}{d_{\mathrm{out}}} + 5\eta^4. \tag{4.166}$$

□

Lemma 4.4.8. *We can upper bound $F((24))$ and $F((1432))$ as follows:*

$$F((24)) = F((1432)) \leq \frac{d_{\mathrm{in}}^2}{d_{\mathrm{out}}^2} + \frac{2m^2}{d_{\mathrm{out}}} + 25\eta^4. \tag{4.167}$$

Proof. Recall that $\langle\Psi| M^\dagger M |\Psi\rangle = \frac{\mathrm{Tr}(\mathcal{N}(\mathbb{I})^2)}{d_{\mathrm{in}}} = \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} + \frac{m^2}{d_{\mathrm{in}}}$. We use the fact that $\mathrm{Tr}(X)\mathrm{Tr}(Y) = \mathrm{Tr}(X \otimes Y)$:

$$F((24)) = \sum_{k,l,k',l'} \mathrm{Tr}(A_k^\dagger A_l A_{k'}^\dagger A_{l'}) \mathrm{Tr}(A_k A_{l'}^\dagger A_{k'} A_l^\dagger) = \sum_{k,l,k',l'} \mathrm{Tr}(A_k^\dagger A_l A_{k'}^\dagger A_{l'} \otimes A_l^\dagger A_k A_{l'}^\dagger A_{k'}) \tag{4.168}$$

$$= \mathrm{Tr}\left(\sum_{k,l} A_k^\dagger A_l \otimes A_l^\dagger A_k\right)\left(\sum_{k,l} A_k^\dagger A_l \otimes A_l^\dagger A_k\right) \tag{4.169}$$

$$= \mathrm{Tr}\left(\sum_{k,l} A_k^\dagger A_l \otimes A_k^\top \bar{A}_l\right)^{\top_2}\left(\sum_{k,l} A_k^\dagger A_l \otimes A_k^\top \bar{A}_l\right)^{\top_2} \tag{4.170}$$

$$= \mathrm{Tr}(M^\dagger M)^{\top_2}(M^\dagger M)^{\top_2} = \mathrm{Tr}(M^\dagger M)(M^\dagger M) \tag{4.171}$$

$$= \mathrm{Tr}\left(M^\dagger M - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} |\Psi\rangle\langle\Psi|\right) M^\dagger M + \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} \mathrm{Tr}(|\Psi\rangle\langle\Psi|\, M^\dagger M) \tag{4.172}$$

$$= \mathrm{Tr}\left(M^\dagger M - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} |\Psi\rangle\langle\Psi|\right)\left(M^\dagger M - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} |\Psi\rangle\langle\Psi|\right) \tag{4.173}$$

$$+ \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} \langle\Psi|\left(M^\dagger M - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} |\Psi\rangle\langle\Psi|\right)|\Psi\rangle + \frac{d_{\mathrm{in}}^2}{d_{\mathrm{out}}^2} + \frac{m^2}{d_{\mathrm{out}}} \tag{4.174}$$

$$\leq \left(\mathrm{Tr}\left|M^\dagger M - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} |\Psi\rangle\langle\Psi|\right|\right)^2 + \frac{d_{\mathrm{in}}^2}{d_{\mathrm{out}}^2} + \frac{2m^2}{d_{\mathrm{out}}} \leq \frac{d_{\mathrm{in}}^2}{d_{\mathrm{out}}^2} + \frac{2m^2}{d_{\mathrm{out}}} + 25\eta^4 \tag{4.175}$$

where we have used the fact that $\mathrm{Tr}\left(A^{\top_2} B^{\top_2}\right) = \mathrm{Tr}(AB)$ and Lemma 4.4.5. □

Lemma 4.4.9. *We can upper bound $F((12)(34)), F((14)(23)), F((234))$ and $F((124))$ as follows:*

$$F((12)(34)) = F((14)(23)) = F((234)) = F((124)) \leq \frac{d_{\text{in}}^3}{d_{\text{out}}^2} + 2\frac{d_{\text{in}}}{d_{\text{out}}}m^2 + \eta^2 m^2. \quad (4.176)$$

Proof. Using the expression of $F(\alpha)$ for the permutation $\alpha = (12)(34)$ and the fact that $\sum_k A_k^\dagger A_k = \mathbb{I}$, we have:

$$F((12)(34)) = \sum_{k,l,k',l'} \text{Tr}(A_k A_k^\dagger A_l A_{l'}^\dagger A_{k'} A_{k'}^\dagger A_{l'} A_l^\dagger) = \sum_{k,l} \text{Tr}(A_l^\dagger \mathcal{N}(\mathbb{I}) A_l A_k^\dagger \mathcal{N}(\mathbb{I}) A_k) \quad (4.177)$$

$$= \sum_{k,l} \text{Tr}(A_l^\dagger \mathcal{M}(\mathbb{I}) A_l A_k^\dagger \mathcal{N}(\mathbb{I}) A_k) + \frac{d_{\text{in}}}{d_{\text{out}}} \sum_{k,l} \text{Tr}(A_l^\dagger A_l A_k^\dagger \mathcal{N}(\mathbb{I}) A_k) \quad (4.178)$$

$$= \sum_{k,l} \text{Tr}(A_l^\dagger \mathcal{M}(\mathbb{I}) A_l A_k^\dagger \mathcal{M}(\mathbb{I}) A_k) + \frac{d_{\text{in}}}{d_{\text{out}}} \sum_{k,l} \text{Tr}(A_l^\dagger \mathcal{M}(\mathbb{I}) A_l A_k^\dagger A_k) + \frac{d_{\text{in}}}{d_{\text{out}}} \sum_k \text{Tr}(A_k^\dagger \mathcal{N}(\mathbb{I}) A_k) \quad (4.179)$$

$$= \sum_{k,l} \text{Tr}(A_l^\dagger \mathcal{M}(\mathbb{I}) A_l A_k^\dagger \mathcal{M}(\mathbb{I}) A_k) + \frac{d_{\text{in}}}{d_{\text{out}}} \text{Tr}(\mathcal{M}(\mathbb{I})\mathcal{N}(\mathbb{I})) + \frac{d_{\text{in}}}{d_{\text{out}}} \text{Tr}(\mathcal{N}(\mathbb{I})^2) \quad (4.180)$$

$$= \sum_{k,l} \text{Tr}(A_l^\dagger \mathcal{M}(\mathbb{I}) A_l A_k^\dagger \mathcal{M}(\mathbb{I}) A_k) + \frac{d_{\text{in}}^3}{d_{\text{out}}^2} + 2\frac{d_{\text{in}}}{d_{\text{out}}} \text{Tr}(\mathcal{M}(\mathbb{I})^2) \quad (4.181)$$

Then, if we focus on the first term,

we can use the replica trick again to obtain:

$$\sum_{k,l} \text{Tr}(A_l^\dagger \mathcal{M}(\mathbb{I}) A_l A_k^\dagger \mathcal{M}(\mathbb{I}) A_k) = \sum_{k,l} \text{Tr}(A_k A_l^\dagger \mathcal{M}(\mathbb{I}) \otimes A_l A_k^\dagger \mathcal{M}(\mathbb{I})\mathbb{F}) \quad (4.182)$$

$$= \text{Tr}\left(\sum_{k,l} A_k A_l^\dagger \otimes A_l A_k^\dagger \cdot \mathcal{M}(\mathbb{I})^{\otimes 2}\mathbb{F}\right) = \text{Tr}\left(\text{T}_2\left(\sum_{k,l} A_k A_l^\dagger \otimes \bar{A}_k A_l^\top\right) \cdot \mathcal{M}(\mathbb{I})^{\otimes 2}\mathbb{F}\right) \quad (4.183)$$

$$= \text{Tr}\left(\left(MM^\dagger\right)^{\top_2} \cdot \mathcal{M}(\mathbb{I})^{\otimes 2}\mathbb{F}\right) \quad (4.184)$$

$$= \text{Tr}\left(\left((\mathbb{I} - |\Psi\rangle\langle\Psi|)MM^\dagger(\mathbb{I} - |\Psi\rangle\langle\Psi|)\right)^{\top_2} \cdot \mathcal{M}(\mathbb{I})^{\otimes 2}\mathbb{F}\right) + \text{Tr}\left(\left(MM^\dagger |\Psi\rangle\langle\Psi|\right)^{\top_2} \cdot \mathcal{M}(\mathbb{I})^{\otimes 2}\mathbb{F}\right) \quad (4.185)$$

$$+ \text{Tr}\left(\left(|\Psi\rangle\langle\Psi| MM^\dagger\right)^{\top_2} \cdot \mathcal{M}(\mathbb{I})^{\otimes 2}\mathbb{F}\right) - \text{Tr}\left(\left(|\Psi\rangle\langle\Psi| MM^\dagger |\Psi\rangle\langle\Psi|\right)^{\top_2} \cdot \mathcal{M}(\mathbb{I})^{\otimes 2}\mathbb{F}\right). \quad (4.186)$$

We can simplify the latter terms. First we have $|\Psi\rangle\langle\Psi|^{\top_2} = \mathbb{F}$ and $\mathbb{F}^2 = \mathbb{I}$ so

$$\text{Tr}\left(\left(|\Psi\rangle\langle\Psi| MM^\dagger |\Psi\rangle\langle\Psi|\right)^{\top_2} \cdot \mathcal{M}(\mathbb{I})^{\otimes 2}\mathbb{F}\right) = \langle\Psi| MM^\dagger |\Psi\rangle \, \text{Tr}\left(\left(|\Psi\rangle\langle\Psi|\right)^{\top_2} \cdot \mathcal{M}(\mathbb{I})^{\otimes 2}\mathbb{F}\right) \quad (4.187)$$

$$= \langle\Psi| MM^\dagger |\Psi\rangle \, \text{Tr}\left(\mathbb{F}\mathcal{M}(\mathbb{I})^{\otimes 2}\mathbb{F}\right) = \langle\Psi| MM^\dagger |\Psi\rangle \, \text{Tr}\left(\mathcal{M}(\mathbb{I})^{\otimes 2}\right) = \langle\Psi| MM^\dagger |\Psi\rangle \, \text{Tr}\left(\mathcal{M}(\mathbb{I})\right)^2 = 0. \quad (4.188)$$

Next by Lemma 4.4.4 we have $MM^\dagger |\Psi\rangle = \frac{1}{\sqrt{d_{\text{in}}}} \sum_{x,y} \langle x| \mathcal{N}(\mathbb{I}) |y\rangle |xy\rangle$ so

$$\operatorname{Tr}\left(\left(MM^\dagger |\Psi\rangle\langle\Psi|\right)^{\top_2} \cdot \mathcal{M}(\mathbb{I})^{\otimes 2}\mathbb{F}\right) = \frac{1}{d_{\text{in}}} \sum_{x,y,z} \langle x| \mathcal{N}(\mathbb{I}) |y\rangle \operatorname{Tr}\left(\left(|xy\rangle \langle zz|\right)^{\top_2} \cdot \mathcal{M}(\mathbb{I})^{\otimes 2}\mathbb{F}\right) \tag{4.189}$$

$$= \frac{1}{d_{\text{in}}} \sum_{x,y,z} \langle x| \mathcal{N}(\mathbb{I}) |y\rangle \operatorname{Tr}\left(|x\rangle \langle z| \otimes |z\rangle \langle y| \cdot \mathcal{M}(\mathbb{I})^{\otimes 2}\mathbb{F}\right) \tag{4.190}$$

$$= \frac{1}{d_{\text{in}}} \sum_{x,y,z} \langle x| \mathcal{N}(\mathbb{I}) |y\rangle \operatorname{Tr}\left(|x\rangle \langle z| \mathcal{M}(\mathbb{I}) |z\rangle \langle y| \mathcal{M}(\mathbb{I})\right) \tag{4.191}$$

$$= \frac{1}{d_{\text{in}}} \sum_{x,y} \langle x| \mathcal{N}(\mathbb{I}) |y\rangle \operatorname{Tr}\left(|x\rangle \operatorname{Tr}(\mathcal{M}(\mathbb{I})) \langle y| \mathcal{M}(\mathbb{I})\right) = 0. \tag{4.192}$$

Similarly we prove:

$$\operatorname{Tr}\left(\left(|\Psi\rangle\langle\Psi| MM^\dagger\right)^{\top_2} \cdot \mathcal{M}(\mathbb{I})^{\otimes 2}\mathbb{F}\right) = \frac{1}{d_{\text{in}}} \sum_{x,y,z} \langle x| \mathcal{N}(\mathbb{I}) |y\rangle \operatorname{Tr}\left(\left(|zz\rangle \langle xy|\right)^{\top_2} \cdot \mathcal{M}(\mathbb{I})^{\otimes 2}\mathbb{F}\right) \tag{4.193}$$

$$= \frac{1}{d_{\text{in}}} \sum_{x,y,z} \langle x| \mathcal{N}(\mathbb{I}) |y\rangle \operatorname{Tr}\left(|z\rangle \langle x| \otimes |y\rangle \langle z| \cdot \mathcal{M}(\mathbb{I})^{\otimes 2}\mathbb{F}\right) \tag{4.194}$$

$$= \frac{1}{d_{\text{in}}} \sum_{x,y,z} \langle x| \mathcal{N}(\mathbb{I}) |y\rangle \operatorname{Tr}\left(|z\rangle \langle x| \mathcal{M}(\mathbb{I}) |y\rangle \langle z| \mathcal{M}(\mathbb{I})\right) \tag{4.195}$$

$$= \frac{1}{d_{\text{in}}} \sum_{x,y} \langle x| \mathcal{N}(\mathbb{I}) |y\rangle \langle y| \mathcal{M}(\mathbb{I}) |x\rangle \operatorname{Tr}(\mathcal{M}(\mathbb{I})) = 0. \tag{4.196}$$

Now the matrix $(\mathbb{I} - |\Psi\rangle\langle\Psi|)MM^\dagger(\mathbb{I} - |\Psi\rangle\langle\Psi|)$ is Hermitian and positive semi-definite so can be written as $(\mathbb{I} - |\Psi\rangle\langle\Psi|)MM^\dagger(\mathbb{I} - |\Psi\rangle\langle\Psi|) = \sum_i \lambda_i |\phi_i\rangle\langle\phi_i|$, and for each i, we can

write the Schmidt's decomposition of $|\phi\rangle = \sum_{x,y} \phi_{x,y} |xy\rangle$. Therefore

$$\sum_{k,l} \mathrm{Tr}(A_l^\dagger \mathcal{M}(\mathbb{I}) A_l A_k^\dagger \mathcal{M}(\mathbb{I}) A_k) = \mathrm{Tr}\left(\left((\mathbb{I} - |\Psi\rangle\langle\Psi|) M M^\dagger (\mathbb{I} - |\Psi\rangle\langle\Psi|) \right)^{\top_2} \cdot \mathcal{M}(\mathbb{I})^{\otimes 2} \mathbb{F} \right) \tag{4.197}$$

$$= \sum_i \lambda_i \mathrm{Tr}\left(\mathrm{T}_2 \left(|\phi_i\rangle\langle\phi_i| \right) \cdot \mathcal{M}(\mathbb{I})^{\otimes 2} \mathbb{F} \right) = \sum_i \sum_{x,y,z,t} \lambda_i \phi_{x,y} \bar{\phi}_{z,t} \mathrm{Tr}\left(\mathrm{T}_2 \left(|xy\rangle \langle zt| \right) \cdot \mathcal{M}(\mathbb{I})^{\otimes 2} \mathbb{F} \right) \tag{4.198}$$

$$= \sum_i \sum_{x,y,z,t} \lambda_i \phi_{x,y} \bar{\phi}_{z,t} \mathrm{Tr}\left(|x\rangle \langle z| \otimes |t\rangle \langle y| \cdot \mathcal{M}(\mathbb{I})^{\otimes 2} \mathbb{F} \right) \tag{4.199}$$

$$= \sum_i \sum_{x,y,z,t} \lambda_i \phi_{x,y} \bar{\phi}_{z,t} \mathrm{Tr}\left(|x\rangle \langle z| \mathcal{M}(\mathbb{I}) \otimes |t\rangle \langle y| \mathcal{M}(\mathbb{I}) \mathbb{F} \right) \tag{4.200}$$

$$= \sum_i \sum_{x,y,z,t} \lambda_i \phi_{x,y} \bar{\phi}_{z,t} \mathrm{Tr}\left(|x\rangle \langle z| \mathcal{M}(\mathbb{I}) |t\rangle \langle y| \mathcal{M}(\mathbb{I}) \right) \tag{4.201}$$

$$= \sum_i \sum_{x,y,z,t} \lambda_i \phi_{x,y} \bar{\phi}_{z,t} \langle z| \mathcal{M}(\mathbb{I}) |t\rangle \langle y| \mathcal{M}(\mathbb{I}) |x\rangle \tag{4.202}$$

$$\leq \sum_i \sum_{x,y,z,t} \lambda_i |\phi_{x,y}|^2 |\langle z| \mathcal{M}(\mathbb{I}) |t\rangle|^2 + \sum_i \sum_{x,y,z,t} \lambda_i |\phi_{z,t}|^2 |\langle y| \mathcal{M}(\mathbb{I}) |x\rangle|^2 \tag{4.203}$$

$$\leq \sum_i \sum_{x,y,z,t} \lambda_i |\phi_{x,y}|^2 |\langle z| \mathcal{M}(\mathbb{I}) |t\rangle|^2 + \sum_i \sum_{x,y,z,t} \lambda_i |\phi_{z,t}|^2 |\langle y| \mathcal{M}(\mathbb{I}) |x\rangle|^2 \tag{4.204}$$

$$= \mathrm{Tr}\left| (\mathbb{I} - |\Psi\rangle\langle\Psi|) M M^\dagger (\mathbb{I} - |\Psi\rangle\langle\Psi|) \right| \mathrm{Tr}(\mathcal{M}(\mathbb{I})^2) \tag{4.205}$$

By Lemma 4.4.4 we have $\langle\Psi| M M^\dagger |\Psi\rangle = \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}$ and $\mathrm{Tr}(M M^\dagger) = \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} + \eta^2$ so we have:

$$\mathrm{Tr}\left| (\mathbb{I} - |\Psi\rangle\langle\Psi|) M M^\dagger (\mathbb{I} - |\Psi\rangle\langle\Psi|) \right| = \mathrm{Tr}(\mathbb{I} - |\Psi\rangle\langle\Psi|) M M^\dagger (\mathbb{I} - |\Psi\rangle\langle\Psi|) = \eta^2. \tag{4.206}$$

Finally

$$F((12)(34)) \leq \frac{d_{\mathrm{in}}^3}{d_{\mathrm{out}}^2} + 2\frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} m^2 + \eta^2 m^2. \tag{4.207}$$

This concludes the proof for $F((12)(34))$. □

Lemma 4.4.10. *We can upper bound $F((14)), F((12)), F((23))$ and $F((34))$ as follows:*

$$F((14)) = F((12)) = F((23)) = F((34)) \leq \frac{d_{\mathrm{in}}^2}{d_{\mathrm{out}}^2} + \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} \eta^2 + \frac{m^2}{d_{\mathrm{out}}} + 5m\eta^3. \tag{4.208}$$

Proof. Recall the notation $N = \sum_{k,l} \mathrm{Tr}(A_k A_l^\dagger) A_k^\dagger A_l$. We can write the spectral decomposition of the Hermitian matrix:

$$N - \frac{\mathbb{I}}{d_{\mathrm{out}}} = \mathrm{Tr}_2 \left(\sum_{k,l} A_l^\dagger A_k \otimes A_l^\top \bar{A}_k - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} |\Psi\rangle\langle\Psi| \right) = \sum_i \lambda_i |\phi_i\rangle\langle\phi_i| . \tag{4.209}$$

Then using the triangle inequality and the fact that $\sum_k A_k^\dagger A_k = \mathbb{I}$:

$$F(14) = \sum_{k,l,k',l'} \mathrm{Tr}(A_k^\dagger A_l)\mathrm{Tr}(A_l^\dagger A_k A_{k'}^\dagger A_{l'} A_{l'}^\dagger A_{k'}) = \sum_{k'} \mathrm{Tr}(N A_{k'}^\dagger \mathcal{N}(\mathbb{I}) A_{k'}) \tag{4.210}$$

$$= \sum_{k'} \mathrm{Tr}(N A_{k'}^\dagger \mathcal{M}(\mathbb{I}) A_{k'}) + \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}\mathrm{Tr}(N) = \mathrm{Tr}(\mathcal{N}(N)\mathcal{M}(\mathbb{I})) + \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}\mathrm{Tr}(N) \tag{4.211}$$

$$= \mathrm{Tr}(\mathcal{M}(N)\mathcal{M}(\mathbb{I})) + \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}\mathrm{Tr}(N) \tag{4.212}$$

$$= \mathrm{Tr}\left(\mathcal{M}\left(N - \frac{\mathbb{I}}{d_{\mathrm{out}}}\right)\mathcal{M}(\mathbb{I})\right) + \frac{\mathrm{Tr}(\mathcal{M}(\mathbb{I})^2)}{d_{\mathrm{out}}} + \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}\mathrm{Tr}(N) \tag{4.213}$$

$$= \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}\left(\frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} + \eta^2\right) + \frac{\mathrm{Tr}(\mathcal{M}(\mathbb{I})^2)}{d_{\mathrm{out}}} + \sum_i \lambda_i \mathrm{Tr}(\mathcal{M}(|\phi_i\rangle\langle\phi_i|)\mathcal{M}(\mathbb{I})) \tag{4.214}$$

$$\leq \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}\left(\frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} + \eta^2\right) + \frac{\mathrm{Tr}(\mathcal{M}(\mathbb{I})^2)}{d_{\mathrm{out}}} + \sum_i |\lambda_i| \|\mathcal{M}(|\phi_i\rangle\langle\phi_i|)\|_2 \|\mathcal{M}(\mathbb{I})\|_2 \tag{4.215}$$

$$\leq \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}\left(\frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} + \eta^2\right) + \frac{m^2}{d_{\mathrm{out}}} + m\eta \sum_i |\lambda_i| \tag{4.216}$$

$$= \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}\left(\frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} + \eta^2\right) + \frac{m^2}{d_{\mathrm{out}}} + m\eta \cdot \mathrm{Tr}\left|N - \frac{\mathbb{I}}{d_{\mathrm{out}}}\right| \tag{4.217}$$

because for all unit vector $|\phi\rangle = \sum_i \phi_i |i\rangle$ we have using the Cauchy Schwarz inequality, the AM-GM inequality, and Lemma 4.4.3:

$$\|\mathcal{M}(|\phi\rangle\langle\phi|)\|_2^2 = \sum_{i,j,k,l} \phi_i \bar{\phi}_j \phi_k \bar{\phi}_l \mathrm{Tr}\left(\mathcal{M}(|i\rangle\langle j|)\mathcal{M}(|k\rangle\langle l|)\right) \tag{4.218}$$

$$\leq \sum_{i,j,k,l} |\phi_i \bar{\phi}_j \phi_k \bar{\phi}_l| \|\mathcal{M}(|i\rangle\langle j|)\|_2 \|\mathcal{M}(|k\rangle\langle l|)\|_2 \tag{4.219}$$

$$\leq \frac{1}{2}\sum_{i,j,k,l} |\phi_i|^2 |\phi_j|^2 \|\mathcal{M}(|k\rangle\langle l|)\|_2^2 + \frac{1}{2}\sum_{i,j,k,l} |\phi_k|^2 |\phi_l|^2 \|\mathcal{M}(|i\rangle\langle j|)\|_2^2 = \eta^2. \tag{4.220}$$

Moreover, using the data processing inequality and Lemma 4.4.5 :

$$\mathrm{Tr}\left|N - \frac{\mathbb{I}}{d_{\mathrm{out}}}\right| = \mathrm{Tr}\left|\mathrm{Tr}_2\left(\sum_{k,l} A_l^\dagger A_k \otimes A_l^\top \bar{A}_k - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}|\Psi\rangle\langle\Psi|\right)\right| \tag{4.221}$$

$$\leq \mathrm{Tr}\left|\sum_{k,l} A_l^\dagger A_k \otimes A_l^\top \bar{A}_k - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}|\Psi\rangle\langle\Psi|\right| = \mathrm{Tr}\left|M^\dagger M - \frac{d_{\mathrm{in}}}{d_{\mathrm{out}}}|\Psi\rangle\langle\Psi|\right| \leq 5\eta^2. \tag{4.222}$$

This concludes the proof. □

For the remaining permutations, we can obtain a closed form for the function F. For

the transposition (13), the image of the function F can be expressed as follows:

$$F((13)) = \sum_{i,j,k,l,k',l'} \mathrm{Tr}(A_{l'}^\dagger |j\rangle \langle i| A_k A_l^\dagger |i\rangle \langle j| A_{k'}) \mathrm{Tr}(A_k^\dagger A_l) \mathrm{Tr}(A_{k'}^\dagger A_{l'}) \tag{4.223}$$

$$= \sum_{k,l,k',l'} \mathrm{Tr}(A_{k'} A_{l'}^\dagger) \mathrm{Tr}(A_k A_l^\dagger) \mathrm{Tr}(A_k^\dagger A_l) \mathrm{Tr}(A_{k'}^\dagger A_{l'}) \tag{4.224}$$

$$= \left(\sum_{k,l} |\mathrm{Tr}(A_k A_l^\dagger)|^2 \right)^2 = \left(\frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} + \eta^2 \right)^2 . \tag{4.225}$$

Then, we remark that the permutations (312) and (134) have the same image:

$$F((312)) = F((134)) = \sum_{i,j,k,l,k',l'} \mathrm{Tr}(A_l^\dagger |i\rangle \langle j| A_{k'} A_{l'}^\dagger |j\rangle \langle i| A_k A_k^\dagger A_l) \mathrm{Tr}(A_{k'}^\dagger A_{l'}) \tag{4.226}$$

$$= \sum_{k,l,k',l'} \mathrm{Tr}(A_{k'} A_{l'}^\dagger) \mathrm{Tr}(A_k A_k^\dagger A_l A_l^\dagger) \mathrm{Tr}(A_{k'}^\dagger A_{l'}) \tag{4.227}$$

$$= \sum_{k,l} \mathrm{Tr}(A_k A_k^\dagger A_l A_l^\dagger) \sum_{k,l} |\mathrm{Tr}(A_k A_l^\dagger)|^2 \tag{4.228}$$

$$= \mathrm{Tr}(\mathcal{N}(\mathbb{I})^2) \left(\frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} + \eta^2 \right) = \left(\frac{d_{\mathrm{in}}^2}{d_{\mathrm{out}}} + m^2 \right) \left(\frac{d_{\mathrm{in}}}{d_{\mathrm{out}}} + \eta^2 \right) . \tag{4.229}$$

Next, the image of the cycle (1234) has also a closed expression:

$$F((1234)) = \sum_{i,j,k,l,k',l'} \mathrm{Tr}(A_{l'}^\dagger |j\rangle \langle i| A_k A_k^\dagger A_l A_l^\dagger |i\rangle \langle j| A_{k'} A_{k'}^\dagger A_{l'}) \tag{4.230}$$

$$= \sum_{k,l,k',l'} \mathrm{Tr}(A_k A_k^\dagger A_l A_l^\dagger) \mathrm{Tr}(A_{k'} A_{k'}^\dagger A_{l'} A_{l'}^\dagger) \tag{4.231}$$

$$= (\mathrm{Tr}(\mathcal{N}(\mathbb{I})^2))^2 = \left(\frac{d_{\mathrm{in}}^2}{d_{\mathrm{out}}} + m^2 \right)^2 . \tag{4.232}$$

To sum up, we have proved so far that:

Lemma 4.4.11. *Let* $m = \|\mathcal{M}(\mathbb{I})\|_2 = \|\mathcal{N}(\mathbb{I}) - \mathcal{D}(\mathbb{I})\|_2$ *and* $\eta = d_{\mathrm{in}} \left\| \mathcal{J} - \frac{\mathbb{I}}{d_{\mathrm{in}} d_{\mathrm{out}}} \right\|_2$. *We have:*

Therefore we have the following upper bound on the second moment using the inequality $m^2 \leq d\eta^2$:

$$\mathbb{E}\left(\left(\mathrm{Tr}(\mathcal{N}(|\phi\rangle\langle\phi|)^2) \right)^2 \right) = \frac{1}{d_{\mathrm{in}}(d_{\mathrm{in}}+1)(d_{\mathrm{in}}+2)(d_{\mathrm{in}}+3)} \sum_{\alpha \in \mathfrak{S}_4} F(\alpha) \tag{4.233}$$

$$\leq \frac{d_{\mathrm{out}}^4 + 6d_{\mathrm{in}}^3 + 2d_{\mathrm{in}}^2 d_{\mathrm{out}} m^2 + 2d_{\mathrm{in}}^2 d_{\mathrm{out}} \eta^2 + 11 d_{\mathrm{in}}^2 + 10 d_{\mathrm{in}} d_{\mathrm{out}} m^2 + 10 d_{\mathrm{in}} d_{\mathrm{out}} \eta^2}{d_{\mathrm{out}}^2 d_{\mathrm{in}}(d_{\mathrm{in}}+1)(d_{\mathrm{in}}+2)(d_{\mathrm{in}}+3)} \tag{4.234}$$

$$+ \frac{6d_{\mathrm{in}} + d_{\mathrm{out}}^2 m^4 + 2d_{\mathrm{out}}^2 \eta^2 + 12 d_{\mathrm{out}} m^2 + 40 d_{\mathrm{out}}^2 m \eta^3 + 81 d_{\mathrm{out}}^2 \eta^4 + 12 d_{\mathrm{out}} \eta^2}{d_{\mathrm{out}}^2 d_{\mathrm{in}}(d_{\mathrm{in}}+1)(d_{\mathrm{in}}+2)(d_{\mathrm{in}}+3)} \tag{4.235}$$

Recall that for the random variable $X = \mathrm{Tr}\left(\mathcal{N}(|\phi\rangle\langle\phi|) - \frac{\mathbb{I}}{d_{\mathrm{out}}} \right)^2 = \mathrm{Tr}\left(\mathcal{N}(|\phi\rangle\langle\phi|) \right)^2 - \frac{1}{d_{\mathrm{out}}}$

Permutation α	**Upper bound on** $F(\alpha)$	**Reference**
(13)	$\left(\frac{d_{\text{in}}}{d_{\text{out}}}+\eta^2\right)^2$	
$\text{id}, (132), (314), (24)(13)$	$\frac{d_{\text{in}}/d_{\text{out}}+\eta^2}{d_{\text{out}}}+\frac{\eta^2}{d_{\text{out}}}+5\eta^4$	(Lemma 4.4.7)
$(312), (134)$	$\left(\frac{d_{\text{in}}^2}{d_{\text{out}}}+m^2\right)\left(\frac{d_{\text{in}}}{d_{\text{out}}}+\eta^2\right)$	
(1234)	$\left(\frac{d_{\text{in}}^2}{d_{\text{out}}}+m^2\right)^2$	
$(24), (1432)$	$\frac{d_{\text{in}}^2}{d_{\text{out}}^2}+\frac{2m^2}{d_{\text{out}}}+25\eta^4$	(Lemma 4.4.8)
$(142), (243)$	$\frac{d_{\text{in}}/d_{\text{out}}+\eta^2}{d_{\text{out}}}+\frac{\eta^2}{d_{\text{out}}}+5\eta^4$	(Lemma 4.4.6)
$(14), (12), (23), (34), (1324), (1423), (1243), (1342)$	$\frac{d_{\text{in}}^2}{d_{\text{out}}^2}+\frac{d_{\text{in}}}{d_{\text{out}}}\eta^2+\frac{m^2}{d_{\text{out}}}+5m\eta^3$	(Lemma 4.4.10)
$(12)(34), (14)(23), (234), (124)$	$\frac{d_{\text{in}}^3}{d_{\text{out}}^2}+2\frac{d_{\text{in}}}{d_{\text{out}}}m^2+m^2\eta^2$	(Lemma 4.4.9)

we have:

$$\mathbb{E}(X) = \frac{\left(\text{Tr}\left(\mathcal{N}(\mathbb{I})-\frac{d_{\text{in}}}{d_{\text{out}}}\mathbb{I}\right)^2+d_{\text{in}}^2\left\|\mathcal{J}-\frac{\mathbb{I}}{d_{\text{in}}d_{\text{out}}}\right\|_2^2\right)}{d_{\text{in}}(d_{\text{in}}+1)} \tag{4.236}$$

$$= \frac{1}{d_{\text{in}}(d_{\text{in}}+1)}\left(\text{Tr}(\mathcal{M}(\mathbb{I})^2)+\eta^2\right) = \frac{m^2+\eta^2}{d_{\text{in}}(d_{\text{in}}+1)}. \tag{4.237}$$

Since $\text{Var}(X) = \text{Var}\left(X+\frac{1}{d_{\text{out}}}\right)$, it can be upper bounded as follows:

$$\text{Var}(X) = \mathbb{E}\left(\left(\text{Tr}(\mathcal{N}(|\phi\rangle\langle\phi|)^2)\right)^2\right)-\left(\mathbb{E}\left(\text{Tr}(\mathcal{N}(|\phi\rangle\langle\phi|)^2)\right)\right)^2 \tag{4.238}$$

$$\leq \frac{d_{\text{out}}^4+6d_{\text{in}}^3+2d_{\text{in}}^2d_{\text{out}}m^2+2d_{\text{in}}^2d_{\text{out}}\eta^2+11d_{\text{in}}^2+10d_{\text{in}}d_{\text{out}}m^2+10d_{\text{in}}d_{\text{out}}\eta^2}{d_{\text{out}}^2d_{\text{in}}(d_{\text{in}}+1)(d_{\text{in}}+2)(d_{\text{in}}+3)} \tag{4.239}$$

$$+\frac{6d_{\text{in}}+d_{\text{out}}^2m^4+2d_{\text{out}}^2\eta^2+12d_{\text{out}}m^2+40d_{\text{out}}^2m\eta^3+81d_{\text{out}}^2\eta^4+12d_{\text{out}}\eta^2}{d_{\text{out}}^2d_{\text{in}}(d_{\text{in}}+1)(d_{\text{in}}+2)(d_{\text{in}}+3)} \tag{4.240}$$

$$-\left(\frac{m^2+\eta^2}{d_{\text{in}}(d_{\text{in}}+1)}+\frac{1}{d_{\text{out}}}\right)^2 \tag{4.241}$$

$$\leq \left(\frac{80d_{\text{in}}^2\eta^4+10d_{\text{in}}^2m^2\eta^2+40d_{\text{in}}^2\eta^3m}{d_{\text{in}}^2(d_{\text{in}}+1)^2(d_{\text{in}}+2)(d_{\text{in}}+3)}\right). \tag{4.242}$$

Therefore the upper bound on the variance becomes using the inequalities $(m^2+\eta^2)^2 \geq 4m^2\eta^2$ and $(m^2+\eta^2)^2 \geq 2m\eta^3$ (successive AM-GM):

$$\frac{\text{Var}(X)}{\mathbb{E}(X)^2} \leq \left(\frac{80d_{\text{in}}^2\eta^4+10d_{\text{in}}^2m^2\eta^2+40d_{\text{in}}^2\eta^3m}{(d_{\text{in}}+2)(d_{\text{in}}+3)(\eta^2+m^2)^2}\right) \tag{4.243}$$

$$\leq 80\frac{d_{\text{in}}^2\eta^4}{d_{\text{in}}^2\eta^4}+10\frac{d_{\text{in}}^2m^2\eta^2}{4d_{\text{in}}^2m^2\eta^2}+40\frac{d_{\text{in}}^2\eta^3m}{2d_{\text{in}}^2m\eta^3} \leq 105. \tag{4.244}$$

□

We have now the required tools to design and prove the correctness of an algorithm for testing identity to the depolarizing channel.

Theorem 4.4.2. *There is an ancilla-free algorithm requiring a number of incoherent measurements* $N = \mathcal{O}\left(\frac{d_{\text{in}}^2 d_{\text{out}}^{1.5}}{\varepsilon^2}\right)$ *to distinguish between* $\mathcal{N} = \mathcal{D}$ *and* $d_\diamond(\mathcal{N}, \mathcal{D}) \geq \varepsilon$ *with a success probability* $2/3$.

As explained before, our algorithm is a reduction to the testing identity of quantum states. Note that we need to test quantum states in the 2-norm which is different than the usual quantum state certification [BCL20]. The algorithm for testing identity to the depolarizing channel is described in Algorithm 8.

Algorithm 8 Testing identity to the depolarizing channel in the diamond norm

$M = 2200$.
for $k = 1 : M$ **do**
  Sample ϕ_k a Haar random vector in $\mathbf{S}^{d_{\text{in}}}$.
  Test whether $h_0 : \mathcal{N}(|\phi_k\rangle\langle\phi_k|) = \frac{\mathbb{I}}{d_{\text{out}}}$ or $h_1 : \left\|\mathcal{N}(|\phi_k\rangle\langle\phi_k|) - \frac{\mathbb{I}}{d_{\text{out}}}\right\|_2 \geq \frac{\varepsilon}{2\sqrt{d_{\text{out}} d_{\text{in}}}}$ using the testing identity of quantum states Algorithm 5, with an error probability $\delta = 1/(3M)$, that answers the hypothesis h_{i_k}, $i_k \in \{0, 1\}$.
end for
if $\exists k : X_k = 1$ **then return** $\mathcal{N}$ is ε-far from $\mathcal{D}$ **else return** $\mathcal{N} = \mathcal{D}$.

We remark that Algorithm 8 uses the channel only on a constant number of random input states $\{|\phi_k\rangle\langle\phi_k|\}_k$. One could think that querying the channel $\mathcal{N}$ on more diverse inputs could lead to a more efficient algorithm. However, it turns out that Algorithm 8 is basically optimal as we prove a matching lower bound up to a poly-logarithmic factor.

Theorem 4.4.3. *Let* $\varepsilon \leq 1/32$ *and* $d_{\text{out}} \geq 10$. *Any ancilla-free non-adaptive algorithm for testing identity to the depolarizing channel requires, in the worst case, a number of measurements satisfying:*

$$N = \Omega\left(\frac{d_{\text{in}}^2 d_{\text{out}}^{1.5}}{\log(d_{\text{in}} d_{\text{out}}/\varepsilon)^2 \varepsilon^2}\right). \tag{4.245}$$

This theorem shows that our proposed Algorithm 8 is almost optimal in the dimensions $(d_{\text{in}}, d_{\text{out}})$ and the precision parameter ε thus the complexity of testing identity to the depolarizing channel is $\tilde{\Theta}(d_{\text{in}}^2 d_{\text{out}}^{1.5}/\varepsilon^2)$ which is slightly surprising. Indeed, we can remark that the complexity of testing identity of discrete distributions and quantum states is the square root (for constant ε) of the complexity of the corresponding learning problems in the same setting. This rule does not apply for quantum channels since we know from [SSKKG22; Ouf23] that the complexity of learning quantum channels in the diamond distance with ancilla-free non-adaptive incoherent measurements is $\tilde{\Theta}(d_{\text{in}}^3 d_{\text{out}}^3/\varepsilon^2)$.

Outine of the proof. Under the null hypothesis $\mathcal{N} = \mathcal{D}$. Under the alternate hypothesis, we construct randomly the quantum channel $\mathcal{N} \sim \mathcal{P}$ of the form: $\mathcal{N}(\rho) = \mathcal{D}(\rho) + \frac{\varepsilon}{d_{\text{out}}} \langle w| \rho |w\rangle U$ where $|w\rangle$ is a Haar distributed vector and U has Gaussian entries: for all $i, j \in [d_{\text{out}}]$, $U_{j,i} = U_{i,j} \sim \mathbb{1}\{i \neq j\}\mathcal{N}(0, 16/d_{\text{out}})$ conditioned on the event $\mathcal{G} = \{\|U\|_1 \geq d_{\text{out}}, \|U\|_\infty \leq 32\}$. Note that the usual construction applied on the Choi state gives a sub-optimal lower bound in the diamond distance. Using a concentration inequality of Lipschitz functions of Gaussian random variables [Wai19], we show that with a high probability $\mathcal{N}$ is ε-far from $\mathcal{D}$ in the diamond distance. Then we use LeCam's method [LeC73] to lower bound the TV distance between the distribution of the observations under the two hypotheses: $\text{TV}\left(\mathbb{P}_{\mathcal{D}}^{I_1,\dots,I_N} \middle\| \mathbb{E}_{\mathcal{N}\sim\mathcal{P}} \mathbb{P}_{\mathcal{N}}^{I_1,\dots,I_N}\right) \geq \frac{1}{3}$ where

$I_1, \dots, I_N$ are the observations the algorithm obtains after the measurements and N is a sufficient number of measurements for the correctness of the algorithm. Next, we condition on the event $\mathcal{E}$ that w satisfies: $\forall t \in [N] : \langle w| \rho_t |w\rangle \leq \frac{20\log(9N)}{d_{\text{in}}}$ which occurs with a probability at least 8/9. For this, we take $W \sim \text{Haar}(d_{\text{in}})$ such that $W|0\rangle = |w\rangle$ and we show that the function $W \mapsto \sqrt{\langle 0| W^* \rho W |0\rangle}$ is 1-Lipschitz using Minkonwski's inequality then we use a concentration inequality of Lipschitz functions of Haar distributed unitary [MM13]. Under the event $\mathcal{E}$, the channel certification problem becomes similar to the state certification problem with a precision parameter $\varepsilon' \approx \frac{20\log(9N)\varepsilon}{d_{\text{in}}}$ and thus we could mimic the proof of [BCL20]. We choose instead to present a shorter proof using the Hypercontractivity Theorem (see e.g., [Jan97; AS17]) which can simplify significantly the analysis of [BCL20]. To carry out the analysis, we need to bound the moments of the random variables: $Z_t(U,V) = \varepsilon^2 \left(\frac{20\log(9N)}{d_{\text{in}}}\right)^2 \left(\sum_{i_t} \frac{\lambda^t_{i_t}}{d_{\text{out}}} \langle \phi^t_{i_t}| U |\phi^t_{i_t}\rangle \langle \phi^t_{i_t}| V |\phi^t_{i_t}\rangle\right)$ where $t \in [N]$ and $\left\{\lambda^t_{i_t} |\phi^t_{i_t}\rangle\langle\phi^t_{i_t}|\right\}_{i_t}$ is the POVM chosen by the algorithm at step t. Since Z_t is a polynomial of degree 2 (in the entries of U and V) of expectation 0, the Hypercontractivity [AS17, Proposition 5.48] implies for all $k \in \{1, \dots, N\} : \mathbb{E}\left(|Z_t|^k\right) \leq k^k \mathbb{E}\left(Z_t^2\right)^{k/2}$. Hence, it is sufficient to upper bound the second moment which can be done easily: $\mathbb{E}\left(Z_t^2\right) \leq \mathcal{O}\left(\frac{\varepsilon^4 \log(N)^4}{d_{\text{in}}^4 d_{\text{out}}^3}\right)$. By a contradiction argument and grouping all these elements, we can prove that $N\log(N)^2 \geq \Omega\left(\frac{d_{\text{in}}^2 d_{\text{out}}^{1.5}}{\varepsilon^2}\right)$ and finally $N \geq \Omega\left(\frac{d_{\text{in}}^2 d_{\text{out}}^{1.5}}{\log(d_{\text{in}} d_{\text{out}}/\varepsilon)^2 \varepsilon^2}\right)$. We present the proof in the following.

Proof. **Construction.** Under the null hypothesis H_0 the quantum channel is $\mathcal{N}(\rho) = \mathcal{D}(\rho) = \text{Tr}(\rho)\frac{\mathbb{I}}{d_{\text{out}}}$. Under the alternate hypothesis H_1, we choose the quantum channel $\mathcal{N} \sim \mathcal{P}$ of the form:

$$\mathcal{N}(\rho) = \text{Tr}(\rho)\frac{\mathbb{I}}{d_{\text{out}}} + \frac{\varepsilon}{d_{\text{out}}} \langle w| \rho |w\rangle U \tag{4.246}$$

where $|w\rangle = W|0\rangle$ and $W \sim \text{Haar}(d_{\text{in}})$ and for all $i, j \in [d_{\text{out}}]$, $U_{j,i} = U_{i,j} \sim \mathbf{1}\{i \neq j\}\mathcal{N}(0, \sigma^2)$ where the parameter σ would be chosen later and we condition on the event $\mathcal{G} = \{\|U\|_1 \geq d_{\text{out}}, \|U\|_\infty \leq 32\}$. We call this distribution $\mathcal{P}$ and use the notation $(w, U) \sim \mathcal{P}$. If we do not condition on the event $\mathcal{G}$, the distribution of U is denoted $\mathcal{P}_0$ and we write $U \sim \mathcal{P}_0$. Random constructions with Gaussian random variables were used for proving lower bounds by [CHLL22; CHLLS22]. Note that $\mathcal{N}$ is trace preserving since $\text{Tr}(U) = 0$. It remains to show that $\mathcal{N}$ is completely positive which is equivalent to proving the corresponding Choi matrix is positive semi-definite. For this we can express the Choi state of the channel $\mathcal{N}$:

$$\mathcal{J}_{\mathcal{N}} = \frac{\mathbb{I}}{d_{\text{in}} d_{\text{out}}} + \frac{\varepsilon}{d_{\text{in}} d_{\text{out}}} \sum_{i,j=1}^{d_{\text{in}}} |i\rangle\langle j| \otimes \langle 0| W^* |i\rangle\langle j| W |0\rangle U \tag{4.247}$$

$$= \frac{\mathbb{I}}{d_{\text{in}} d_{\text{out}}} + \frac{\varepsilon}{d_{\text{in}} d_{\text{out}}} \sum_{i,j} |i\rangle\langle j| \otimes \langle i| \bar{W} |0\rangle\langle 0| W^\top |j\rangle U \tag{4.248}$$

$$= \frac{\mathbb{I}}{d_{\text{in}} d_{\text{out}}} + \frac{\varepsilon}{d_{\text{in}} d_{\text{out}}} \bar{W}|0\rangle\langle 0| W^\top \otimes U. \tag{4.249}$$

Observe that $\|\bar{W}|0\rangle\langle 0| W^\top \otimes U\|_\infty = \|U\|_\infty \leq 32$ thus $\mathcal{J}_{\mathcal{N}} \geq 0$ if $\varepsilon \leq 1/32$. So under the event $\mathcal{G}$, the map $\mathcal{N}$ is a quantum channel. The parameter σ should be chosen so that

$d_\diamond(\mathcal{N},\mathcal{D}) \geq \varepsilon$. Recall that $|w\rangle = W|0\rangle$, the definition of the diamond distance implies

$$d_\diamond(\mathcal{N},\mathcal{D}) = \max_\rho \|\mathrm{id}\otimes(\mathcal{N}-\mathcal{D})(\rho)\|_1 \geq \|(\mathcal{N}-\mathcal{D})(|w\rangle\langle w|)\|_1 = \frac{\varepsilon}{d_{\text{out}}}\|U\|_1. \tag{4.250}$$

Lemma 4.4.12. *There is a constant $c>0$ such that we have:*

$$\mathbb{P}\left(\left|\|U\|_1 - \mathbb{E}\left(\|U\|_1\right)\right| > s\right) \leq \exp\left(-\frac{cs^2}{d_{\text{out}}\sigma^2}\right). \tag{4.251}$$

Proof. The function $U \mapsto \|U\|_1$ is $\sqrt{d_{\text{out}}}$-Lipschitz w.r.t. the Hilbert-Schmidt norm. Indeed, by the triangle inequality and the Cauchy Schwarz inequality $|\|U\|_1 - \|V\|_1| \leq \|U-V\|_1 \leq \sqrt{d_{\text{out}}}\|U-V\|_2$. The concentration of Lipschitz functions of Gaussian random variables [Wai19, Theorem 2.26] yields exactly the desired statement. □

Next, we need a lower bound on the expectation of $\|U\|_1$. By the Hölder's inequality:

$$\mathbb{E}\left(\|U\|_1\right) \geq \sqrt{\frac{\mathbb{E}\left(\|U\|_2^2\right)^3}{\mathbb{E}\left(\|U\|_4^4\right)}} \geq \sqrt{\frac{(d_{\text{out}}^2\sigma^2)^3}{4d_{\text{out}}^3\sigma^4}} \geq \frac{d_{\text{out}}\sqrt{d_{\text{out}}}\sigma}{2}. \tag{4.252}$$

Since $d_\diamond(\mathcal{N},\mathcal{D}) \geq \frac{\varepsilon}{d_{\text{out}}}\|U\|_1$, it is sufficient to choose $\sigma = \frac{4}{\sqrt{d_{\text{out}}}}$ so that $\mathbb{E}\left(\|U\|_1\right) \geq 2d_{\text{out}}$ and by Lemma 4.4.12, we have $\|U\|_1 \geq d_{\text{out}}$ with a probability $1-\exp(-\Omega(d_{\text{out}}^2))$. Therefore with an overwhelming probability we have $d_\diamond(\mathcal{N},\mathcal{D}) \geq \frac{\varepsilon}{d_{\text{out}}}\|U\|_1 \geq \varepsilon$. It remains to see that the event $\{\|U\|_\infty \leq 32\}$ also occurs with high probability. Indeed, let $\mathcal{S}$ be a $1/4$-net of $\mathbf{S}^{d_{\text{out}}}$ of size at most $8^{d_{\text{out}}}$. By the union bound:

$$\begin{aligned}
\mathbb{P}\left(\|U\|_\infty > 32\right) &= \mathbb{P}\left(\exists\phi\in\mathbf{S}^{d_{\text{out}}} : \langle\phi|U|\phi\rangle = \|U\|_\infty, \|U\|_\infty > 32\right) && (4.253)\\
&\leq \mathbb{P}\left(\exists\phi\in\mathcal{S} : \langle\phi|U|\phi\rangle > \frac{1}{2}\|U\|_\infty, \|U\|_\infty > 32\right) && (4.254)\\
&\leq |\mathcal{S}|\mathbb{P}\left(\langle\phi|U|\phi\rangle > 16\right) \leq 8^{d_{\text{out}}}e^{-8d_{\text{out}}} \leq e^{-4d_{\text{out}}}. && (4.255)
\end{aligned}$$

Finally, with a probability at least $1-\exp(-\Omega(d_{\text{out}}^2)) - \exp(-\Omega(d_{\text{out}}))$ the event $\mathcal{G}$ is satisfied and we have a quantum channel $\mathcal{N}$ that is ε-far in the diamond distance from the depolarizing channel $\mathcal{D}$. A $1/3$-correct algorithm $\mathcal{A}$ should distinguish between the channels $\mathcal{N}$ and $\mathcal{D}$ with a probability of error at most $1/3$. Let N be a sufficient number of measurements for this task and $I_1,\ldots,I_N$ be the observations of the algorithm $\mathcal{A}$. The Data-Processing inequality applied on the TV-distance gives LeCam's method [LeC73]:

$$\begin{aligned}
\mathrm{TV}\left(\mathbb{P}_{H_0}^{I_1,\ldots,I_N}\middle\|\mathbb{P}_{H_1}^{I_1,\ldots,I_N}\right) &\geq \mathrm{TV}\left(\mathrm{Bern}(\mathbb{P}_{H_0}(\mathcal{A}=1))\|\,\mathrm{Bern}(\mathbb{P}_{H_1}(\mathcal{A}=1))\right) && (4.256)\\
&\geq \mathrm{TV}(\mathrm{Bern}(1/3)\|\,\mathrm{Bern}(2/3)) = \frac{1}{3}. && (4.257)
\end{aligned}$$

Now, we need to upper bound this TV distance with an expression involving $N, d_{\text{in}}, d_{\text{out}}$ and ε.

Upper bound on the TV distance. The non-adaptive algorithm $\mathcal{A}$ would choose at step t the input ρ_t and the measurement device $\mathcal{M}_t = \{\lambda_i^t|\phi_i^t\rangle\langle\phi_i^t|\}_{i\in\mathcal{I}_t}$. Observe that we can always reduce w.l.o.g. to such a POVM. Moreover, we have $\sum_i \lambda_i^t = d_{\text{out}}$. Under the null hypothesis H_0, the quantum channel $\mathcal{N}=\mathcal{D}$ so the probability of the outcomes is exactly:

$$\mathbb{P}_{H_0}^{I_1,\ldots,I_N} = \left\{\prod_{t=1}^N \frac{\lambda_{i_t}^t}{d_{\text{out}}}\right\}_{i_1,\ldots,i_N}. \tag{4.258}$$

On the other hand, under the alternate hypothesis H_1, the probability of the outcomes is exactly:

$$\mathbb{P}_{H_1}^{I_1,\dots,I_N} = \left\{ \mathbb{E}_{(w,U)\sim\mathcal{P}} \prod_{t=1}^{N} \frac{\lambda_{i_t}^t}{d_{\text{out}}^N} \left(1 + \varepsilon \langle w| \rho_t |w\rangle \left\langle \phi_{i_t}^t \right| U \left| \phi_{i_t}^t \right\rangle\right) \right\}_{i_1,\dots,i_N} \tag{4.259}$$

We can express the TV distance as follows:

$$2\,\mathrm{TV}\left(\mathbb{P}_{H_0}^{I_1,\dots,I_N} \middle\| \mathbb{P}_{H_1}^{I_1,\dots,I_N}\right) \tag{4.260}$$

$$= \sum_{i_1,\dots,i_N} \left| \mathbb{E}_{(w,U)\sim\mathcal{P}} \prod_{t=1}^{N} \frac{\lambda_{i_t}^t}{d_{\text{out}}} \left(1 + \varepsilon \langle w| \rho_t |w\rangle \left\langle \phi_{i_t}^t \right| U \left| \phi_{i_t}^t \right\rangle\right) - \prod_{t=1}^{N} \frac{\lambda_{i_t}^t}{d_{\text{out}}} \right| \tag{4.261}$$

$$= \mathbb{E}_{\leq N} \left| \mathbb{E}_{(w,U)\sim\mathcal{P}} \prod_{t=1}^{N} \left(1 + \varepsilon \langle w| \rho_t |w\rangle \left\langle \phi_{i_t}^t \right| U \left| \phi_{i_t}^t \right\rangle\right) - 1 \right| \tag{4.262}$$

where we use the notation $\mathbb{E}_{\leq N}(X(i_1,\dots,i_N)) = \sum_{i_1,\dots,i_N} \left(\prod_{t=1}^N \frac{\lambda_{i_t}^t}{d_{\text{out}}}\right) X(i_1,\dots,i_N)$. Now, we need first to remove the terms $\{\langle w| \rho_t |w\rangle\}_t$ because, as we will see later, we would like to construct a polynomial of a small degree. If these terms remain, they would increase the degree. Since the algorithm is non-adaptive, we have at most N distinct input states $\{\rho_t\}_t$. So let us condition on the event $\mathcal{E}$ that w satisfies:

$$\forall t \in [N]: \qquad \langle 0| W^* \rho_t W |0\rangle = \langle w| \rho_t |w\rangle \leq \frac{20\log(9N)}{d_{\text{in}}}. \tag{4.263}$$

The function $f : W \mapsto \sqrt{\langle 0| W^* \rho W |0\rangle}$ is 1-Lipschitz. Indeed, we can write $\rho = \sum_i \lambda_i |\phi_i\rangle\langle\phi_i|$ then since $\{\lambda_i\}_i$ is a probability, we have by Minkowski's inequality:

$$|f(W) - f(V)| = |\sqrt{\mathbb{E}_{i\sim\lambda}(|\langle\phi_i| W |0\rangle|^2)} - \sqrt{\mathbb{E}_{i\sim\lambda}(|\langle\phi_i| V |0\rangle|^2)}| \tag{4.264}$$

$$\leq \sqrt{\mathbb{E}_{i\sim\lambda}(|\langle\phi_i| (W-V) |0\rangle|^2)} \leq \|W - U\|_2. \tag{4.265}$$

Hence by the concentration inequality for Lipschitz functions of a Haar-distributed matrix [MM13]:

$$\mathbb{P}\left(|f(W) - \mathbb{E}(f)| > \sqrt{\frac{12\log(9N)}{d_{\text{in}}}} \right) \leq \exp\left(-\frac{12 d_{\text{in}} \log(9N)}{12 d_{\text{in}}}\right) = \frac{1}{9N}. \tag{4.266}$$

Moreover, we can upper bound the expectation $\mathbb{E}(f(W)) \leq \sqrt{\mathbb{E}(f(W)^2)} = \sqrt{1/d_{\text{in}}}$. Therefore by the union bound:

$$\mathbb{P}(\bar{\mathcal{E}}) = \mathbb{P}\left(\exists t \in [N] : \langle 0| W^* \rho_t W |0\rangle \geq \frac{20\log(9N)}{d_{\text{in}}}\right) \leq N\mathbb{P}\left(f(W) \geq \sqrt{\frac{20\log(9N)}{d_{\text{in}}}} \right) \tag{4.267}$$

$$\leq N\mathbb{P}\left(f(W) - \mathbb{E}(f) \geq \sqrt{\frac{12\log(9N)}{d_{\text{in}}}} \right) \leq \frac{1}{9}. \tag{4.268}$$

Now, we can distinguish whether the event $\mathcal{E}$ is verified or not in the TV distance. Let $\Psi_{i,w,U} = \prod_{t=1}^{N} \left(1 + \varepsilon \langle w| \rho_t |w\rangle \left\langle \phi_{i_t}^t \right| U \left| \phi_{i_t}^t \right\rangle\right)$, by the triangle inequality:

$$\mathbb{E}_{\leq N} \left| \mathbb{E}_{(w,U)\sim\mathcal{P}} \prod_{t=1}^{N} \left(1 + \varepsilon \langle w| \rho_t |w\rangle \left\langle \phi_{i_t}^t \right| U \left| \phi_{i_t}^t \right\rangle\right) - 1 \right| \tag{4.269}$$

$$\leq \mathbb{E}_{\leq N} \left| \mathbb{E}_{(w,U)\sim\mathcal{P}} \left[\mathbf{1}\{\mathcal{E}\}(\Psi_{i,w,U} - 1)\right] \right| + \mathbb{E}_{\leq N} \left| \mathbb{E}_{(w,U)\sim\mathcal{P}} \left[\mathbf{1}\{\bar{\mathcal{E}}\}(\Psi_{i,w,U} - 1)\right] \right| \tag{4.270}$$

$$\leq \mathbb{E}_{\leq N} \left| \mathbb{E}_{(w,U)\sim\mathcal{P}} \left[\mathbf{1}\{\mathcal{E}\}(\Psi_{i,w,U} - 1)\right] \right| + \mathbb{E}_{\leq N} \mathbb{E}_{(w,U)\sim\mathcal{P}} \left[\mathbf{1}\{\bar{\mathcal{E}}\}(\Psi_{i,w,U} + 1)\right] \tag{4.271}$$

$$= \mathbb{E}_{\leq N} \left| \mathbb{E}_{(w,U)\sim\mathcal{P}} \left[\mathbf{1}\{\mathcal{E}\}(\Psi_{i,w,U} - 1)\right] \right| + 2\mathbb{P}\left(\bar{\mathcal{E}}\right) \tag{4.272}$$

where we use the fact that $\mathbb{E}_{\leq N} \Psi_{i,w,U} = \sum_{i_1,\ldots,i_N} \left(\prod_{t=1}^{N} \frac{\lambda_{i_t}^t}{d_{\text{out}}}\right) \prod_{t=1}^{N} \left(1 + \varepsilon \langle w| \rho_t |w\rangle \left\langle \phi_{i_t}^t \right| U \left| \phi_{i_t}^t \right\rangle\right) = \prod_{t=1}^{N} \left(1 + \varepsilon \langle w| \rho_t |w\rangle \operatorname{Tr}(U)\right) = 1$. It remains to upper bound this latter expectation. For this, we follow [BCL20] and apply the Cauchy Schwarz inequality and Hölder's inequality:

$$\left(\mathbb{E}_{\leq N} \left| \mathbb{E}_{(w,U)\sim\mathcal{P}} \left[\mathbf{1}\{\mathcal{E}\}(\Psi_{i,w,U} - 1)\right] \right|\right)^2 + \mathbb{P}\left(\mathcal{E}\right) \leq \mathbb{E}_{\leq N} \left(\mathbb{E}_{(w,U)\sim\mathcal{P}} \left[\mathbf{1}\{\mathcal{E}(w)\}(\Psi_{i,w,U} - 1)\right]\right)^2 + \mathbb{P}\left(\mathcal{E}\right) \tag{4.273}$$

$$= \mathbb{E}_{\leq N} \mathbb{E}_{(w,U)\sim\mathcal{P}} \mathbb{E}_{(z,V)\sim\mathcal{P}} \mathbf{1}\{\mathcal{E}(w)\} \Psi_{i,w,U} \mathbf{1}\{\mathcal{E}(z)\} \Psi_{i,z,V} \tag{4.274}$$

$$\leq \mathbb{E}_{\substack{(w,U)\sim\mathcal{P} \\ (z,V)\sim\mathcal{P}}} \mathbf{1}\{\mathcal{E}(w), \mathcal{E}(z)\} \prod_{t=1}^{N} \sum_{i_t} \frac{\lambda_{i_t}^t}{d_{\text{out}}} \left(1 + \varepsilon \langle w| \rho_t |w\rangle \left\langle \phi_{i_t}^t \right| U \left| \phi_{i_t}^t \right\rangle\right) \left(1 + \varepsilon \langle z| \rho_t |z\rangle \left\langle \phi_{i_t}^t \right| V \left| \phi_{i_t}^t \right\rangle\right) \tag{4.275}$$

$$= \mathbb{E}_{\substack{(w,U)\sim\mathcal{P} \\ (z,V)\sim\mathcal{P}}} \mathbf{1}\{\mathcal{E}(w), \mathcal{E}(z)\} \prod_{t=1}^{N} \left(1 + \varepsilon^2 \langle w| \rho_t |w\rangle \langle z| \rho_t |z\rangle \sum_{i_t} \frac{\lambda_{i_t}^t}{d_{\text{out}}} \left\langle \phi_{i_t}^t \right| U \left| \phi_{i_t}^t \right\rangle \left\langle \phi_{i_t}^t \right| V \left| \phi_{i_t}^t \right\rangle\right) \tag{4.276}$$

$$\leq \mathbb{E}_{U\sim\mathcal{P}} \mathbb{E}_{V\sim\mathcal{P}} \prod_{t=1}^{N} \left(1 + \varepsilon^2 \left(\frac{20\log(9N)}{d_{\text{in}}}\right)^2 \left| \sum_{i_t} \frac{\lambda_{i_t}^t}{d_{\text{out}}} \left\langle \phi_{i_t}^t \right| U \left| \phi_{i_t}^t \right\rangle \left\langle \phi_{i_t}^t \right| V \left| \phi_{i_t}^t \right\rangle \right|\right) \tag{4.277}$$

$$\leq \frac{1}{\mathbb{P}\left(\mathcal{G}\right)} \max_{1\leq t\leq N} \mathbb{E}_{U,V\sim\mathcal{P}_0} \left(1 + \varepsilon^2 \left(\frac{20\log(9N)}{d_{\text{in}}}\right)^2 \left| \sum_{i_t} \frac{\lambda_{i_t}^t}{d_{\text{out}}} \left\langle \phi_{i_t}^t \right| U \left| \phi_{i_t}^t \right\rangle \left\langle \phi_{i_t}^t \right| V \left| \phi_{i_t}^t \right\rangle \right|\right)^N \tag{4.278}$$

$$\leq \frac{1}{\left(1 - e^{-\Omega(d_{\text{out}})}\right)} \max_{1\leq t\leq N} \mathbb{E}_{U,V\sim\mathcal{P}_0} \left(1 + \varepsilon^2 \left(\frac{20\log(9N)}{d_{\text{in}}}\right)^2 \left| \sum_{i_t} \frac{\lambda_{i_t}^t}{d_{\text{out}}} \left\langle \phi_{i_t}^t \right| U \left| \phi_{i_t}^t \right\rangle \left\langle \phi_{i_t}^t \right| V \left| \phi_{i_t}^t \right\rangle \right|\right)^N \tag{4.279}$$

because for all $t \in [N]$ we have $\sum_{i_t} \frac{\lambda_{i_t}^t}{d_{\text{out}}} \left\langle \phi_{i_t}^t \right| U \left| \phi_{i_t}^t \right\rangle = \operatorname{Tr}(U) = 0$ and under the event $\mathcal{E}$, $\langle w| \rho_t |w\rangle \langle z| \rho_t |z\rangle \leq \left(\frac{20\log(9N)}{d_{\text{in}}}\right)^2$. Note that at the last inequality, we do not require anymore that U satisfies $\|U\|_\infty \leq 32$ and $\|U\|_1 \geq d_{\text{out}}$. This is possible because the integrand is positive and $\mathbb{P}\left(\mathcal{G}\right) \geq 1 - e^{-\Omega(d_{\text{out}})}$.

For $t \in [N]$, let Z_t be the polynomial in $\{U_{i,j}, V_{i,j}\}_{i,j=1}^{d_{\text{out}}}$ defined as follows:

$$Z_t = \varepsilon^2 \left(\frac{20\log(9N)}{d_{\text{in}}}\right)^2 \left(\sum_{i_t} \frac{\lambda_{i_t}^t}{d_{\text{out}}} \left\langle \phi_{i_t}^t \right| U \left| \phi_{i_t}^t \right\rangle \left\langle \phi_{i_t}^t \right| V \left| \phi_{i_t}^t \right\rangle\right). \tag{4.280}$$

[BCL20] gives a different method to control the moments of a similar function of Haar distributed unitaries. However, our method is shorter compared to theirs since we only need to control the second moment. Note that Z_t is a polynomial of degree 2 of expectation 0. The Hypercontractivity [AS17, Proposition 5.48] implies for all $k \in \{1, \dots, N\}$:

$$\mathbb{E}\left(|Z_t|^k\right) \leq k^k \mathbb{E}\left(Z_t^2\right)^{k/2}. \tag{4.281}$$

This means that we only need to control the second moment of Z_t. We have:

$$\mathbb{E}\left(Z_t^2\right) = \varepsilon^4 \left(\frac{20\log(9N)}{d_{\text{in}}}\right)^4 \mathbb{E}_{U,V}\left(\sum_{i_t,j_t} \frac{\lambda_{i_t}^t \lambda_{j_t}^t}{d_{\text{out}}^2} \left\langle \phi_{i_t}^t \right| U \left| \phi_{i_t}^t \right\rangle \left\langle \phi_{i_t}^t \right| V \left| \phi_{i_t}^t \right\rangle \left\langle \phi_{j_t}^t \right| U \left| \phi_{j_t}^t \right\rangle \left\langle \phi_{j_t}^t \right| V \left| \phi_{j_t}^t \right\rangle\right). \tag{4.282}$$

For given i_t, j_t, we can upper bound the expectation:

$$\begin{aligned}\mathbb{E}\left(\left\langle \phi_{i_t}^t \right| U \left| \phi_{i_t}^t \right\rangle \left\langle \phi_{j_t}^t \right| U \left| \phi_{j_t}^t \right\rangle\right) &= \sum_{x,y,x',y'} \mathbb{E}\left(U_{x,y} U_{x',y'}\right) \langle \phi_{i_t}^t | x\rangle\langle y | \phi_{i_t}^t\rangle \langle \phi_{j_t}^t | x'\rangle\langle y' | \phi_{j_t}^t\rangle \\ &\leq \frac{32}{d_{\text{out}}}(|\langle \phi_{i_t}^t | \phi_{j_t}^t\rangle|^2 + |\langle \phi_{i_t}^t | \bar{\phi}_{j_t}^t\rangle|^2)\end{aligned} \tag{4.283}$$

Therefore we can upper bound the expectation of Z_t^2 using the inequality $\sum_{i_t} \lambda_{i_t}^t |\langle \phi_{i_t}^t | \phi_{j_t}^t\rangle|^4 \leq \sum_{i_t} \lambda_{i_t}^t |\langle \phi_{i_t}^t | \phi_{j_t}^t\rangle|^2 = 1$ and the equality $\sum_{j_t} \lambda_{j_t}^t = d_{\text{out}}$:

$$\mathbb{E}\left(Z_t^2\right) \leq \varepsilon^4 \left(\frac{20\log(9N)}{d_{\text{in}}}\right)^4 \sum_{i_t,j_t} \frac{\lambda_{i_t}^t \lambda_{j_t}^t}{d_{\text{out}}^2} \left(\frac{32}{d_{\text{out}}}(|\langle \phi_{i_t}^t | \phi_{j_t}^t\rangle|^2 + |\langle \phi_{i_t}^t | \bar{\phi}_{j_t}^t\rangle|^2)\right)^2 \tag{4.284}$$

$$\leq \varepsilon^4 \left(\frac{20\log(9N)}{d_{\text{in}}}\right)^4 \left(\frac{2 \cdot 32^2}{d_{\text{out}}^2}\right) \sum_{i_t,j_t} \frac{\lambda_{i_t}^t \lambda_{j_t}^t}{d_{\text{out}}^2} \left(|\langle \phi_{i_t}^t | \phi_{j_t}^t\rangle|^4 + |\langle \phi_{i_t}^t | \bar{\phi}_{j_t}^t\rangle|^4\right) \tag{4.285}$$

$$\leq \frac{C\varepsilon^4 \log(N)^4}{d_{\text{in}}^4 d_{\text{out}}^3} \tag{4.286}$$

where $C > 0$ is a universal constant. This implies an upper bound for every moment:

$$\mathbb{E}\left(|Z_t|^k\right) \leq k^k \left(\frac{C\varepsilon^4 \log(N)^4}{d_{\text{in}}^4 d_{\text{out}}^3}\right)^{k/2}. \tag{4.287}$$

Now, grouping the lower bound and upper bounds on the TV distance, we obtain:

$$(1 - e^{-\Omega(d_{\text{out}})})\left(\frac{4^2}{9^2} + \frac{8}{9}\right) \leq \max_t \mathbb{E}\left((1 + |Z_t|)^N\right) \tag{4.288}$$

$$\leq \max_t \sum_{k=0}^{N} \binom{N}{k} k^k \left(\frac{C\varepsilon^4 \log(N)^4}{d_{\text{in}}^4 d_{\text{out}}^3}\right)^{k/2} \tag{4.289}$$

$$\leq \max_t \sum_{k=0}^{N} \left(\frac{C' N \varepsilon^2 \log(N)^2}{d_{\text{in}}^2 d_{\text{out}}^{1.5}}\right)^k \tag{4.290}$$

where we used $\binom{N}{k} \leq \frac{N^k e^k}{k^k}$ and $C' = \sqrt{C}e$. If $N\log(N)^2 \leq \frac{d_{\text{in}}^2 d_{\text{out}}^{1.5}}{101 C' \varepsilon^2}$ the RHS is upper bounded by $\sum_{k \geq 0} \frac{1}{101^k} = 1.01$ but the LHS is at least $(1 - e^{-\Omega(d_{\text{out}})})\left(\frac{4^2}{9^2} + \frac{8}{9}\right) \geq 1.05$ for $d_{\text{out}} \geq \Omega(1)$ which is a contradiction. Hence $N\log(N)^2 \geq \frac{d_{\text{in}}^2 d_{\text{out}}^{1.5}}{101 C' \varepsilon^2}$ and finally:

$$N \geq \Omega\left(\frac{d_{\text{in}}^2 d_{\text{out}}^{1.5}}{\log(d_{\text{in}} d_{\text{out}}/\varepsilon)^2 \varepsilon^2}\right). \tag{4.291}$$

□

This proof relies crucially on the non-adaptiveness of the strategy. A natural question arises then, can adaptive strategies outperform their non-adaptive counterpart? We do not settle completely this question. Yet, we propose a lower bound for adaptive strategies showing that, if a separation exists, the advantage would be at most $\mathcal{O}(\sqrt{d_{\text{out}}})$.

Theorem 4.4.4. *Let $\varepsilon \leq 1/32$ and $d_{\text{out}} \geq 10$. Any ancilla-free adaptive algorithm for testing identity to the depolarizing channel requires, in the worst case, $N = \Omega\left(\frac{d_{\text{in}}^2 d_{\text{out}} + d_{\text{out}}^{1.5}}{\varepsilon^2}\right)$ incoherent measurements.*

Proof. We use the same construction as in the proof of Theorem 4.4.3. Mainly, under the null hypothesis H_0 the quantum channel is $\mathcal{N}(\rho) = \mathcal{D}(\rho) = \text{Tr}(\rho)\frac{\mathbb{I}}{d_{\text{out}}}$. Under the alternate hypothesis H_1, a quantum channel $\mathcal{N} \sim \mathcal{P}$ has the form:

$$\mathcal{N}(\rho) = \text{Tr}(\rho)\frac{\mathbb{I}}{d_{\text{out}}} + \frac{\varepsilon}{d_{\text{out}}} \langle w| \rho |w\rangle U \tag{4.292}$$

where $|w\rangle = W|0\rangle$, $W \sim \text{Haar}(d_{\text{in}})$, for all $i, j \in [d_{\text{out}}]$,

$$U_{j,i} = U_{i,j} \sim \mathbf{1}\{i \neq j\}\mathcal{N}\left(0, \sigma^2 = \frac{16}{d_{\text{out}}}\right)$$

and we condition on the event $\mathcal{G} = \{\|U\|_1 \geq d_{\text{out}}, \|U\|_\infty \leq 32\}$. We call this distribution $\mathcal{P}$ and use the notation $(w, U) \sim \mathcal{P}$. Recall that $\mathbb{P}(\mathcal{G}) \geq 1 - \exp(-\Omega(d_{\text{out}})) - \exp(-\Omega(d_{\text{out}}^2))$ and under $\mathcal{G}$, the map $\mathcal{N}$ is a valid quantum channel ε-far from the quantum channel $\mathcal{D}$.

Now, given a set of observations $i_{<t} = (i_1, \ldots, i_{t-1})$. The adaptive algorithm $\mathcal{A}$ would choose at step t the input $\rho_t^{i_{<t}}$ and the measurement device $\mathcal{M}_t^{i_{<t}} = \left\{\lambda_{i_t}^{i_{<t}} \left|\phi_{i_t}^{i_{<t}}\right\rangle\left\langle\phi_{i_t}^{i_{<t}}\right|\right\}_{i_t \in \mathcal{I}_t}$. Such POVM implies $\sum_{i_t} \lambda_{i_t}^{i_{<t}} = d$. Under the null hypothesis H_0, the quantum channel $\mathcal{N} = \mathcal{D}$ so the probability of the outcomes is exactly:

$$\mathbb{P}_{H_0}^{I_1,\ldots,I_N} = \left\{\prod_{t=1}^N \frac{\lambda_{i_t}^{i_{<t}}}{d_{\text{out}}}\right\}_{i_1,\ldots,i_N} \tag{4.293}$$

On the other hand, under the alternate hypothesis H_1, the probability of the outcomes is exactly:

$$\mathbb{P}_{H_1}^{I_1,\ldots,I_N} = \left\{\mathbb{E}_{(w,U)\sim\mathcal{P}} \prod_{t=1}^N \frac{\lambda_{i_t}^{i_{<t}}}{d_{\text{out}}^N}\left(1 + \varepsilon \langle w| \rho_t^{i_{<t}} |w\rangle \left\langle\phi_{i_t}^{i_{<t}}\right| U \left|\phi_{i_t}^{i_{<t}}\right\rangle\right)\right\}_{i_1,\ldots,i_N} \tag{4.294}$$

We can express the KL divergence as follows:

$$\text{KL}\left(\mathbb{P}_{H_0}^{I_1,\ldots,I_N} \middle\| \mathbb{P}_{H_1}^{I_1,\ldots,I_N}\right) \tag{4.295}$$

$$= \sum_{i_1,\ldots,i_N} \prod_{t=1}^N \frac{\lambda_{i_t}^{i_{<t}}}{d_{\text{out}}} \log\left(\frac{\prod_{t=1}^N \frac{\lambda_{i_t}^{i_{<t}}}{d_{\text{out}}}}{\mathbb{E}_{(w,U)\sim\mathcal{P}} \prod_{t=1}^N \frac{\lambda_{i_t}^{i_{<t}}}{d_{\text{out}}^N}\left(1 + \varepsilon \langle w| \rho_t^{i_{<t}} |w\rangle \left\langle\phi_{i_t}^{i_{<t}}\right| U \left|\phi_{i_t}^{i_{<t}}\right\rangle\right)}\right) \tag{4.296}$$

$$= \mathbb{E}_{\leq N}(-\log)\left(\mathbb{E}_{(w,U)\sim\mathcal{P}} \prod_{t=1}^N \left(1 + \varepsilon \langle w| \rho_t^{i_{<t}} |w\rangle \left\langle\phi_{i_t}^{i_{<t}}\right| U \left|\phi_{i_t}^{i_{<t}}\right\rangle\right)\right) \tag{4.297}$$

where we use the notation $\mathbb{E}_{\leq N}(X(i_1,\ldots,i_N)) = \sum_{i_1,\ldots,i_N}\left(\prod_{t=1}^N \frac{\lambda_{i_t}^{i<t}}{d_{\text{out}}}\right) X(i_1,\ldots,i_N)$. The function $(-\log)$ is convex so by Jensen inequality:

$$\mathrm{KL}\left(\mathbb{P}_{H_0}^{I_1,\ldots,I_N}\middle\|\mathbb{P}_{H_1}^{I_1,\ldots,I_N}\right) = \mathbb{E}_{\leq N}(-\log)\left(\mathbb{E}_{(w,U)\sim\mathcal{P}}\prod_{t=1}^N\left(1+\varepsilon\left\langle w\right|\rho_t^{i<t}\left|w\right\rangle\left\langle\phi_{i_t}^{i<t}\right|U\left|\phi_{i_t}^{i<t}\right\rangle\right)\right) \tag{4.298}$$

$$\leq \mathbb{E}_{\leq N}\mathbb{E}_{(w,U)\sim\mathcal{P}}(-\log)\prod_{t=1}^N\left(1+\varepsilon\left\langle w\right|\rho_t^{i<t}\left|w\right\rangle\left\langle\phi_{i_t}^{i<t}\right|U\left|\phi_{i_t}^{i<t}\right\rangle\right) \tag{4.299}$$

$$= \sum_{t=1}^N \mathbb{E}_{\leq N}\mathbb{E}_{(w,U)\sim\mathcal{P}}(-\log)\left(1+\varepsilon\left\langle w\right|\rho_t^{i<t}\left|w\right\rangle\left\langle\phi_{i_t}^{i<t}\right|U\left|\phi_{i_t}^{i<t}\right\rangle\right). \tag{4.300}$$

Using the inequality $(-\log(1+x)) \leq -x+2x^2$ valid for $x \geq -1/2$ and since for $\varepsilon \leq 64$, we have $\varepsilon\left\langle w\right|\rho_t^{i<t}\left|w\right\rangle\left\langle\phi_{i_t}^{i<t}\right|U\left|\phi_{i_t}^{i<t}\right\rangle \geq -1/2$, we can upper bound the previous integrand as follows:

$$(-\log)\left(1+\varepsilon\left\langle w\right|\rho_t^{i<t}\left|w\right\rangle\left\langle\phi_{i_t}^{i<t}\right|U\left|\phi_{i_t}^{i<t}\right\rangle\right) \tag{4.301}$$

$$\leq \varepsilon\left\langle w\right|\rho_t^{i<t}\left|w\right\rangle\left\langle\phi_{i_t}^{i<t}\right|U\left|\phi_{i_t}^{i<t}\right\rangle + 2\varepsilon^2\left\langle w\right|\rho_t^{i<t}\left|w\right\rangle^2\left\langle\phi_{i_t}^{i<t}\right|U\left|\phi_{i_t}^{i<t}\right\rangle^2. \tag{4.302}$$

Observe that the first term vanishes under the expectation:

$$\mathbb{E}_{\leq N}\left(\varepsilon\left\langle w\right|\rho_t^{i<t}\left|w\right\rangle\left\langle\phi_{i_t}^{i<t}\right|U\left|\phi_{i_t}^{i<t}\right\rangle\right) = \mathbb{E}_{\leq t-1}\sum_{i_t}\frac{\lambda_{i_t}^{i<t}}{d_{\text{out}}}\varepsilon\left\langle w\right|\rho_t^{i<t}\left|w\right\rangle\left\langle\phi_{i_t}^{i<t}\right|U\left|\phi_{i_t}^{i<t}\right\rangle \tag{4.303}$$

$$= \mathbb{E}_{\leq t-1}\varepsilon\left\langle w\right|\rho_t^{i<t}\left|w\right\rangle\mathrm{Tr}(U) = 0. \tag{4.304}$$

For the second term, we will instead upper bound its expectation under U. Observe that this term is nonnegative, so we can safely remove the condition on the event $\mathcal{G}$ and then we compute the expectation under Haar distributed vector w and Gaussians $\{U_{i,j}\}$ similar to Equation (4.283).

$$\mathbb{E}\left(2\varepsilon^2\left\langle w\right|\rho_t^{i<t}\left|w\right\rangle^2\left\langle\phi_{i_t}^{i<t}\right|U\left|\phi_{i_t}^{i<t}\right\rangle^2\right) = 2\varepsilon^2\mathbb{E}\left(\left\langle w\right|\rho_t^{i<t}\left|w\right\rangle^2\right)\mathbb{E}\left(\left\langle\phi_{i_t}^{i<t}\right|U\left|\phi_{i_t}^{i<t}\right\rangle^2\right) \tag{4.305}$$

$$\leq 2\varepsilon^2\cdot\left(\frac{\mathrm{Tr}(\rho_t)^2+\mathrm{Tr}(\rho_t^2)}{d_{\text{in}}(d_{\text{in}}+1)}\right)\cdot\left(\frac{64}{d_{\text{out}}}\right) \leq \frac{256\varepsilon^2}{d_{\text{in}}^2 d_{\text{out}}}. \tag{4.306}$$

Therefore

$$\mathrm{KL}\left(\mathbb{P}_{H_0}^{I_1,\ldots,I_N}\middle\|\mathbb{P}_{H_1}^{I_1,\ldots,I_N}\right) \leq \sum_{t=1}^N \mathbb{E}_{\leq N}\mathbb{E}_{(w,U)\sim\mathcal{P}}(-\log)\left(1+\varepsilon\left\langle w\right|\rho_t^{i<t}\left|w\right\rangle\left\langle\phi_{i_t}^{i<t}\right|U\left|\phi_{i_t}^{i<t}\right\rangle\right) \tag{4.307}$$

$$\leq \sum_{t=1}^N \mathbb{E}_{\leq N}\frac{256\varepsilon^2}{d_{\text{in}}^2 d_{\text{out}}} = \frac{256N\varepsilon^2}{d_{\text{in}}^2 d_{\text{out}}}. \tag{4.308}$$

On the other hand, the Data-Processing inequality applied on the KL divergence writes:

$$\mathrm{KL}\left(\mathbb{P}_{H_0}^{I_1,\dots,I_N}\middle\|\mathbb{P}_{H_1}^{I_1,\dots,I_N}\right) \geq \mathrm{KL}(\mathbb{P}_{H_0}(\mathcal{A}=0)\,\|\mathbb{P}_{H_1}(\mathcal{A}=0)) \tag{4.309}$$

$$\geq \mathrm{KL}(2/3\|1/3) = \frac{2}{3}\log(2) - \frac{1}{3}\log(2) = \frac{1}{3}\log(2). \tag{4.310}$$

Grouping the lower and upper bounds on the KL divergence:

$$\frac{256N\varepsilon^2}{d_{\mathrm{in}}^2 d_{\mathrm{out}}} \geq \mathrm{KL}\left(\mathbb{P}_{H_0}^{I_1,\dots,I_N}\middle\|\mathbb{P}_{H_1}^{I_1,\dots,I_N}\right) \geq \frac{1}{3}\log(2) \tag{4.311}$$

which yields the lower bound:

$$N = \Omega\left(\frac{d_{\mathrm{in}}^2 d_{\mathrm{out}}}{\varepsilon^2}\right). \tag{4.312}$$

□

4.5 Conclusion and open problems

We have generalized the problem of testing identity to quantum channels. We have in particular identified the optimal complexity $\Theta(d/\varepsilon^2)$ for testing identity to a unitary channel in the adaptive setting. Moreover, we have shown that the complexity for testing identity to the depolarizing channel in the non-adaptive setting is $\tilde{\Theta}(d_{\mathrm{in}}^2 d_{\mathrm{out}}^{1.5}/\varepsilon^2)$. These results open up several interesting questions: can the gap between non-adaptive and adaptive strategies for certification of the depolarizing channel be closed? How to achieve the instance optimality (as in [VV16; CLO22])? This would allow to adapt the complexity to the tested process $\mathcal{N}_0$. Another interesting issue deals with the fact that 4-designs can replace Haar distributed unitaries in Algorithm 8. But can the same complexity be achieved for 3 (and lower) designs? Finally, it would be interesting to consider general strategies allowing entanglement between the uses of the channel, as was done for states in [OW15].

Chapter 5

Sample-Optimal Quantum Process Tomography with Non-Adaptive Incoherent Strategies

5.1 Introduction

In this chapter, we consider the problem of quantum process tomography which consists of approximating an arbitrary quantum channel–any linear map that preserves the axioms of quantum mechanics. This task is an important tool in quantum information processing and quantum control which has been performed in actual experiments (see e.g. [OPGJLRW04; BAHL+10; YW10]). Given a quantum channel $\mathcal{N} : \mathbb{C}^{d_{\text{in}} \times d_{\text{in}}} \to \mathbb{C}^{d_{\text{out}} \times d_{\text{out}}}$ as a black box, a learner could choose the input state and send it through the unknown quantum channel. Then, it can only extract classical information by performing a measurement on the output state. It repeats this procedure at different steps. After collecting a sufficient amount of classical data, the goal is to return a quantum channel $\tilde{\mathcal{N}}$ satisfying:

$$\forall \rho \in \mathbb{C}^{d_{\text{in}} \times d_{\text{in}}} \otimes \mathbb{C}^{d_{\text{in}} \times d_{\text{in}}} : \|\text{id} \otimes (\mathcal{N} - \tilde{\mathcal{N}})(\rho)\|_1 \leq \varepsilon \|\rho\|_1 \tag{5.1}$$

with high probability. In this chapter, we investigate the optimal complexity of non-adaptive strategies using incoherent strategies. These strategies can only use one copy of the unknown channel at each step and must specify the input states and measurement devices before starting the learning procedure.

Contribution The main contribution of this chapter is to show that the optimal complexity of the quantum process tomography with non-adaptive incoherent strategies is $\tilde{\Theta}(d_{\text{in}}^3 d_{\text{out}}^3/\varepsilon^2)$. First, we prove a general lower bound of $\Omega(d_{\text{in}}^3 d_{\text{out}}^3/\varepsilon^2)$ on the number of incoherent measurements for every non-adaptive process learning algorithm. To do so, we construct an $\Omega(\varepsilon)$-separated family of quantum channels close to the completely depolarizing channel of cardinal $M = \exp(\Omega(d_{\text{in}}^2 d_{\text{out}}^2))$ by choosing random Choi states of a specific form. This family is used to encode a message from $\{1, \ldots, M\}$. A process tomography algorithm can be used to decode this message with the same error probability. Hence, the encoder and decoder should share at least $\Omega(d_{\text{in}}^2 d_{\text{out}}^2)$ nats of information. On the other hand, we show that the correlation between the encoder and decoder can only increase by at most $\mathcal{O}(\varepsilon^2/d_{\text{in}} d_{\text{out}})$ nats after each measurement. Note that the naive upper bound on this correlation is $\mathcal{O}(\varepsilon^2)$, we obtain an improvement by a factor $d_{\text{in}} d_{\text{out}}$ by exploiting the randomness in the construction of the quantum channel. This result is stated in Theorem 5.3.1. Next, we show that the process tomography algorithm of

[SSKKG22] can be generalized to approximate an unknown quantum channel to within ε in the diamond norm (5.1) using a number of incoherent measurements $\tilde{\mathcal{O}}(d_{\text{in}}^3 d_{\text{out}}^3/\varepsilon^2)$ (Theorem 5.4.2). For this, we relate the diamond norm between two quantum channels and the operator norm between their corresponding Choi states which improves on the usual inequality with the 1-norm: $\|\mathcal{M}\|_\diamond \leq d_{\text{in}}\|\mathcal{J}_\mathcal{M}\|_1$ (see e.g. [JP16]).

Related work The first works on process tomography including [CN97; PCZ97] follow the strategy of learning the quantum states images of a complete set of basis states then obtaining the quantum channel by an inversion. The problem of state tomography using incoherent measurements is fully understood even for adaptive strategies [HHJWY16; GKKT20; LN22; CHLLS22]: the optimal complexity is $\Theta(d^3/\varepsilon^2)$. So, learning a quantum channel can be done using $\mathcal{O}(d_{\text{in}}^2 d_{\text{out}}^3)$ measurements, but this complexity does not take into account the accumulation of errors. The same drawback can be seen in the resource analysis of different strategies by [MRL08]. Another reductive approach is to use the Choi–Jamiołkowski isomorphism [Cho75; Jam72] to reduce the process tomography to state tomography with a higher dimension [Leu00; DP01]. However, this requires an ancilla and only implies a sub-optimal upper bound $\mathcal{O}((d_{\text{in}}d_{\text{out}})^3/(\varepsilon/d_{\text{in}})^2) = \mathcal{O}(d_{\text{in}}^5 d_{\text{out}}^3/\varepsilon^2)$ for learning in the diamond norm.
For low Kraus rank, [KKEG19] solve quantum process tomography using compressed sensing-based methods. Moreover, [SSKKG22] propose an algorithm for estimating the Choi state in the 2-norm that requires only $\tilde{\mathcal{O}}(d^4/\varepsilon^2)$ ancilla-free incoherent measurements (when $d_{\text{in}} = d_{\text{out}} = d$). This chapter generalizes this result to the diamond norm and general input/output dimensions and shows that this algorithm is optimal up to a logarithmic factor.
A special case of quantum process tomography is learning Pauli channels. These channels have weighted Pauli matrices as Kraus operators and can be learned in diamond norm using $\tilde{\mathcal{O}}(d^3/\varepsilon^2)$ measurements [FW20] (here $d_{\text{in}} = d_{\text{out}} = d$). Furthermore, it is shown that $\Omega(d^3/\varepsilon^2)$ are necessary for any non-adaptive strategy (Chapter 6). While the techniques of the lower bound of this chapter are similar to the ones in Chapter 6, we obtain here a larger lower bound because, in general, we are not restricted to weighted Pauli matrices in the Kraus operators and these latter are implicitly chosen at random.

5.2 Preliminaries

We consider the problem of learning quantum channels of input dimension d_{in} and output dimension d_{out} in diamond distance. The diamond distance can be thought of as a worst-case distance, while the average case distance is given by the Hilbert-Schmidt or Schatten 2-norm between the corresponding Choi states. However, to have comparable distances, we will normalize the 2-norm which is equivalent to unnormalizing the maximally entangled state and we define the 2-distance between two quantum channels $\mathcal{N}$ and $\mathcal{M}$ as follows:

$$d_2(\mathcal{N},\mathcal{M}) := d_{\text{in}}\|\mathcal{J}_\mathcal{N} - \mathcal{J}_\mathcal{M}\|_2 = \|\text{id}\otimes(\mathcal{N}-\mathcal{M})(d_{\text{in}}\,|\Psi\rangle\langle\Psi|)\|_2. \tag{5.2}$$

This is a valid distance since the map $\mathcal{J} : \mathcal{N} \mapsto \text{id}\otimes\mathcal{N}(|\Psi\rangle\langle\Psi|)$ is an isomorphism [Cho75; Jam72].
The channel tomography problem consists of learning a quantum channel $\mathcal{N}$ in the diamond distance. Given a precision parameter $\varepsilon > 0$, the goal is to construct a quantum

channel $\tilde{\mathcal{N}}$ satisfying with at least a probability 2/3:

$$d_\diamond(\mathcal{N}, \tilde{\mathcal{N}}) \leq \varepsilon. \tag{5.3}$$

An algorithm $\mathcal{A}$ is 1/3-correct for this problem if it outputs a quantum channel ε-close to $\mathcal{N}$ with a probability of error at most 1/3. We choose to learn in the diamond distance because it characterizes the minimal error probability to distinguish between two quantum channels when auxiliary systems are allowed [Wat18].
The learner can only extract classical information from the unknown quantum channel $\mathcal{N}$ by performing a measurement on the output state. Throughout this chapter, we only consider unentangled or incoherent measurements. That is, the learner can only use d_{in} (or $d_{\text{anc}} \times d_{\text{in}}$) input states and measure with a d_{out} (or $d_{\text{anc}} \times d_{\text{out}}$)-dimensional measurement devices. We refer to Section 1.4.3 for the definition of different settings of learning quantum channels.
Note that ancilla-assisted strategies were proven to provide an exponential (in the number of qubits $n = \log_2(d)$) advantage over ancilla-free strategies for some problems [CZSJ22; CCHL22]. However, in this chapter, we show that ancilla-assisted strategies cannot overcome ancilla-free strategies for process tomography. Finally, we only consider non-adaptive strategies: the input states and measurement devices should be chosen before starting the learning procedure and thus cannot depend on the observations.

5.3 Lower bound

In this section, we investigate the intrinsic limitations of learning quantum channels using incoherent measurements. To avoid repetition, we consider only ancilla-assisted strategies since they contain ancilla-free strategies as a special case: one can map every d_{in}-dimensional input state ρ to the $d \times d_{\text{in}}$-dimensional input state $\tilde{\rho} = \frac{\mathbb{I}}{d} \otimes \rho$ and every d_{out}-dimensional POVM $\mathcal{M} = \{M_x\}_{x \in \mathcal{X}}$ to the $d \times d_{\text{out}}$-dimensional POVM $\tilde{\mathcal{M}} = \{\mathbb{I}_d \otimes M_x\}_{x \in \mathcal{X}}$.
Mainly, we prove the following theorem:

Theorem 5.3.1. *Let $\varepsilon \leq 1/16$ and $d_{\text{out}} \geq 8$. Any non-adaptive ancilla-assisted algorithm for process tomography in diamond distance requires*

$$N = \Omega\left(\frac{d_{\text{in}}^3 d_{\text{out}}^3}{\varepsilon^2}\right) \tag{5.4}$$

incoherent measurements.

Proof. For the proof, we use the construction of the Choi state:

$$\mathcal{J}_U = \frac{\mathbb{I}}{d_{\text{in}} d_{\text{out}}} + \frac{\varepsilon}{d_{\text{in}} d_{\text{out}}}(U + U^\dagger) - \frac{\varepsilon}{d_{\text{in}} d_{\text{out}}} \text{Tr}_2(U + U^\dagger) \otimes \frac{\mathbb{I}}{d_{\text{out}}} \tag{5.5}$$

where $U \sim \text{Haar}(d_{\text{in}} d_{\text{out}})$. $\mathcal{J}_U$ is Hermitian and satisfies $\text{Tr}_2(\mathcal{J}) = \frac{\mathbb{I}}{d_{\text{in}}}$. Moreover, $\mathcal{J}_U \succcurlyeq 0$ for $\varepsilon \leq 1/4$. Indeed, U is a unitary so it has an operator norm 1 thus $\|U + U^\dagger\|_\infty \leq 2$. Besides, $\|\text{Tr}_2(U + U^\dagger) \otimes \frac{\mathbb{I}}{d_{\text{out}}}\|_\infty = \frac{1}{d_{\text{out}}} \|\text{Tr}_2(U + U^\dagger)\|_\infty \leq \max_i \|\mathbb{I} \otimes \langle i| (U + U^\dagger) \mathbb{I} \otimes |i\rangle\|_\infty \leq 2$.
We claim that:

Lemma 5.3.1. *We can construct an $\varepsilon/2$-separated (according to the diamond distance) family $\{\mathcal{N}_x\}_{x \in [M]}$ of cardinal $M = \exp(\Omega(d_{\text{in}}^2 d_{\text{out}}^2))$.*

Proof. It is sufficient to show that for $U, V \sim \mathrm{Haar}(d_{\mathrm{in}} d_{\mathrm{out}})$:

$$\mathbb{P}\left(\|\mathcal{J}_U - \mathcal{J}_V\|_1 \leq \varepsilon/2\right) \leq \exp\left(-\Omega(d_{\mathrm{in}}^2 d_{\mathrm{out}}^2)\right). \tag{5.6}$$

because, once this concentration inequality holds, we can choose our family randomly, and by the union bound, it will be $\varepsilon/2$-separated with an overwhelming probability $(1 - \exp\left(-\Omega(d_{\mathrm{in}}^2 d_{\mathrm{out}}^2)\right))$ using the inequality $d_\diamond(\mathcal{N}_U, \mathcal{N}_V) \geq \|\mathcal{J}_U - \mathcal{J}_V\|_1$. First, let us lower bound the expected value.

$$\mathbb{E}\left(\|\mathcal{J}_U - \mathcal{J}_V\|_1\right) \geq \frac{\varepsilon}{d_{\mathrm{in}} d_{\mathrm{out}}} \mathbb{E}\left(\|U + U^\dagger - V - V^\dagger\|_1\right) \tag{5.7}$$

$$- \frac{\varepsilon}{d_{\mathrm{in}} d_{\mathrm{out}}^2} \mathbb{E}\left(\|\mathrm{Tr}_2(U + U^\dagger - V - V^\dagger) \otimes \mathbb{I}\|_1\right). \tag{5.8}$$

On one hand, we can upper bound the second expectation using the triangle and the Cauchy-Schwarz inequalities:

$$\mathbb{E}\left(\|\mathrm{Tr}_2(U + U^\dagger - V - V^\dagger) \otimes \mathbb{I}\|_1\right) \leq 4\mathbb{E}\left(\|\mathrm{Tr}_2(U) \otimes \mathbb{I}\|_1\right) \tag{5.9}$$

$$\leq 4\sqrt{d_{\mathrm{in}} d_{\mathrm{out}}}\, \mathbb{E}\left(\|\mathrm{Tr}_2(U) \otimes \mathbb{I}\|_2\right) \leq 4\sqrt{d_{\mathrm{in}} d_{\mathrm{out}}} \sqrt{\mathbb{E}\left(\mathrm{Tr}(\mathrm{Tr}_2(U)\mathrm{Tr}_2(U^\dagger) \otimes \mathbb{I})\right)} \tag{5.10}$$

$$= 4\sqrt{d_{\mathrm{in}} d_{\mathrm{out}}} \sqrt{d_{\mathrm{out}}} \sqrt{\mathbb{E}\left(\sum_i \sum_{k,l} \langle i| \otimes \langle k| U\mathbb{I} \otimes |k\rangle \langle l| U^\dagger |i\rangle \otimes |l\rangle\right)} \tag{5.11}$$

$$= 4\sqrt{d_{\mathrm{in}} d_{\mathrm{out}}} \sqrt{d_{\mathrm{out}}} \sqrt{\mathbb{E}\left(\sum_{i=1}^{d_{\mathrm{in}}} \sum_{k,l=1}^{d_{\mathrm{out}}} \frac{d_{\mathrm{in}} \delta_{k,l}}{d_{\mathrm{in}} d_{\mathrm{out}}}\right)} = 4 d_{\mathrm{in}} d_{\mathrm{out}}. \tag{5.12}$$

On the other hand, we can lower bound the first expectation using Hölder's inequality.

$$\mathbb{E}\left(\|U + U^\dagger - V - V^\dagger\|_1\right) \geq \sqrt{\frac{\left(\mathbb{E}\left(\mathrm{Tr}(U + U^\dagger - V - V^\dagger)^2\right)\right)^3}{\mathbb{E}\left(\mathrm{Tr}(U + U^\dagger - V - V^\dagger)^4\right)}} \tag{5.13}$$

$$\geq \sqrt{\frac{(4 d_{\mathrm{in}} d_{\mathrm{out}})^3}{28 d_{\mathrm{in}} d_{\mathrm{out}}}} \geq \frac{3}{2} d_{\mathrm{in}} d_{\mathrm{out}}. \tag{5.14}$$

Therefore:

$$\mathbb{E}\left(\|\mathcal{J}_U - \mathcal{J}_V\|_1\right) \geq \frac{\varepsilon}{d_{\mathrm{in}} d_{\mathrm{out}}} \mathbb{E}\left(\|U + U^\dagger - V - V^\dagger\|_1\right) - \frac{4\varepsilon}{d_{\mathrm{in}} d_{\mathrm{out}}^2} \mathbb{E}\left(\|\mathrm{Tr}_2 U \otimes \mathbb{I}\|_1\right) \tag{5.15}$$

$$\geq \frac{3}{2}\varepsilon - \frac{4\varepsilon}{d_{\mathrm{out}}} \geq \varepsilon \qquad \text{for} \qquad d_{\mathrm{out}} \geq 8. \tag{5.16}$$

Now, we claim that the function $(U, V) \mapsto \|\mathcal{J}_U - \mathcal{J}_V\|_1$ is $\frac{8\varepsilon}{\sqrt{d_{\mathrm{in}} d_{\mathrm{out}}}}$-Lipschitz. Indeed, we have $\|\mathrm{Tr}_2(X) \otimes \mathbb{I}\|_1 \leq \sqrt{d_{\mathrm{in}} d_{\mathrm{out}}} \|\mathrm{Tr}_2(X) \otimes \mathbb{I}\|_2 = \sqrt{d_{\mathrm{in}}} d_{\mathrm{out}} \|\mathrm{Tr}_2(X)\|_2 \leq \sqrt{d_{\mathrm{in}} d_{\mathrm{out}}} d_{\mathrm{out}} \|X\|_2$ where the last inequality can be found in [LZK08]. Therefore, by letting $X = U - U'$ and

$Y = V - V'$ and using the triangle inequality we obtain:

$$
\begin{aligned}
&|\|\mathcal{J}_U - \mathcal{J}_V\|_1 - \|\mathcal{J}_{U'} - \mathcal{J}_{V'}\|_1| && (5.17)\\
&\leq \frac{2\varepsilon}{d_{\text{in}} d_{\text{out}}} \left[\|X\|_1 + \|Y\|_1 + \left\| \mathrm{Tr}_2(X) \otimes \frac{\mathbb{I}}{d_{\text{out}}} \right\|_1 + \left\| \mathrm{Tr}_2(Y) \otimes \frac{\mathbb{I}}{d_{\text{out}}} \right\|_1 \right] && (5.18)\\
&\leq \frac{2\sqrt{d_{\text{in}} d_{\text{out}}}\varepsilon}{d_{\text{in}} d_{\text{out}}} \left(\|U - U'\|_2 + \|V - V'\|_2 \right) + \frac{2\sqrt{d_{\text{in}} d_{\text{out}}} d_{\text{out}}\varepsilon}{d_{\text{in}} d_{\text{out}}^2} \left(\|U - U'\|_2 + \|V - V'\|_2 \right) && (5.19)\\
&\leq \frac{8\varepsilon}{\sqrt{d_{\text{in}} d_{\text{out}}}} \|(U, V) - (U', V')\|_2 && \text{(Cauchy-Schwarz)} \\
& && (5.20)
\end{aligned}
$$

so by the concentration inequality for Lipschitz functions of Haar measure [MM13]:

$$
\begin{aligned}
&\mathbb{P}\left(\|\mathcal{J}_U - \mathcal{J}_V\|_1 \leq \varepsilon/2 \right) && (5.21)\\
&\leq \mathbb{P}\left(\|\mathcal{J}_U - \mathcal{J}_V\|_1 - \mathbb{E}\left(\|\mathcal{J}_U - \mathcal{J}_V\|_1 \right) \leq -\varepsilon/2 \right) && (5.22)\\
&\leq \exp\left(- \frac{d_{\text{in}} d_{\text{out}} \varepsilon^2}{48 \times 64 \varepsilon^2 / d_{\text{in}} d_{\text{out}}} \right) = \exp\left(-\Omega(d_{\text{in}}^2 d_{\text{out}}^2) \right). && (5.23)
\end{aligned}
$$

□

Now we follow a standard strategy for proving lower bounds of learning problems (see e.g., [FGLE12; HHJWY16]). We use this $\varepsilon/2$-separated family of quantum channels $\{\mathcal{N}_x\}_{x \in [M]}$ (corresponding to the Choi states $\{\mathcal{J}_x\}_{x \in [M]}$ found in Lemma 5.3.1 and $M = \exp(\Omega(d_{\text{in}}^2 d_{\text{out}}^2))$) to encode a uniformly random message $X \sim \text{Uniform}([M])$ by the map $X \mapsto \mathcal{N}_X$. Using a learning algorithm for process tomography with precision $\varepsilon/4$ and an error probability at most $1/3$, a decoder Y can find X with the same error probability because the family $\{\mathcal{N}_x\}_{x \in [M]}$ is $\varepsilon/2$-separated. By Fano's inequality, the encoder and decoder should share at least $\Omega(\log(M))$ nats of information.

Lemma 5.3.2. *[Fan61] We have*

$$\mathcal{I}(X : Y) \geq 2/3 \log(M) - \log(2) \geq \Omega(d_{\text{in}}^2 d_{\text{out}}^2). \tag{5.24}$$

The remaining part of the proof is to upper bound this mutual information in terms of the number of measurements N, the dimensions $d_{\text{in}}, d_{\text{out}}$, and the precision parameter ε. Intuitively, the mutual information, after a few measurements, is very small and then it increases when the number of measurements increases. To make this intuition formal, let N be a number of measurements sufficient for process tomography and let $(I_1, \ldots, I_N)$ be the observations of the learning algorithm, we apply first the data processing inequality to relate the mutual information between the encoder and the decoder with the mutual information between the uniform random variable X and the observations $(I_1, \ldots, I_N)$:

$$\mathcal{I}(X : Y) \leq \mathcal{I}(X : I_1, \ldots, I_N). \tag{5.25}$$

Then we apply the chain rule for the mutual information:

$$\mathcal{I}(X : I_1, \ldots, I_N) = \sum_{t=1}^{N} \mathcal{I}(X : I_t | I_{\leq t-1}) \tag{5.26}$$

where we use the notation $I_{\leq t} = (I_1, \ldots, I_t)$ and $\mathcal{I}(X : I_t|I_{\leq t-1})$ is the conditional mutual information between X and I_t given $I_{\leq t-1}$. A learning algorithm $\mathcal{A}$ would choose the input states $\{\rho_t\}_{t\in[N]}$ and measurement devices $\{\mathcal{M}_t\}_{t\in[N]}$ which can be chosen to have the form $\mathcal{M}_t = \{\mu_i^t |\phi_i^t\rangle\langle\phi_i^t|\}_{i\in\mathcal{I}_t}$ where $\mu_i^t \geq 0$ and $\langle\phi_i^t|\phi_i^t\rangle = 1$ for all t, i. Using Jensen's inequality, we can prove the following upper bound on the conditional mutual information:

Lemma 5.3.3. *For $x \in [M]$, let $\mathcal{M}_x = \mathcal{N}_x - \mathcal{D}$ where $\mathcal{D}(\rho) = \mathrm{Tr}(\rho)\frac{\mathbb{I}}{d_{\text{out}}}$ is the completely depolarizing channel. We have for all $t \in \{1, \ldots, N\}$:*

$$\mathcal{I}(X : I_t|I_{\leq t-1}) \leq \frac{3}{M} \sum_{i\in\mathcal{I}_t, x\in[M]} \mu_i^t \left\langle \phi_i^t \right| \mathrm{id} \otimes \mathcal{D}(\rho_t) \left| \phi_i^t \right\rangle \left(\frac{\langle \phi_i^t| \,\mathrm{id} \otimes \mathcal{M}_x(\rho_t) \,|\phi_i^t\rangle}{\langle \phi_i^t| \,\mathrm{id} \otimes \mathcal{D}(\rho_t) \,|\phi_i^t\rangle} \right)^2 \tag{5.27}$$

Proof. Let $t \in \{1, \ldots, N\}$ and $x \in [M]$. Let $i = (i_1, \ldots, i_t) \in (\mathcal{I}_1, \ldots, \mathcal{I}_t)$, we can express the joint probability p of $(X, I_1, \ldots, I_t)$ as follows:

$$p(x, i_1, \ldots, i_t) = \frac{1}{M} \prod_{k=1}^{t} \mu_{i_k}^k \left\langle \phi_{i_k}^k \right| \mathrm{id} \otimes \mathcal{N}_x(\rho_k) \left| \phi_{i_k}^k \right\rangle \tag{5.28}$$

We can remark that, for all $1 \leq k \leq t$:

$$\begin{aligned} p(x, i_{\leq k}) &= \mu_{i_k}^k \left\langle \phi_{i_k}^k \right| \mathrm{id} \otimes \mathcal{N}_x(\rho_k) \left| \phi_{i_k}^k \right\rangle p(x, i_{\leq k-1}) && (5.29)\\ &= \mu_{i_k}^k \left\langle \phi_{i_k}^k \right| \mathrm{id} \otimes \mathcal{D}(\rho_k) \left| \phi_{i_k}^k \right\rangle (1 + \Phi_{x,i_k}^k) p(x, i_{\leq k-1}) && (5.30) \end{aligned}$$

where $\Phi_{x,i_k}^k = \frac{\left\langle \phi_{i_k}^k \right| \mathrm{id}\otimes\mathcal{M}_x(\rho_k) \left| \phi_{i_k}^k \right\rangle}{\left\langle \phi_{i_k}^k \right| \mathrm{id}\otimes\mathcal{D}(\rho_k) \left| \phi_{i_k}^k \right\rangle}$ because $\mathcal{D} + \mathcal{M}_x = \mathcal{N}_x$. So, the ratio of conditional probabilities can be written as:

$$\begin{aligned} \frac{p(x, i_t|i_{\leq t-1})}{p(x|i_{\leq t-1})p(i_t|i_{\leq t-1})} &= \frac{p(x, i_{\leq t})p(i_{\leq t-1})}{p(x, i_{\leq t-1})p(i_{\leq t})} && (5.31)\\ &= \frac{\mu_{i_t}^t \left\langle \phi_{i_t}^t \right| \mathrm{id} \otimes \mathcal{D}(\rho_t) \left| \phi_{i_t}^t \right\rangle (1 + \Phi_{x,i_t}^t) p(x, i_{\leq t-1}) p(i_{\leq t-1})}{p(x, i_{\leq t-1}) \sum_y p(y, i_{\leq t})} && (5.32)\\ &= \frac{\mu_{i_t}^t \left\langle \phi_{i_t}^t \right| \mathrm{id} \otimes \mathcal{D}(\rho_t) \left| \phi_{i_t}^t \right\rangle (1 + \Phi_{x,i_t}^t) p(i_{\leq t-1})}{\sum_y p(y, i_{\leq t})} && (5.33)\\ &= \frac{\mu_{i_t}^t \left\langle \phi_{i_t}^t \right| \mathrm{id} \otimes \mathcal{D}(\rho_t) \left| \phi_{i_t}^t \right\rangle (1 + \Phi_{x,i_t}^t) p(i_{\leq t-1})}{\sum_y \mu_{i_t}^t \left\langle \phi_{i_t}^t \right| \mathrm{id} \otimes \mathcal{D}(\rho_t) \left| \phi_{i_t}^t \right\rangle (1 + \Phi_{y,i_t}^t) p(y, i_{\leq t-1})} && (5.34)\\ &= \frac{(1 + \Phi_{x,i_t}^t) p(i_{\leq t-1})}{\sum_y (1 + \Phi_{y,i_t}^t) p(y, i_{\leq t-1})} = \frac{(1 + \Phi_{x,i_t}^t)}{\sum_y (1 + \Phi_{y,i_t}^t) p(y|i_{\leq t-1})} && (5.35) \end{aligned}$$

Therefore by Jensen's inequality:

$$\begin{aligned} \mathcal{I}(X : I_t|I_{\leq t-1}) &= \mathbb{E}\left(\log\left(\frac{p(x, i_t|i_{\leq t-1})}{p(x|i_{\leq t-1})p(i_t|i_{\leq t-1})} \right) \right) && (5.36)\\ &= \mathbb{E}\left(\log\left(\frac{(1 + \Phi_{x,i_t}^t)}{\sum_y p(y|i_{\leq t-1})(1 + \Phi_{y,i_t}^t)} \right) \right) && (5.37)\\ &\leq \mathbb{E}\left(\log(1 + \Phi_{x,i_t}^t) - \sum_y p(y|i_{\leq t-1}) \log(1 + \Phi_{y,i_t}^t) \right) && (5.38)\\ &= \mathbb{E}\left(\log(1 + \Phi_{x,i_t}^t) \right) - \sum_y \mathbb{E}\left(p(y|i_{\leq t-1}) \log(1 + \Phi_{y,i_t}^t) \right). && (5.39) \end{aligned}$$

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

$$\mathbb{E}\left(\log(1+\Phi^t_{x,i_t})\right) = \mathbb{E}_{x,i\sim p} \log(1+\Phi^t_{x,i_t}) \tag{5.40}$$
$$\leq \mathbb{E}_{x,i\sim p}\Phi^t_{x,i_t} = \mathbb{E}_{x,i\sim p_{\leq t}}\Phi^t_{x,i_t} \tag{5.41}$$
$$= \mathbb{E}_{x,i\sim p_{\leq t-1}} \sum_{i_t} \mu^t_{i_t} \left\langle \phi^t_{i_t}\right| \mathrm{id}\otimes\mathcal{D}(\rho_t) \left|\phi^t_{i_t}\right\rangle (1+\Phi^t_{x,i_t})\Phi^t_{x,i_t} \tag{5.42}$$
$$= \mathbb{E}_{x,i\sim p_{\leq t-1}} \sum_{i_t} \mu^t_{i_t} \left\langle \phi^t_{i_t}\right| \mathrm{id}\otimes\mathcal{D}(\rho_t) \left|\phi^t_{i_t}\right\rangle (\Phi^t_{x,i_t})^2 \tag{5.43}$$
$$= \frac{1}{M}\sum_{x=1}^{M}\sum_{i_t} \mu^t_{i_t} \left\langle \phi^t_{i_t}\right| \mathrm{id}\otimes\mathcal{D}(\rho_t) \left|\phi^t_{i_t}\right\rangle (\Phi^t_{x,i_t})^2 \tag{5.44}$$

because $\sum_{i_t} \mu^t_{i_t} \left\langle \phi^t_{i_t}\right| \mathrm{id}\otimes\mathcal{D}(\rho_t) \left|\phi^t_{i_t}\right\rangle \Phi^t_{x,i_t} = \mathrm{Tr}(\mathrm{id}\otimes\mathcal{M}_x(\rho_t)) = \mathrm{Tr}(\mathrm{id}\otimes\mathcal{N}_x(\rho_t)) - \mathrm{Tr}(\mathrm{id}\otimes\mathcal{D}(\rho_t)) = \mathrm{Tr}(\rho_t) - \mathrm{Tr}(\rho_t) = 0$ and we use the condition that the algorithm is non-adaptive in the last line.
On the other hand, the second term can be upper bounded using the inequality $-\log(1+x) \leq -x + x^2$ verified for all $x \in (-1/2, \infty)$. Let $\lambda^t_{i_t} = \mu^t_{i_t} \left\langle \phi^t_{i_t}\right| \mathrm{id}\otimes\mathcal{D}(\rho_t) \left|\phi^t_{i_t}\right\rangle$, we have :

$$\mathbb{E}\left(-\sum_y p(y|i_{\leq t-1}) \log(1+\Phi^t_{y,i_t})\right) \tag{5.45}$$
$$= \sum_y \mathbb{E}_{x,i\sim p} p(y|i_{\leq t-1})(-\log)(1+\Phi^t_{y,i_t}) \tag{5.46}$$
$$= \sum_y \mathbb{E}_{x,i\sim p_{\leq t}} p(y|i_{\leq t-1})(-\log)(1+\Phi^t_{y,i_t}) \tag{5.47}$$
$$\leq \sum_y \mathbb{E}_{x,i\sim p_{\leq t}} p(y|i_{\leq t-1})(-\Phi^t_{y,i_t} + (\Phi^t_{y,i_t})^2) \tag{5.48}$$
$$\leq \sum_y \mathbb{E}_{x,i\sim p_{\leq t-1}} p(y|i_{\leq t-1}) \sum_{i_t} \lambda^t_{i_t}((2\Phi^t_{x,i_t})^2 + 2(\Phi^t_{y,i_t})^2) \tag{5.49}$$
$$= 4\sum_y \mathbb{E}_{x,i\sim p_{\leq t-1}} p(y|i_{\leq t-1}) \sum_{i_t} \lambda^t_{i_t}(\Phi^t_{x,i_t})^2 \tag{5.50}$$
$$= 4\mathbb{E}_{x,i\sim p_{\leq t-1}} \sum_{i_t} \mu^t_{i_t} \left\langle \phi^t_{i_t}\right| \mathrm{id}\otimes\mathcal{D}(\rho_t) \left|\phi^t_{i_t}\right\rangle (\Phi^t_{x,i_t})^2 \tag{5.51}$$
$$= 4\mathbb{E}_{x,i\sim p_{\leq t-1}} \sum_{i_t} \lambda^t_{i_t}(\Phi^t_{x,i_t})^2 \tag{5.52}$$
$$= \frac{4}{M}\sum_{x=1}^{M}\sum_{i_t} \mu^t_{i_t} \left\langle \phi^t_{i_t}\right| \mathrm{id}\otimes\mathcal{D}(\rho_t) \left|\phi^t_{i_t}\right\rangle (\Phi^t_{x,i_t})^2 \tag{5.53}$$

where we use the condition that the algorithm is non-adaptive in the last line. Since the conditional mutual information is upper bounded by the sum of these two terms, the upper bound on the conditional mutual information follows. □

It remains to approximate every mean $\frac{1}{M}\sum_{x=1}^{M}$ by the expectation $\mathbb{E}_U$.

Lemma 5.3.4. *We have with at least a probability* 9/10*:*

$$\frac{1}{M}\sum_{t,i,x}\mu_i^t \left\langle\phi_i^t\right| \mathrm{id}\otimes\mathcal{D}(\rho_t)\left|\phi_i^t\right\rangle \left(\frac{\left\langle\phi_i^t\right| \mathrm{id}\otimes\mathcal{M}_x(\rho_t)\left|\phi_i^t\right\rangle}{\left\langle\phi_i^t\right| \mathrm{id}\otimes\mathcal{D}(\rho_t)\left|\phi_i^t\right\rangle}\right)^2 \tag{5.54}$$

$$\leq \sum_{t,i}\mu_i^t \left\langle\phi_i^t\right| \mathrm{id}\otimes\mathcal{D}(\rho_t)\left|\phi_i^t\right\rangle \mathbb{E}_U \left(\frac{\left\langle\phi_i^t\right| \mathrm{id}\otimes\mathcal{M}_U(\rho_t)\left|\phi_i^t\right\rangle}{\left\langle\phi_i^t\right| \mathrm{id}\otimes\mathcal{D}(\rho_t)\left|\phi_i^t\right\rangle}\right)^2 + 16N\varepsilon^2\sqrt{\frac{\log(10)}{M}}. \tag{5.55}$$

Proof. Denote by f_x^t the function $|\phi\rangle \mapsto \frac{\langle\phi|\mathrm{id}\otimes\mathcal{M}_x(\rho_t)|\phi\rangle^2}{\langle\phi|\mathrm{id}\otimes\mathcal{D}(\rho_t)|\phi\rangle^2}$. We claim that the functions f_x^t are bounded. Indeed, we can write $\rho_t = \sum_i \lambda_i |\psi_i\rangle\langle\psi_i|$ and every $|\psi_i\rangle$ can be written as $|\psi_i\rangle = A_i \otimes \mathbb{I}\,|\Psi\rangle$ so for a unit vector $|\phi\rangle$, we have:

$$f_x^t(|\phi\rangle) = \frac{\langle\phi| \mathrm{id}\otimes\mathcal{M}_x(\rho_t)\left|\phi\right\rangle^2}{\langle\phi| \mathrm{id}\otimes\mathcal{D}(\rho_t)\left|\phi\right\rangle^2} \tag{5.56}$$

$$= \frac{4\varepsilon^2\left(\langle\phi|\sum_i \lambda_i(A_i\otimes\mathbb{I})\left(U_x - \mathrm{Tr}_2(U_x)\otimes\frac{\mathbb{I}}{d_{\mathrm{out}}}\right)(A_i^\dagger\otimes\mathbb{I})\left|\phi\right\rangle\right)^2}{\langle\phi|\sum_i \lambda_i A_i A_i^\dagger\otimes\mathbb{I}\left|\phi\right\rangle^2} \tag{5.57}$$

$$\leq \frac{4\varepsilon^2\langle\phi|\sum_i \lambda_i(A_i\otimes\mathbb{I})(A_i^\dagger\otimes\mathbb{I})\left|\phi\right\rangle^2 \left\|U_x - \mathrm{Tr}_2(U_x)\otimes\frac{\mathbb{I}}{d_{\mathrm{out}}}\right\|_\infty^2}{\langle\phi|\sum_i \lambda_i A_i A_i^\dagger\otimes\mathbb{I}\left|\phi\right\rangle^2} \tag{5.58}$$

$$\leq \frac{16\varepsilon^2\langle\phi|\sum_i \lambda_i(A_i\otimes\mathbb{I})(A_i^\dagger\otimes\mathbb{I})\left|\phi\right\rangle^2}{\langle\phi|\sum_i \lambda_i A_i A_i^\dagger\otimes\mathbb{I}\left|\phi\right\rangle^2} = 16\varepsilon^2 \tag{5.59}$$

where we used that $\|U_x\|_\infty = 1$ and $\|\mathrm{Tr}_2(U_x)\|_\infty \leq d_{\mathrm{out}}$ for a unitary U_x. But we have $\sum_i \mu_i^t \left\langle\phi_i^t\right| \mathrm{id}\otimes\mathcal{D}(\rho_t)\left|\phi_i^t\right\rangle = \mathrm{Tr}(\mathrm{id}\otimes\mathcal{D}(\rho_t)) = 1$ so for all $x \in [M]$:

$$\sum_{t,i}\mu_i^t \left\langle\phi_i^t\right| \mathrm{id}\otimes\mathcal{D}(\rho_t)\left|\phi_i^t\right\rangle \left(\frac{\left\langle\phi_i^t\right| \mathrm{id}\otimes\mathcal{M}_x(\rho_t)\left|\phi_i^t\right\rangle}{\left\langle\phi_i^t\right| \mathrm{id}\otimes\mathcal{D}(\rho_t)\left|\phi_i^t\right\rangle}\right)^2 \leq 16N\varepsilon^2. \tag{5.60}$$

Therefore, by Hoeffding's inequality [Hoe63] and the union bound, we have with a probability at least 9/10:

$$\frac{1}{M}\sum_{x,t,i}\mu_i^t \left\langle\phi_i^t\right| \mathrm{id}\otimes\mathcal{D}(\rho_t)\left|\phi_i^t\right\rangle \left(\frac{\left\langle\phi_i^t\right| \mathrm{id}\otimes\mathcal{M}_x(\rho_t)\left|\phi_i^t\right\rangle}{\left\langle\phi_i^t\right| \mathrm{id}\otimes\mathcal{D}(\rho_t)\left|\phi_i^t\right\rangle}\right)^2 \tag{5.61}$$

$$\leq \sum_{t,i}\mu_i^t \left\langle\phi_i^t\right| \mathrm{id}\otimes\mathcal{D}(\rho_t)\left|\phi_i^t\right\rangle \mathbb{E}_U \left(\frac{\left\langle\phi_i^t\right| \mathrm{id}\otimes\mathcal{M}_U(\rho_t)\left|\phi_i^t\right\rangle}{\left\langle\phi_i^t\right| \mathrm{id}\otimes\mathcal{D}(\rho_t)\left|\phi_i^t\right\rangle}\right)^2 + 16N\varepsilon^2\sqrt{\frac{\log(10)}{M}}. \tag{5.62}$$

□

These two Lemmas 5.3.3, 5.3.4 imply:

$$\begin{aligned}
\mathcal{I}(X : I_1, \ldots, I_N) &= \sum_{t=1}^{N} \mathcal{I}(X : I_t | I_{\leq t-1}) \\
&\leq \frac{5}{M} \sum_{x,t,i} \mu_i^t \left\langle \phi_i^t \right| \mathrm{id} \otimes \mathcal{D}(\rho_t) \left| \phi_i^t \right\rangle \mathbb{E} \left(\frac{\left\langle \phi_i^t \right| \mathrm{id} \otimes \mathcal{M}_x(\rho_t) \left| \phi_i^t \right\rangle}{\left\langle \phi_i^t \right| \mathrm{id} \otimes \mathcal{D}(\rho_t) \left| \phi_i^t \right\rangle} \right)^2 \\
&\leq 5 \sum_{t,i} \mu_i^t \left\langle \phi_i^t \right| \mathrm{id} \otimes \mathcal{D}(\rho_t) \left| \phi_i^t \right\rangle \mathbb{E} \left(\frac{\left\langle \phi_i^t \right| \mathrm{id} \otimes \mathcal{M}_U(\rho_t) \left| \phi_i^t \right\rangle}{\left\langle \phi_i^t \right| \mathrm{id} \otimes \mathcal{D}(\rho_t) \left| \phi_i^t \right\rangle} \right)^2 + 80 N \varepsilon^2 \sqrt{\frac{\log(10)}{M}} \\
&\leq 5N \sup_{t,i} \mathbb{E} \left(\left(\frac{\left\langle \phi_i^t \right| \mathrm{id} \otimes \mathcal{M}_U(\rho_t) \left| \phi_i^t \right\rangle}{\left\langle \phi_i^t \right| \mathrm{id} \otimes \mathcal{D}(\rho_t) \left| \phi_i^t \right\rangle} \right)^2 \right) + 80 N \varepsilon^2 \sqrt{\frac{\log(10)}{M}} \qquad (5.63)
\end{aligned}$$

where we used that fact that for all $t \in [N]$: $\sum_i \mu_i^t \left\langle \phi_i^t \right| \mathrm{id} \otimes \mathcal{D}(\rho_t) \left| \phi_i^t \right\rangle = \mathrm{Tr}(\mathrm{id} \otimes \mathcal{D}(\rho_t)) = \mathrm{Tr}(\rho_t) = 1$. The error probability $1/10$ of this approximation can be absorbed in the construction above by asking the unitaries $\{U_x\}_{x \in [M]}$ not only to satisfy the separability condition, but also to satisfy the inequalities in Lemma 5.3.4:

$$\frac{1}{M} \sum_{t,i,x} \mu_i^t \left\langle \phi_i^t \right| \mathrm{id} \otimes \mathcal{D}(\rho_t) \left| \phi_i^t \right\rangle \left(\frac{\left\langle \phi_i^t \right| \mathrm{id} \otimes \mathcal{M}_x(\rho_t) \left| \phi_i^t \right\rangle}{\left\langle \phi_i^t \right| \mathrm{id} \otimes \mathcal{D}(\rho_t) \left| \phi_i^t \right\rangle} \right)^2 \qquad (5.64)$$

$$\leq \sum_{t,i} \mu_i^t \left\langle \phi_i^t \right| \mathrm{id} \otimes \mathcal{D}(\rho_t) \left| \phi_i^t \right\rangle \mathbb{E}_U \left(\frac{\left\langle \phi_i^t \right| \mathrm{id} \otimes \mathcal{M}_U(\rho_t) \left| \phi_i^t \right\rangle}{\left\langle \phi_i^t \right| \mathrm{id} \otimes \mathcal{D}(\rho_t) \left| \phi_i^t \right\rangle} \right)^2 + 16 N \varepsilon^2 \sqrt{\frac{\log(10)}{M}}. \qquad (5.65)$$

Now fix $t \in [N], i_t \in \mathcal{I}_t$ and $|\phi\rangle = \left| \phi_{i_t}^t \right\rangle$. Recall that we can write $\rho_t = \sum_i \lambda_i |\psi_i\rangle\langle\psi_i|$, the maximally entangled state is denoted $|\Psi\rangle = \frac{1}{\sqrt{d_{\mathrm{in}}}} \sum_{i=1}^{d_{\mathrm{in}}} |ii\rangle$ and every $|\psi_i\rangle$ can be written as $|\psi_i\rangle = A_i \otimes \mathbb{I} |\Psi\rangle$ so:

$$\begin{aligned}
\mathrm{id} \otimes \mathcal{D}(\rho_t) &= \sum_i \lambda_i (\mathrm{id} \otimes \mathcal{D})(A_i \otimes \mathbb{I} |\Psi\rangle\langle\Psi| A_i^\dagger \otimes \mathbb{I}) \\
&= \sum_i \lambda_i (A_i \otimes \mathbb{I}) \mathrm{id} \otimes \mathcal{D}(|\Psi\rangle\langle\Psi|)(A_i^\dagger \otimes \mathbb{I}) \\
&= \sum_i \lambda_i (A_i \otimes \mathbb{I}) \frac{\mathbb{I}}{d_{\mathrm{in}} d_{\mathrm{out}}} (A_i^\dagger \otimes \mathbb{I}) \\
&= \frac{\sum_i \lambda_i A_i A_i^\dagger}{d_{\mathrm{in}}} \otimes \frac{\mathbb{I}}{d_{\mathrm{out}}}. \qquad (5.66)
\end{aligned}$$

On the other hand, using the notation $V = U - \mathrm{Tr}_2(U) \otimes \frac{\mathbb{I}}{d_{\mathrm{out}}}$, we can write:

$$\begin{aligned}
\mathrm{id} \otimes \mathcal{M}(\rho_t) &= \sum_i \lambda_i \mathrm{id} \otimes \mathcal{M}(A_i \otimes \mathbb{I} |\Psi\rangle\langle\Psi| A_i^\dagger \otimes \mathbb{I}) \qquad (5.67) \\
&= \sum_i \lambda_i (A_i \otimes \mathbb{I}) \mathrm{id} \otimes (\mathcal{N} - \mathcal{D})(|\Psi\rangle\langle\Psi|)(A_i^\dagger \otimes \mathbb{I}) \qquad (5.68) \\
&= \sum_i \lambda_i (A_i \otimes \mathbb{I}) \left(\mathcal{J}_{\mathcal{N}} - \frac{\mathbb{I}}{d_{\mathrm{in}} d_{\mathrm{out}}} \right) (A_i^\dagger \otimes \mathbb{I}) \qquad (5.69) \\
&= \frac{\varepsilon}{d_{\mathrm{in}} d_{\mathrm{out}}} \sum_i \lambda_i A_i \otimes \mathbb{I} \left(U + U^\dagger - \mathrm{Tr}_2(U + U^\dagger) \otimes \frac{\mathbb{I}}{d_{\mathrm{out}}} \right) A_i^\dagger \otimes \mathbb{I}) \qquad (5.70) \\
&= \frac{\varepsilon}{d_{\mathrm{in}} d_{\mathrm{out}}} \sum_i \lambda_i \left[(A_i \otimes \mathbb{I}) V (A_i^\dagger \otimes \mathbb{I}) + (A_i \otimes \mathbb{I}) V^\dagger (A_i^\dagger \otimes \mathbb{I}) \right]. \qquad (5.71)
\end{aligned}$$

By Equation (5.63), we need to control the expectation $\mathbb{E}_U \langle\phi| \operatorname{id} \otimes \mathcal{M}_U(\rho_t) |\phi\rangle^2$. First, we replace $\operatorname{id} \otimes \mathcal{M}(\rho_t)$ with the latter expression, then we apply the inequality $(x+y)^2 \leq 2x^2 + 2y^2$ to separate the terms involving U and the terms involving $\operatorname{Tr}_2(U)$. The first term can be computed and bounded as follows.

$$\begin{aligned}
&\frac{4\varepsilon^2}{d_{\text{in}}^2 d_{\text{out}}^2} \mathbb{E} \left(\left| \langle\phi| \left(\sum_i \lambda_i (A_i \otimes \mathbb{I}) U (A_i^\dagger \otimes \mathbb{I}) \right) |\phi\rangle \right|^2 \right) \\
&= \frac{4\varepsilon^2}{d_{\text{in}}^2 d_{\text{out}}^2} \sum_{i,j} \frac{\lambda_i \lambda_j}{d_{\text{in}} d_{\text{out}}} \left| \operatorname{Tr} \left(A_i^\dagger \otimes \mathbb{I} |\phi\rangle \langle\phi| A_j \otimes \mathbb{I} \right) \right|^2 \\
&\overset{\text{(CS)}}{\leq} \frac{4\varepsilon^2}{d_{\text{in}}^2 d_{\text{out}}^2} \sum_{i,j} \frac{\lambda_i \lambda_j}{d_{\text{in}} d_{\text{out}}} \langle\phi| A_i A_i^\dagger \otimes \mathbb{I} |\phi\rangle \langle\phi| A_j A_j^\dagger \otimes \mathbb{I} |\phi\rangle \\
&= \frac{4\varepsilon^2}{d_{\text{in}}^3 d_{\text{out}}^3} \left(\langle\phi| \sum_i \lambda_i A_i A_i^\dagger \otimes \mathbb{I} |\phi\rangle \right)^2 . \qquad (5.72)
\end{aligned}$$

Let's move to the second term which involves the partial trace. Let $M_{ij} = (A_i^\dagger \otimes \mathbb{I}) |\phi\rangle\langle\phi| (A_j \otimes \mathbb{I})$.

$$\frac{4\varepsilon^2}{d_{\text{in}}^2 d_{\text{out}}^2} \mathbb{E} \left(\left| \langle\phi| \sum_i \lambda_i (A_i \otimes \mathbb{I}) \left(\operatorname{Tr}_2(U) \otimes \frac{\mathbb{I}}{d_{\text{out}}} \right) (A_i^\dagger \otimes \mathbb{I}) |\phi\rangle \right|^2 \right) \qquad (5.73)$$

$$= \frac{4\varepsilon^2}{d_{\text{in}}^2 d_{\text{out}}^4} \sum_{i,j} \lambda_i \lambda_j \mathbb{E} \left(\operatorname{Tr} \left[(\operatorname{Tr}_2(U) \otimes \mathbb{I}) M_{i,j} \left(\operatorname{Tr}_2(U^\dagger) \otimes \mathbb{I} \right) M_{ij}^\dagger \right] \right) \qquad (5.74)$$

$$= \frac{4\varepsilon^2}{d_{\text{in}}^2 d_{\text{out}}^4} \sum_{i,j} \lambda_i \lambda_j \sum_{x,y,z,t=1}^{d_{\text{in}}} \sum_{k,l=1}^{d_{\text{out}}} \mathbb{E} \left(\langle xk| U |yk\rangle \langle zl| U^\dagger |tl\rangle \right) \operatorname{Tr} \left[\left(|y\rangle \langle x| \otimes \mathbb{I} M_{ij} |t\rangle \langle z| \otimes \mathbb{I} M_{ij}^\dagger \right) \right] \qquad (5.75)$$

$$= \frac{4\varepsilon^2}{d_{\text{in}}^3 d_{\text{out}}^5} \sum_{i,j} \lambda_i \lambda_j \sum_{x=t,y=z=1}^{d_{\text{in}}} \sum_{k=l=1}^{d_{\text{out}}} \operatorname{Tr} \left[\left(|y\rangle \langle x| \otimes \mathbb{I} M_{ij} |x\rangle \langle y| \otimes \mathbb{I} M_{ij}^\dagger \right) \right] \qquad (5.76)$$

$$= \frac{4\varepsilon^2}{d_{\text{in}}^3 d_{\text{out}}^4} \sum_{i,j} \lambda_i \lambda_j \sum_{x,y=1}^{d_{\text{in}}} \operatorname{Tr} \left[\left(|y\rangle \langle x| \otimes \mathbb{I} M_{ij} |x\rangle \langle y| \otimes \mathbb{I} M_{ij}^\dagger \right) \right]. \qquad (5.77)$$

To control the latter expression, we write $|\phi\rangle = B^\dagger \otimes \mathbb{I} |\Psi\rangle$ so that $M_{i,j} = (A_i^\dagger \otimes \mathbb{I}) |\phi\rangle\langle\phi| (A_j \otimes \mathbb{I}) = (A_i^\dagger B^\dagger \otimes \mathbb{I}) |\Psi\rangle\langle\Psi| (B A_j \otimes \mathbb{I})$. Using the property of the maximally

entangled state $\langle\Psi| M \otimes \mathbb{I} |\Psi\rangle = \frac{1}{d_{\text{in}}}\text{Tr}(M)$ we obtain:

$$\sum_{x,y=1}^{d_{\text{in}}} \text{Tr}\Big[\Big(|y\rangle \langle x| \otimes \mathbb{I} M_{ij} |x\rangle \langle y| \otimes \mathbb{I} M_{ij}^\dagger \Big) \tag{5.78}$$

$$= \sum_{x,y=1}^{d_{\text{in}}} \text{Tr}\left(|y\rangle \langle x| \otimes \mathbb{I}(A_i^\dagger B^\dagger \otimes \mathbb{I}) |\Psi\rangle\langle\Psi| (BA_j \otimes \mathbb{I}) |x\rangle \langle y| \otimes \mathbb{I}(A_j^\dagger B^\dagger \otimes \mathbb{I}) |\Psi\rangle\langle\Psi| (BA_i \otimes \mathbb{I}) \right) \tag{5.79}$$

$$= \sum_{x,y=1}^{d_{\text{in}}} \langle\Psi| (BA_j \otimes \mathbb{I}) |x\rangle \langle y| \otimes \mathbb{I}(A_j^\dagger B^\dagger \otimes \mathbb{I}) |\Psi\rangle \langle\Psi| (BA_i \otimes \mathbb{I})^\dagger |y\rangle \langle x| \otimes \mathbb{I}(A_i^\dagger B^\dagger \otimes \mathbb{I}) |\Psi\rangle \tag{5.80}$$

$$= \frac{1}{d_{\text{in}}^2} \sum_{x,y=1}^{d_{\text{in}}} \text{Tr}(BA_j |x\rangle \langle y| A_j^\dagger B^\dagger) \text{Tr}(BA_i |y\rangle \langle x| A_i^\dagger B^\dagger) \tag{5.81}$$

$$= \frac{1}{d_{\text{in}}^2} \sum_{x,y=1}^{d_{\text{in}}} \langle y| A_j^\dagger B^\dagger BA_j |x\rangle \langle x| A_i^\dagger B^\dagger BA_i |y\rangle = \frac{1}{d_{\text{in}}^2} \text{Tr}\left(A_j^\dagger B^\dagger BA_j A_i^\dagger B^\dagger BA_i \right). \tag{5.82}$$

On the other hand, we can write

$$\langle\phi| \sum_i \lambda_i A_i A_i^\dagger \otimes \mathbb{I} |\phi\rangle = \langle\Psi| \sum_i \lambda_i BA_i A_i^\dagger B^\dagger \otimes \mathbb{I} |\Psi\rangle = \frac{1}{d_{\text{in}}} \text{Tr}\left(\sum_i \lambda_i A_i^\dagger B^\dagger BA_i \right). \tag{5.83}$$

Note that the matrix $\sum_i \lambda_i A_i^\dagger B^\dagger BA_i$ is positive semi-definite so:

$$\begin{aligned} &\sum_{i,j} \lambda_i \lambda_j \frac{1}{d_{\text{in}}^2} \text{Tr}\left(A_j^\dagger B^\dagger BA_j A_i^\dagger B^\dagger BA_i \right) = \frac{1}{d_{\text{in}}^2} \text{Tr}\left(\sum_i \lambda_i A_i^\dagger B^\dagger BA_i \right)^2 \\ &\leq \left[\frac{1}{d_{\text{in}}} \text{Tr}\left(\sum_i \lambda_i A_i^\dagger B^\dagger BA_i \right) \right]^2 = \langle\phi| \sum_i \lambda_i A_i A_i^\dagger \otimes \mathbb{I} |\phi\rangle^2 . \end{aligned} \tag{5.84}$$

Hence

$$\begin{aligned} &\frac{4\varepsilon^2}{d_{\text{in}}^2 d_{\text{out}}^2} \mathbb{E}\left(\left| \langle\phi| \sum_i \lambda_i (A_i \otimes \mathbb{I}) \left(\text{Tr}_2(U) \otimes \frac{\mathbb{I}}{d_{\text{out}}} \right) (A_i^\dagger \otimes \mathbb{I}) |\phi\rangle \right|^2 \right) \\ &= \frac{4\varepsilon^2}{d_{\text{in}}^3 d_{\text{out}}^4} \sum_{i,j} \lambda_i \lambda_j \sum_{x,y=1}^{d_{\text{in}}} \text{Tr}\Big[\Big(|y\rangle \langle x| \otimes \mathbb{I} M_{ij} |x\rangle \langle y| \otimes \mathbb{I} M_{ij}^\dagger \Big) \Big] \\ &= \frac{4\varepsilon^2}{d_{\text{in}}^3 d_{\text{out}}^4} \sum_{i,j} \lambda_i \lambda_j \frac{1}{d_{\text{in}}^2} \text{Tr}\left(A_j^\dagger B^\dagger BA_j A_i^\dagger B^\dagger BA_i \right) \\ &\leq \frac{4\varepsilon^2}{d_{\text{in}}^3 d_{\text{out}}^4} \langle\phi| \sum_i \lambda_i A_i A_i^\dagger \otimes \mathbb{I} |\phi\rangle^2 \end{aligned} \tag{5.85}$$

Using the equality (5.66) and the two inequalities (5.72) and (5.85), we deduce:

$$\mathbb{E}\left(\left(\frac{\langle\phi| \text{id} \otimes \mathcal{M}_U(\rho_t) |\phi\rangle}{\langle\phi| \text{id} \otimes \mathcal{D}(\rho_t) |\phi\rangle} \right)^2 \right) \leq \frac{\frac{8\varepsilon^2}{d_{\text{in}}^3 d_{\text{out}}^3} \left(\langle\phi| \sum_i \lambda_i A_i A_i^\dagger \otimes \mathbb{I} |\phi\rangle \right)^2}{\langle\phi| \frac{\sum_i \lambda_i A_i A_i^\dagger}{d_{\text{in}}} \otimes \frac{\mathbb{I}}{d_{\text{out}}} |\phi\rangle^2} = \frac{8\varepsilon^2}{d_{\text{in}} d_{\text{out}}}. \tag{5.86}$$

Therefore using the inequality (5.63):

$$\mathcal{I}(X : I_1, \dots, I_N) = \sum_{t=1}^{N} \mathcal{I}(X : I_t | I_{\leq t-1}) \tag{5.87}$$

$$\leq 5N \sup_{t, i_t} \mathbb{E} \left(\left(\frac{\left\langle \phi_{i_t}^t \right| \mathrm{id} \otimes \mathcal{M}_U(\rho_t) \left| \phi_{i_t}^t \right\rangle}{\left\langle \phi_{i_t}^t \right| \mathrm{id} \otimes \mathcal{D}(\rho_t) \left| \phi_{i_t}^t \right\rangle} \right)^2 \right) + 80N\varepsilon^2 \sqrt{\frac{\log(10)}{M}} \tag{5.88}$$

$$\leq 40N \frac{\varepsilon^2}{d_{\mathrm{in}} d_{\mathrm{out}}} + 80N\varepsilon^2 \sqrt{\frac{\log(10)}{M}} \leq \mathcal{O}\left(N \frac{\varepsilon^2}{d_{\mathrm{in}} d_{\mathrm{out}}} \right) \tag{5.89}$$

because $M = \exp(\Omega(d_{\mathrm{in}}^2 d_{\mathrm{out}}^2))$. But from the data processing inequality and Lemma 5.3.2, $\mathcal{I}(X : I_1, \dots, I_N) \geq \mathcal{I}(X : Y) \geq \Omega(d_{\mathrm{in}}^2 d_{\mathrm{out}}^2)$, we deduce that:

$$\mathcal{O}\left(N \frac{\varepsilon^2}{d_{\mathrm{in}} d_{\mathrm{out}}} \right) \geq \mathcal{I}(X : I_1, \dots, I_N) \geq \Omega(d_{\mathrm{in}}^2 d_{\mathrm{out}}^2). \tag{5.90}$$

Finally, the lower bound follows:

$$N \geq \Omega\left(\frac{d_{\mathrm{in}}^3 d_{\mathrm{out}}^3}{\varepsilon^2} \right). \tag{5.91}$$

□

To assess this lower bound, it is necessary to design an algorithm for quantum process tomography. This will be the object of the following section.

5.4 Upper bound

In this section, we propose an upper bound on the complexity of the quantum process tomography problem. We generalize the algorithm proposed by [SSKKG22] which is ancilla-free.

Theorem 5.4.1. *[SSKKG22] There is an ancilla-free process tomography algorithm that learns a quantum channel (of $d_{\mathrm{in}} = d_{\mathrm{out}} = d$) in the distance d_2 using only a number of measurements:*

$$N = \mathcal{O}\left(\frac{d^6 \log(d)}{\varepsilon^2} \right). \tag{5.92}$$

This algorithm proceeds by providing an unbiased estimator for the Choi state $\mathcal{J}_{\mathcal{N}}$, then projecting this matrix to the space of Choi states (PSD and partial trace $\mathbb{I}/d$) and finally by invoking the Choi–Jamiołkowski isomorphism we obtain an approximation of the channel. This reduction from learning the Choi state in the operator norm to learning the quantum channel in the d_2 distance uses mainly the inequality $d_2(\mathcal{N}, \mathcal{M}) = d\|\mathcal{J}_{\mathcal{N}} - \mathcal{J}_{\mathcal{M}}\|_2 \leq d^2 \|\mathcal{J}_{\mathcal{N}} - \mathcal{J}_{\mathcal{M}}\|_\infty$ when $d_{\mathrm{in}} = d_{\mathrm{out}} = d$. We generalize this result to the diamond norm and any input/output dimensions. For this we show the following inequality:

Lemma 5.4.1. *Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be two quantum channels. We have:*

$$d_\diamond(\mathcal{N}_1, \mathcal{N}_2) \leq d_{\mathrm{in}} d_{\mathrm{out}} \|\mathcal{J}_{\mathcal{N}_1} - \mathcal{J}_{\mathcal{N}_2}\|_\infty. \tag{5.93}$$

This inequality can also be obtained by applying the inequality (3) of [NPPŻ18] and the triangle inequality. We provide a simpler proof for completeness.

Proof. Denote by $\mathcal{M} = \mathcal{N}_1 - \mathcal{N}_2$. Let $|\phi\rangle$ be a maximizing unit vector of the diamond norm, i.e., $\|\mathrm{id} \otimes \mathcal{M}(|\phi\rangle\langle\phi|)\|_1 = d_\diamond(\mathcal{N}_1, \mathcal{N}_2)$. We can write $|\phi\rangle = A \otimes \mathbb{I} |\Psi\rangle$ where $|\Psi\rangle = \frac{1}{\sqrt{d_{\mathrm{in}}}} \sum_{i=1}^{d_{\mathrm{in}}} |ii\rangle$ is the maximally entangled state. $|\phi\rangle$ has norm 1 so $\frac{1}{d_{\mathrm{in}}}\mathrm{Tr}(A^\dagger A) = \langle\Psi| A^\dagger A \otimes \mathbb{I} |\Psi\rangle = \langle\phi|\phi\rangle = 1$. On the other hand we can write

$$\begin{aligned} d_\diamond(\mathcal{N}_1, \mathcal{N}_2) &= \|\mathrm{id} \otimes \mathcal{M}(|\phi\rangle\langle\phi|)\|_1 && (5.94)\\ &= \|\mathbb{I} \otimes \mathcal{M}(A \otimes \mathbb{I}_{d_{\mathrm{in}}} |\Psi\rangle\langle\Psi| A^\dagger \otimes \mathbb{I}_{d_{\mathrm{in}}})\|_1 && (5.95)\\ &= \|(A \otimes \mathbb{I}_{d_{\mathrm{out}}})\mathrm{id} \otimes \mathcal{M}(|\Psi\rangle\langle\Psi|)(A^\dagger \otimes \mathbb{I}_{d_{\mathrm{out}}})\|_1 && (5.96)\\ &= \|(A \otimes \mathbb{I}_{d_{\mathrm{out}}})\mathcal{J}_{\mathcal{M}}(A^\dagger \otimes \mathbb{I}_{d_{\mathrm{out}}})\|_1. && (5.97) \end{aligned}$$

$\mathcal{J}_{\mathcal{M}}$ is Hermitian so it can be written as : $\mathcal{J}_{\mathcal{M}} = \sum_i \lambda_i |\psi_i\rangle\langle\psi_i|$. Using the triangle inequality, we obtain:

$$\begin{aligned} &\|(A \otimes \mathbb{I}_{d_{\mathrm{out}}})\mathcal{J}_{\mathcal{M}}(A^\dagger \otimes \mathbb{I}_{d_{\mathrm{out}}})\|_1 && (5.98)\\ &= \left\| (A \otimes \mathbb{I}_{d_{\mathrm{out}}}) \sum_i \lambda_i |\psi_i\rangle\langle\psi_i| (A^\dagger \otimes \mathbb{I}_{d_{\mathrm{out}}}) \right\|_1 && (5.99)\\ &\leq \sum_i |\lambda_i| \|(A \otimes \mathbb{I}_{d_{\mathrm{out}}}) |\psi_i\rangle\langle\psi_i| (A^\dagger \otimes \mathbb{I}_{d_{\mathrm{out}}})\|_1 && (5.100)\\ &\leq \max_i |\lambda_i| \sum_i \|(A \otimes \mathbb{I}_{d_{\mathrm{out}}}) |\psi_i\rangle\langle\psi_i| (A^\dagger \otimes \mathbb{I}_{d_{\mathrm{out}}})\|_1 && (5.101)\\ &= \|\mathcal{J}\|_\infty \sum_i \mathrm{Tr}((A \otimes \mathbb{I}_{d_{\mathrm{out}}}) |\psi_i\rangle\langle\psi_i| (A^\dagger \otimes \mathbb{I}_{d_{\mathrm{out}}})) && (5.102)\\ &= \|\mathcal{J}\|_\infty \mathrm{Tr}(AA^\dagger \otimes \mathbb{I}_{d_{\mathrm{out}}}) = d_{\mathrm{in}} d_{\mathrm{out}} \|\mathcal{J}\|_\infty. && (5.103) \end{aligned}$$

□

This Lemma shows that the diamond and 2 distances satisfy the same inequality with respect to the infinity norm between the Choi states when $d_{\mathrm{in}} = d_{\mathrm{out}} = d$. Since the algorithm of [SSKKG22] approximates first the Choi state in the infinity norm, we obtain the same upper bound for the diamond distance. For general dimensions, we obtain the following complexity:

Theorem 5.4.2. *There is a non-adaptive ancilla-free process tomography algorithm that learns a quantum channel in the distance $d_\diamond$ using only a number of measurements:*

$$N = \mathcal{O}\left(\frac{d_{\mathrm{in}}^3 d_{\mathrm{out}}^3 \log(d_{\mathrm{in}} d_{\mathrm{out}})}{\varepsilon^2}\right). \tag{5.104}$$

This complexity was expected for process tomography with incoherent measurements since the complexity of state tomography with incoherent measurements is $\Theta\left(\frac{d^3}{\varepsilon^2}\right)$ [HHJWY16] and learning $(d_{\mathrm{in}}, d_{\mathrm{out}})$-dimensional channels can be thought of as learning states of dimension $d_{\mathrm{in}} \times d_{\mathrm{out}}$. We believe that the $\log(d_{\mathrm{in}} d_{\mathrm{out}})$-factor can be removed from the upper bound in Theorem 5.4.2 using the techniques of [GKKT20]. The algorithm is formally described in Algorithm 9 and is similar to the one in [SSKKG22]. By Theorem 5.3.1, Algorithm 9 is almost optimal.

Its analysis is also similar to the one in [SSKKG22].

Algorithm 9 Learning a quantum channel in the diamond distance using ancilla-free independent measurements.

$N = \mathcal{O}(d_{\text{in}}^3 d_{\text{out}}^3 \log(d_{\text{in}} d_{\text{out}})/\varepsilon^2)$.
for $t = 1 : N$ **do**
Sample two independent copies of Haar distributed unitaries $V \sim \text{Haar}(d_{\text{in}})$ and $U \sim \text{Haar}(d_{\text{out}})$.
Let $|v\rangle = V|0\rangle$ be a Haar distributed vector.
Take the input states $\rho_t = |v\rangle\langle v|$ and $\sigma_t = \frac{\mathbb{I}}{d_{\text{in}}}$, the output states are respectively $\mathcal{N}(|v\rangle\langle v|)$ and $\mathcal{N}\left(\frac{\mathbb{I}}{d_{\text{in}}}\right)$.
Perform a measurement on $\mathcal{N}(|v\rangle\langle v|)$ and $\mathcal{N}\left(\frac{\mathbb{I}}{d_{\text{in}}}\right)$ using the POVM $\mathcal{M}_U := \{U|i\rangle\langle i|U^\dagger\}_{i\in[d_{\text{out}}]}$ and observe $i_t \sim p_{U,V} := \{\langle i|U^\dagger \mathcal{N}(|v\rangle\langle v|)U|i\rangle\}_{i\in[d_{\text{out}}]}$ and $j_t \sim q_U := \{\langle i|U^\dagger \mathcal{N}\left(\frac{\mathbb{I}}{d}\right)U|i\rangle\}_{i\in[d_{\text{out}}]}$.
Define $\mathcal{J}_t := (d_{\text{in}}+1)|v\rangle\langle v|^\top \otimes ((d_{\text{out}}+1)(U|i_t\rangle\langle i_t|U^\dagger) - \mathbb{I}) - \mathbb{I}\otimes((d_{\text{out}}+1)(U|j_t\rangle\langle j_t|U^\dagger) - \mathbb{I})$
end for
Define the estimator $\hat{\mathcal{J}} = \frac{1}{N}\sum_{t=1}^N \mathcal{J}_t$.
Find a valid Choi state $\mathcal{J}_{\mathcal{M}}$ such that $\|\mathcal{J}_{\mathcal{M}} - \hat{\mathcal{J}}\|_\infty \leq \frac{\varepsilon}{2d_{\text{in}}d_{\text{out}}}$.
return the quantum channel $\mathcal{M}$ corresponding to the Choi state $\mathcal{J}_{\mathcal{M}}$.

Correctness Let us prove that Algorithm 9 is 1/3-correct. First we show that $\hat{\mathcal{J}} = \frac{1}{N}\sum_{t=1}^N \mathcal{J}_t$ is an unbiased estimator of $\mathcal{J}_{\mathcal{N}}$. For this, we prove the following lemma relating the Choi state to the average of the tensor product of a random rank-1 projector and its image by the quantum channel.

Lemma 5.4.2. *Let $|\phi\rangle$ be a* Haar*-distributed random vector. We have the following equality:*

$$\mathcal{J}_{\mathcal{N}} = (d_{\text{in}}+1)\mathbb{E}\left(|\phi\rangle\langle\phi|^\top \otimes \mathcal{N}(|\phi\rangle\langle\phi|)\right) - \mathbb{I}\otimes\mathcal{N}\left(\frac{\mathbb{I}}{d_{\text{in}}}\right). \tag{5.105}$$

Proof. We use the Kraus decomposition of the quantum channel $\mathcal{N}(\rho) = \sum_k A_k \rho A_K^\dagger$. We start by writing the following expectation:

$$\mathbb{E}\left(|\phi\rangle\langle\phi| \otimes \mathcal{N}(|\phi\rangle\langle\phi|)\right) = \sum_k \mathbb{E}\left(|\phi\rangle\langle\phi| \otimes A_k|\phi\rangle\langle\phi|A_k^\dagger\right) \tag{5.106}$$

$$= \sum_k \mathbb{I}\otimes A_k \mathbb{E}\left(|\phi\rangle\langle\phi|\otimes|\phi\rangle\langle\phi|\right)\mathbb{I}\otimes A_k^\dagger \tag{5.107}$$

Let F be the flip operator $F = \sum_{i,j=1}^{d_{\text{in}}} |ij\rangle\langle ji|$, if we take the transpose on the fist tensor we obtain the unnormalized maximally entangled state:

$$F^{T_1} = \sum_{i,j=1}^{d_{\text{in}}} |i\rangle\langle j|^\top \otimes |j\rangle\langle i| = \sum_{i,j=1}^{d_{\text{in}}} |j\rangle\langle i| \otimes |j\rangle\langle i| = d_{\text{in}}|\Psi\rangle\langle\Psi| \tag{5.108}$$

where $|\Psi\rangle\langle\Psi| = \frac{1}{d_{\text{in}}}\sum_{i,j=1}^{d_{\text{in}}} |ii\rangle\langle jj| = \frac{1}{d_{\text{in}}}\sum_{i,j=1}^{d_{\text{in}}} |i\rangle\langle j|\otimes|i\rangle\langle j|$ is the maximally entangled state. It is known that there is constants α and β such that:

$$\mathbb{E}\left(|\phi\rangle\langle\phi|\otimes|\phi\rangle\langle\phi|\right) = \alpha\mathbb{I} + \beta F. \tag{5.109}$$

Taking the trace we have the first relation $1 = \alpha d_{\text{in}}^2 + \beta d_{\text{in}}$, then taking the trace after multiplying with F we obtain the second relation $\text{Tr}(|\phi\rangle\langle\phi| \otimes |\phi\rangle\langle\phi| F) = \text{Tr}(|\phi\rangle\langle\phi| |\phi\rangle\langle\phi|) = 1 = \alpha d_{\text{in}} + \beta d_{\text{in}}^2$. These relations imply $\alpha = \beta = \frac{1}{d_{\text{in}}(d_{\text{in}}+1)}$. Hence:

$$\mathbb{E}\left(|\phi\rangle\langle\phi| \otimes |\phi\rangle\langle\phi|\right) = \frac{\mathbb{I} + F}{d_{\text{in}}(d_{\text{in}}+1)}. \tag{5.110}$$

Replacing this expectation on the first expectation yields:

$$\mathbb{E}\left(|\phi\rangle\langle\phi|^\top \otimes \mathcal{N}(|\phi\rangle\langle\phi|)\right) = \sum_k \mathbb{E}\left(|\phi\rangle\langle\phi|^\top \otimes A_k |\phi\rangle\langle\phi| A_k^\dagger\right) \tag{5.111}$$

$$= \sum_k \mathbb{I} \otimes A_k \mathbb{E}\left(|\phi\rangle\langle\phi|^\top \otimes |\phi\rangle\langle\phi|\right) \mathbb{I} \otimes A_k^\dagger \tag{5.112}$$

$$= \frac{1}{d_{\text{in}}(d_{\text{in}}+1)} \sum_k \mathbb{I} \otimes A_k A_k^\dagger + \frac{1}{d_{\text{in}}(d_{\text{in}}+1)} \sum_k \mathbb{I} \otimes A_k (d_{\text{in}} |\Psi\rangle\langle\Psi|) \mathbb{I} \otimes A_k^\dagger \tag{5.113}$$

$$= \frac{1}{d_{\text{in}}(d_{\text{in}}+1)} \mathbb{I} \otimes \mathcal{N}(\mathbb{I}) + \frac{1}{d_{\text{in}}+1} \mathbb{I} \otimes \mathcal{N}(|\Psi\rangle\langle\Psi|) \tag{5.114}$$

$$= \frac{1}{d_{\text{in}}(d_{\text{in}}+1)} \mathbb{I} \otimes \mathcal{N}(\mathbb{I}) + \frac{1}{d_{\text{in}}+1} \mathcal{J}_{\mathcal{N}}. \tag{5.115}$$

□

Then we compute another expectation:

Lemma 5.4.3. *Let $U \sim \text{Haar}(d)$ and $x \sim p_{U,\rho} := \{\langle i| U^\dagger \rho U |i\rangle\}_{i\in[d]}$, we have*

$$\mathbb{E}\left((d+1) U |x\rangle\langle x| U^\dagger - \mathbb{I}\right) = \rho \tag{5.116}$$

Proof. Since the equality is linear in ρ we can without loss of generality restrict ourselves to a pure state $\rho = |\phi\rangle\langle\phi|$. Now $x \sim \{\langle i| U^\dagger |\phi\rangle\langle\phi| U |i\rangle\}_{i\in[d]}$ hence for $k, l \in [d]$, by Weingarten calculus:

$$\mathbb{E}_{U, x\sim p_{U,\phi}} \left(\langle k| U |x\rangle\langle x| U^\dagger |l\rangle\right) \tag{5.117}$$

$$= \mathbb{E}_U \left(\sum_{x=1}^d \langle x| U^\dagger |\phi\rangle\langle\phi| U |x\rangle \langle k| U |x\rangle\langle x| U^\dagger |l\rangle\right) \tag{5.118}$$

$$= \mathbb{E}_U \left(\sum_{x=1}^d \langle x| U^\dagger |\phi\rangle\langle\phi| U |x\rangle \langle x| U^\dagger |l\rangle \langle k| U |x\rangle\right) \tag{5.119}$$

$$= \sum_{x=1}^d \frac{1}{d(d+1)} \left(\delta_{l,k} + \langle\phi|l\rangle\langle k|\phi\rangle\right) \tag{5.120}$$

$$= \frac{1}{(d+1)} \left(\langle k| \mathbb{I} + |\phi\rangle\langle\phi| |l\rangle\right)) \tag{5.121}$$

Therefore

$$\mathbb{E}\left((d+1) U |x\rangle\langle x| U^\dagger - \mathbb{I}\right) = |\phi\rangle\langle\phi| = \rho. \tag{5.122}$$

□

Using Lemmas 5.4.2 and 5.4.3 we deduce:

$$\mathbb{E}\left(\hat{\mathcal{J}}\right) = \mathbb{E}\left(\mathcal{J}_1\right) \tag{5.123}$$

$$= \mathbb{E}_{V,U,i_t,j_t}\left[(d_{\text{in}}+1)\left|v\right\rangle\!\left\langle v\right|^{\top} \otimes ((d_{\text{out}}+1)(U\left|i_t\right\rangle\!\left\langle i_t\right|U^{\dagger}) - \mathbb{I})\right] \tag{5.124}$$

$$- \mathbb{E}_{V,U,i_t,j_t}\left[\mathbb{I} \otimes ((d_{\text{out}}+1)(U\left|j_t\right\rangle\!\left\langle j_t\right|U^{\dagger}) - \mathbb{I})\right] \tag{5.125}$$

$$= \mathbb{E}_{V}\left[(d_{\text{in}}+1)\left|v\right\rangle\!\left\langle v\right|^{\top} \otimes \mathbb{E}_{U,i_t}((d_{\text{out}}+1)(U\left|i_t\right\rangle\!\left\langle i_t\right|U^{\dagger}) - \mathbb{I})\right] \tag{5.126}$$

$$- \left[\mathbb{I} \otimes \mathbb{E}_{U,j_t}((d_{\text{out}}+1)(U\left|j_t\right\rangle\!\left\langle j_t\right|U^{\dagger}) - \mathbb{I})\right] \tag{5.127}$$

$$= \mathbb{E}_{V}\left((d_{\text{in}}+1)\left|v\right\rangle\!\left\langle v\right|^{\top} \otimes \mathcal{N}(\left|v\right\rangle\!\left\langle v\right|) - \mathbb{I} \otimes \mathcal{N}\left(\frac{\mathbb{I}}{d_{\text{in}}}\right)\right) = \mathcal{J}_{\mathcal{N}}. \tag{5.128}$$

So the estimator $\hat{\mathcal{J}} = \frac{1}{N}\sum_{t=1}^{N}\mathcal{J}_t$ is unbiased. It remains to show a concentration inequality for the random variable $\hat{\mathcal{J}}$ so that we can estimate how much steps we need in order to achieve the precision and confidence we aim to. For this, we use the matrix Bernstein inequality [Tro12]:

Theorem 5.4.3. *[Tro12] Consider a sequence of n independent Hermitian random matrices $A_1, \ldots, A_n \in \mathbb{C}^{d\times d}$. Assume that each A_i satisfies*

$$\mathbb{E}\left(A_i\right) = 0 \quad \text{and} \quad \|A_i\|_{\infty} \leq R \text{ as.} \tag{5.129}$$

Let $\sigma^2 = \|\sum_{i=1}^{n}\mathbb{E}\left(A_i^2\right)\|_{\infty}$. Then for any $t \geq \frac{\sigma^2}{R}$:

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n}(A_i - \mathbb{E}\left(A_i\right))\right\|_{\infty} \geq t\right) \leq d\exp\left(-\frac{3t}{8R}\right). \tag{5.130}$$

Moreover for any $t \leq \frac{\sigma^2}{R}$:

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n}(A_i - \mathbb{E}\left(A_i\right))\right\|_{\infty} \geq t\right) \leq d\exp\left(-\frac{3t^2}{8\sigma^2}\right). \tag{5.131}$$

Let $\mathcal{J} = \mathcal{J}_{\mathcal{N}} = \mathbb{E}\left(\mathcal{J}_t\right)$. We apply this theorem to the estimator $\hat{\mathcal{J}} - \mathcal{J} = \frac{1}{N}\sum_{t=1}^{N}(\mathcal{J}_t - \mathcal{J})$. Recall that

$$\mathcal{J}_t = (d_{\text{in}}+1)\left|v\right\rangle\!\left\langle v\right|^{\top} \otimes ((d_{\text{out}}+1)(U\left|i_t\right\rangle\!\left\langle i_t\right|U^{\dagger}) - \mathbb{I}) - \mathbb{I} \otimes ((d_{\text{out}}+1)(U\left|j_t\right\rangle\!\left\langle j_t\right|U^{\dagger}) - \mathbb{I}). \tag{5.132}$$

Let $A_t = \frac{\mathcal{J}_t - \mathcal{J}}{N}$, we have proven that $\mathbb{E}\left(A_t\right) = \frac{1}{N}\mathbb{E}\left(\mathcal{J}_t - \mathcal{J}\right) = 0$. Moreover

$$\|A_t\|_{\infty} = \frac{1}{N}\|\mathcal{J}_t - \mathcal{J}\|_{\infty} \leq \frac{1}{N}(\|\mathcal{J}_t\|_{\infty} + \|\mathcal{J}\|_{\infty}) \leq \frac{8d_{\text{in}}d_{\text{out}}}{N} := R. \tag{5.133}$$

Besides

$$\sigma^2 = \left\|\sum_{t=1}^{N}\mathbb{E}\left(A_t^2\right)\right\|_{\infty} = \frac{1}{N}\left\|\mathbb{E}\left((\mathcal{J}_1 - \mathcal{J})^2\right)\right\|_{\infty} = \frac{1}{N}\left\|\mathbb{E}\left((\mathcal{J}_1)^2\right)\right\|_{\infty} + \Theta\left(\frac{1}{N}\right). \tag{5.134}$$

Using the identity $(a\left|\phi\right\rangle\!\left\langle\phi\right| - \mathbb{I})^2 = (a^2 - 2a)\left|\phi\right\rangle\!\left\langle\phi\right| + \mathbb{I}$, we have:

$$\mathbb{E}\left(\left[\mathbb{I}\otimes((d_{\text{out}}+1)(U\left|j_t\right\rangle\!\left\langle j_t\right|U^\dagger) - \mathbb{I})\right]^2\right) \tag{5.135}$$
$$= \mathbb{E}\left((\mathbb{I}\otimes((d_{\text{out}}^2-1)(U\left|j_t\right\rangle\!\left\langle j_t\right|U^\dagger) + \mathbb{I})\right) \tag{5.136}$$
$$= \mathbb{E}\left((\mathbb{I}\otimes((d_{\text{out}}^2-1)(U\left|j_t\right\rangle\!\left\langle j_t\right|U^\dagger - \mathbb{I}/(d_{\text{out}}+1)) + d_{\text{out}}\ \mathbb{I})\right) \tag{5.137}$$
$$= (d_{\text{out}}-1)\mathbb{I}\otimes\mathcal{N}(\mathbb{I}/d_{\text{in}}) + d_{\text{out}}\ \mathbb{I}\otimes\mathbb{I} \tag{5.138}$$

has an operator norm at most $\mathcal{O}(d_{\text{out}})$ so we can focus on the first term in the definition of $\mathcal{J}_1$ which has the main contribution. We have using again the identity $(a\left|\phi\right\rangle\!\left\langle\phi\right| - \mathbb{I})^2 = (a^2-2a)\left|\phi\right\rangle\!\left\langle\phi\right| + \mathbb{I}$:

$$\mathbb{E}\left[(d_{\text{in}}+1)\left|v\right\rangle\!\left\langle v\right|^\top\otimes((d_{\text{out}}+1)(U\left|i_t\right\rangle\!\left\langle i_t\right|U^\dagger) - \mathbb{I})\right]^2 \tag{5.139}$$
$$= (d_{\text{in}}+1)^2\mathbb{E}\left(\left|v\right\rangle\!\left\langle v\right|^\top\otimes((d_{\text{out}}+1)(U\left|i_t\right\rangle\!\left\langle i_t\right|U^\dagger) - \mathbb{I})^2\right) \tag{5.140}$$
$$= (d_{\text{in}}+1)^2\mathbb{E}\left(\left|v\right\rangle\!\left\langle v\right|^\top\otimes((d_{\text{out}}^2-1)(U\left|i_t\right\rangle\!\left\langle i_t\right|U^\dagger) + \mathbb{I})\right) \tag{5.141}$$
$$= (d_{\text{out}}-1)(d_{\text{in}}+1)(\mathcal{J}+\mathbb{I}\otimes\mathcal{N}(\mathbb{I}/d_{\text{in}})) + \left(\frac{d_{\text{out}}(d_{\text{in}}+1)^2}{d_{\text{in}}}\right)\mathbb{I} \tag{5.142}$$

which has an operator norm $\Theta(d_{\text{in}}d_{\text{out}})$. Therefore

$$\sigma^2 = \frac{1}{N}\left\|\mathbb{E}\left(\mathcal{J}_1^2\right)\right\|_\infty + \Theta\left(\frac{1}{N}\right) = \Theta\left(\frac{d_{\text{in}}d_{\text{out}}}{N}\right). \tag{5.143}$$

Since we have $\frac{\sigma^2}{R} \geq \Omega(1)$ we can use the matrix-Bernstein inequality in the regime $t = \frac{\varepsilon}{2d_{\text{in}}d_{\text{out}}} \leq \mathcal{O}(1)$:

$$\mathbb{P}\left(\left\|\sum_{t=1}^N(A_t - \mathbb{E}\left(A_t\right))\right\|_\infty \geq \frac{\varepsilon}{2d_{\text{in}}d_{\text{out}}}\right) \leq d_{\text{in}}d_{\text{out}}\exp\left(-\frac{3\varepsilon^2}{8d_{\text{in}}^2d_{\text{out}}^2\sigma^2}\right) \tag{5.144}$$
$$\leq d_{\text{in}}d_{\text{out}}\exp\left(-\frac{CN\varepsilon^2}{d_{\text{in}}^3d_{\text{out}}^3}\right) \tag{5.145}$$

where $C > 0$ is a universal constant. Hence if $N = d_{\text{in}}^3d_{\text{out}}^3\log(3d_{\text{in}}d_{\text{out}})/(C\varepsilon^2) = \mathcal{O}\left(d_{\text{in}}^3d_{\text{out}}^3\log(d_{\text{in}}d_{\text{out}})/\varepsilon^2\right)$ then with a probability at least $2/3$ we have

$$\|\hat{\mathcal{J}} - \mathcal{J}_{\mathcal{N}}\|_\infty = \left\|\sum_{t=1}^N(A_t - \mathbb{E}\left(A_t\right))\right\|_\infty \leq \frac{\varepsilon}{2d_{\text{in}}d_{\text{out}}}. \tag{5.146}$$

This implies that $\|\mathcal{J}_{\mathcal{M}} - \mathcal{J}_{\mathcal{N}}\|_\infty \leq \frac{\varepsilon}{d_{\text{in}}d_{\text{out}}}$ and finally $\|\mathcal{M}-\mathcal{N}\|_\diamond \leq \varepsilon$ by Lemma 5.4.1. This finishes the proof of the correctness of Algorithm 9.

5.5 Conclusion and open questions

In this chapter, we find the optimal complexity of quantum process tomography using non-adaptive incoherent measurements. Furthermore, we show that ancilla-assisted strategies cannot outperform their ancilla-free counterparts contrary to Pauli channel tomography [CZSJ22]. Still, many questions remain open. First, it is known that adaptive strategies

have the same complexity as non-adaptive ones for state tomography [CHLLS22], could adaptive strategies overcome non-adaptive ones for quantum process tomography? Secondly, can entangled strategies exploit the symmetry and show a polynomial (in $d_{\text{in}}, d_{\text{out}}$) speedup as they do for state tomography [HHJWY16]? Lastly, what would be the potential improvements for simpler problems such as learning the expectations of some given input states and observables?

Chapter 6

Lower bounds on learning Pauli channels

6.1 Introduction

In spite of their impressive progress over the last few years [AABB+19; ZWDC+20; SSWE+21; EWLK+21], 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 [EHWRMPCK20]. 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 [Wat18, 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 [WE16] 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 [CZSJ22].

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 [FW20; MGE12; FH18; HROWE22; HXVW19]. Finally, Pauli noise model reflects experiments that are actually done in practice (see e.g., [HFW20] 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 6.1

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 channel uses. In particular, this shows that the randomized benchmarking algorithm of [FW20, 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 6.4.1. 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 does not 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 6.3.1. 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 6.5.1. 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. [FW20] 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.

Model	Lower bound	Upper bound
Non-adaptive, ℓ_1-distance	$N = \Omega(d^3/\varepsilon^2)$ or $\sum_{t=1}^{N} m_t = \Omega(d^4/\varepsilon^6)$ [Thm. 6.4.1]	$N = \tilde{\mathcal{O}}(d^3/\varepsilon^2)$ [FW20]
Non-adaptive, ℓ_∞-distance	$N = \Omega(1/\varepsilon^2)$ [FO21]	$N = \tilde{\mathcal{O}}(1/\varepsilon^2)$ [FO21]
Adaptive, ℓ_1-distance	$N = \Omega(d^2/\varepsilon^2)$ [Thm. 6.3.1]	$N = \tilde{\mathcal{O}}(d^3/\varepsilon^2)$ [FW20]
Adaptive, ℓ_1-distance $\varepsilon \leq 1/(20d)$, $\forall t \in [N] : m_t = 1$	$N = \Omega(d^{2.5}/\varepsilon^2)$ [Thm. 6.5.1]	$N = \tilde{\mathcal{O}}(d^3/\varepsilon^2)$ [FW20]

Table 6.1: 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.

For completeness, we reproduce this argument in Section 6.7. In this chapter, we address an open question posed in [FW20] about a lower bound for learning Pauli channels. In particular we show that the algorithm of [FW20] 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 [FO21] and this is optimal up to logarithmic factors. The previous settings did not allow for ancillas. The work of [CZSJ22] 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 do not apply in this setting since we only consider ancilla-free strategies. We also note that [CZSJ22] 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 [BKGNMSM13] and techniques based on compressed sensing [RKKLGEK18]. 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 [MW13] for results on testing quantum channels and to [SSKKG22] and Chapter 5 for quantum channel learning in the non-adaptive setting. It is shown that $\tilde{\Theta}(d^6/\varepsilon^2)$ copies are necessary and sufficient to learn a (d, d)-dimensional quantum channel in the diamond norm. However, if we add the Pauli structure to the channel, our lower bound along with the upper bound of [FW20] show that the optimal copy complexity becomes only $\tilde{\Theta}(d^3/\varepsilon^2)$. On the other hand, the question of optimal quantum channel tomography remains open in the adaptive setting.

6.2 Preliminaries

Let $d = 2^n$ be the dimension of an n-qubit system. Recall that 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 $\sum_k A_k^\dagger A_k = \mathbb{I}$. If the quantum channel $\mathcal{N}$ satisfies further $\mathcal{N}(\mathbb{I}) = \mathbb{I}$, it is called *unital*. Pauli channels are special quantum channels whose Kraus operators are weighted

Pauli operators. Formally, a Pauli quantum channel $\mathcal{P}$ can be written as follows:

$$\mathcal{P}(\rho) = \sum_{P \in \{\mathbb{I},X,Y,Z\}^{\otimes n}} p(P) P\rho P \tag{6.1}$$

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 $2/3$:

$$\|\mathcal{P} - \tilde{\mathcal{P}}\|_\diamond \leq \varepsilon. \tag{6.2}$$

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 [Wat18]. Since the diamond norm between two Pauli channels is exactly twice the TV-distance between their corresponding probability distributions [MGE12], approximating the Pauli channel $\mathcal{P}$ in diamond norm is equivalent to approximating the probability distribution p in TV-distance.

The learner can only extract classical information from the unknown Pauli channel $\mathcal{P}$ by performing a measurement on the output state. Throughout the chapter, 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. We refer to Section 1.4.3 for the definitions of different settings for learning channels.

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 chapter 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 $\mathcal{M} = \{|0\rangle\langle 0|, \mathbb{I} - |0\rangle\langle 0|\}$ 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. However, if we do not allow reusing the channel, then the number of measurements needed is approximately $\log(1/\delta)/\varepsilon$.

6.3 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 $2/3$:

$$\mathrm{TV}(p, \hat{p}) \leq \varepsilon \tag{6.3}$$

with as few steps as possible.

Let N be a sufficient number of steps to learn $\mathcal{P}$ as defined in Equation (6.1). 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}:

$$\rho_t^{\text{output}} = \underbrace{\mathcal{P} \circ \mathcal{N}_{m_t-1} \circ \mathcal{P} \circ \cdots \circ \mathcal{P} \circ \mathcal{N}_1 \circ \mathcal{P}(\rho_t)}_{m_t \text{ times}}. \tag{6.4}$$

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 Figure 6.1 for an illustration). However, to not overload the expressions we do not 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

$$x \sim \{\mathrm{Tr}(\rho_t^{\text{output}} M_i^t)\}_{i \in \mathcal{I}}. \tag{6.5}$$

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

$$N = \Omega\left(\frac{d^2}{\varepsilon^2}\right). \tag{6.6}$$

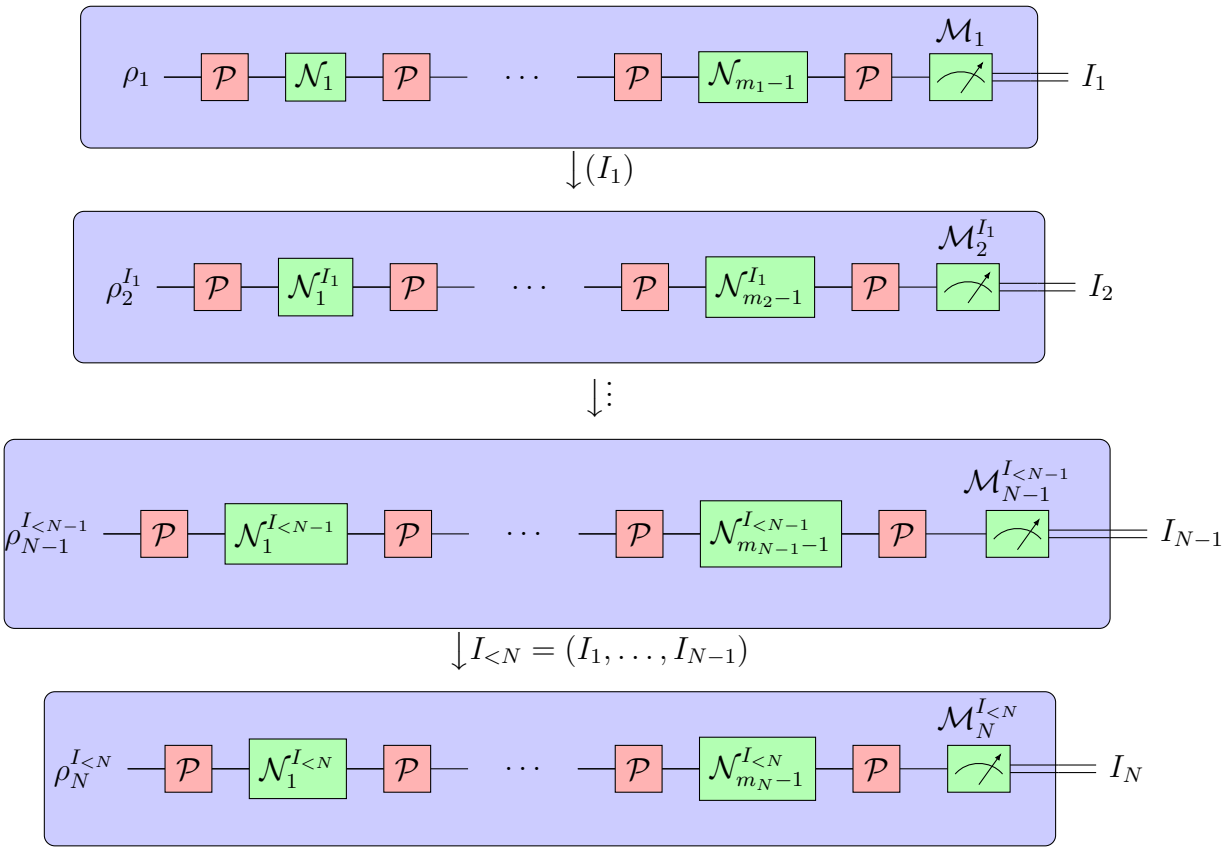

Figure 6.1: Illustration of an adaptive strategy for learning Pauli channel.

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 further increases the noise and does not aid in extracting additional 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. For the proof, we follow a standard strategy for proving lower bounds of learning problems (see e.g., [FGLE12; HHJWY16]).

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 $x \neq y \in [\![1, M]\!] : \mathrm{TV}(p_x, p_y) \geq \varepsilon$, we say that the family $\mathcal{F}$ is ε-separated. These quantum channels have the form for $x \in [\![1, M]\!]$:

$$\mathcal{P}_x(\rho) = \sum_{P\in\{\mathbb{I},X,Y,Z\}^{\otimes n}} p_x(P) P\rho P = \sum_{P\in\{\mathbb{I},X,Y,Z\}^{\otimes n}} \left(\frac{1 + 4\alpha_x(P)\varepsilon}{d^2}\right) P\rho P \tag{6.7}$$

where $\alpha_x(P) = \pm 1$ to be chosen randomly so that $\alpha_x(P) = -\alpha_x(\sigma(P))$ for some matching σ of $\{\mathbb{I}, X, Y, Z\}^{\otimes n}$ [1]. Suppose that we have already constructed an ε-separated family of

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

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 is not ε-far from the family $\mathcal{F}$. By the union bound and Chernoff-Hoeffding inequality [Hoe63]:

$$\mathbb{P}\left(\exists \mathcal{P}_x \in \mathcal{F} : \mathrm{TV}(p, p_x) < \varepsilon\right) \leq \sum_{x=1}^{M} \mathbb{P}\left(\sum_{P \in \mathbb{P}_n} |p(P) - p_x(P)| < 2\varepsilon\right) \tag{6.8}$$

$$= \sum_{x=1}^{M} \mathbb{P}\left(\sum_{P \in \mathbb{P}_n} 4|\alpha(P) - \alpha_x(P)| < 2d^2\right) \tag{6.9}$$

$$= \sum_{x=1}^{M} \mathbb{P}\left(\sum_{P \in \mathbb{P}_n} \mathbf{1}_{\alpha(P) \neq \alpha_x(P)} < \frac{d^2}{4}\right) \tag{6.10}$$

$$= \sum_{x=1}^{M} \mathbb{P}\left(\sum_{P \in \mathbb{P}_n/\sigma} \mathbb{E}\left(\mathbf{1}_{\alpha(P) \neq \alpha_x(P)}\right) - \mathbf{1}_{\alpha(P) \neq \alpha_x(P)} > \frac{d^2}{8}\right) \tag{6.11}$$

$$\leq \sum_{x=1}^{M} \exp(-2(d^2/2)(1/4)^2) = M \exp\left(-d^2/16\right) \tag{6.12}$$

which is strictly smaller than 1 if $M < \exp\left(d^2/16\right)$. So far, we have proven the following lemma:

Lemma 6.3.1. *There exists an ε-separated family of quantum Pauli channels of the form 6.7 and size at least $e^{d^2/16}$.*

Hence, we can use this family to encode a message $X \sim \mathrm{Uniform}\,[\![1, M]\!]$ 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 6.3.2 ([Fan61])**.** *The mutual information between the index of the actual channel X and the estimated index $\hat{X}$ is at least*

$$\mathcal{I}(X : \hat{X}) \geq 2/3 \log(M) - \log(2) = \Omega(d^2). \tag{6.13}$$

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:

$$\mathcal{I}(X : \hat{X}) \leq \mathcal{I}(X : I_1, \ldots, I_N). \tag{6.14}$$

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 gives:

$$\mathcal{I}(X : I_1, \ldots, I_N) = \sum_{k=1}^{N} \mathcal{I}(X : I_k | I_{\leq k-1}) \tag{6.15}$$

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 $\mathcal{P}^{m_t}(\rho_t) = \underbrace{\mathcal{P} \circ \mathcal{N}_{m_t-1}^t \circ \mathcal{P} \ldots \mathcal{P} \circ \mathcal{N}_1^t \circ \mathcal{P}}_{m_t \text{ times}}(\rho_t)$ the quantum channel applied to the input quantum state ρ_t. We denote by q the joint distribution of $(X, I_1, \ldots, I_N)$:

$$q(x, i_1, \ldots, i_N) = \frac{1}{M} \prod_{t=1}^{N} \lambda_{i_t}^t \left\langle \phi_{i_t}^t \right| \mathcal{P}_x^{m_t}(\rho_t) \left| \phi_{i_t}^t \right\rangle . \tag{6.16}$$

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:

$$\begin{aligned}
q_{\leq k}(x, i_1, \ldots, i_k) &= \sum_{i_{k+1}, \ldots, i_N} \frac{1}{M} \prod_{t=1}^{N} \lambda_{i_t}^t \left\langle \phi_{i_t}^t \right| \mathcal{P}_x^{m_t}(\rho_t) \left| \phi_{i_t}^t \right\rangle & (6.17)\\
&= \frac{1}{M} \prod_{t=1}^{k} \lambda_{i_t}^t \left\langle \phi_{i_t}^t \right| \mathcal{P}_x^{m_t}(\rho_t) \left| \phi_{i_t}^t \right\rangle \prod_{t=k+1}^{N} \sum_{i_t} \lambda_{i_t}^t \left\langle \phi_{i_t}^t \right| \mathcal{P}_x^{m_t}(\rho_t) \left| \phi_{i_t}^t \right\rangle & (6.18)\\
&= \frac{1}{M} \prod_{t=1}^{k} \lambda_{i_t}^t \left\langle \phi_{i_t}^t \right| \mathcal{P}_x^{m_t}(\rho_t) \left| \phi_{i_t}^t \right\rangle \prod_{t=k+1}^{N} \operatorname{Tr}(\mathcal{P}_x^{m_t}(\rho_t)) & (6.19)\\
&= \frac{1}{M} \prod_{t=1}^{k} \lambda_{i_t}^t \left\langle \phi_{i_t}^t \right| \mathcal{P}_x^{m_t}(\rho_t) \left| \phi_{i_t}^t \right\rangle . & (6.20)
\end{aligned}$$

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 6.3.3. *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:*

$$\mathcal{I}(X : I_k | I_{\leq k-1}) \leq 5 \mathbb{E}_x \mathbb{E}_{i \sim q_{\leq k-1}} \left[\sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,x})^2 \right]. \tag{6.21}$$

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:*

$$\mathcal{I}(X : I_k | I_{\leq k-1}) \leq 5\mathbb{E}_x \left[\sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,x})^2 \right]. \tag{6.22}$$

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

$$\frac{q(x, i_k | i_{\leq k-1})}{q(x|i_{\leq k-1}) q(i_k|i_{\leq k-1})} = \frac{q(x, i_{\leq k}) q(i_{\leq k-1})}{q(x, i_{\leq k-1}) q(i_{\leq k})} = \frac{\lambda_{i_k}^k \left(\frac{1+u_{i_k}^{k,x}}{d} \right) q(x, i_{\leq k-1}) q(i_{\leq k-1})}{q(x, i_{\leq k-1}) \sum_y q(y, i_{\leq k})} \tag{6.23}$$

$$= \frac{\lambda_{i_k}^k \left(\frac{1+u_{i_k}^{k,x}}{d} \right) q(i_{\leq k-1})}{\sum_y q(y, i_{\leq k})} = \frac{\lambda_{i_k}^k \left(\frac{1+u_{i_k}^{k,x}}{d} \right) q(i_{\leq k-1})}{\sum_y q(y, i_{\leq k-1}) \lambda_{i_k}^k \left(\frac{1+u_{i_k}^{k,y}}{d} \right)} \tag{6.24}$$

$$= \frac{(1 + u_{i_k}^{k,x}) q(i_{\leq k-1})}{\sum_y q(y, i_{\leq k-1})(1 + u_{i_k}^{k,y})} = \frac{(1 + u_{i_k}^{k,x})}{\sum_y q(y|i_{\leq k-1})(1 + u_{i_k}^{k,y})}. \tag{6.25}$$

Therefore by Jensen's inequality:

$$\mathcal{I}(X : I_k | I_{\leq k-1}) = \mathbb{E} \left(\log \left(\frac{q(x, i_k | i_{\leq k-1})}{q(x|i_{\leq k-1}) q(i_k|i_{\leq k-1})} \right) \right) \tag{6.26}$$

$$= \mathbb{E} \left(\log \left(\frac{(1 + u_{i_k}^{k,x})}{\sum_y q(y|i_{\leq k-1})(1 + u_{i_k}^{k,y})} \right) \right) \tag{6.27}$$

$$\leq \mathbb{E} \left(\log(1 + u_{i_k}^{k,x}) - \sum_y q(y|i_{\leq k-1}) \log(1 + u_{i_k}^{k,y}) \right) \tag{6.28}$$

$$= \mathbb{E} \left(\log(1 + u_{i_k}^{k,x}) \right) - \sum_y \mathbb{E} \left(q(y|i_{\leq k-1}) \log(1 + u_{i_k}^{k,y}) \right). \tag{6.29}$$

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

$$\mathbb{E} \left(\log(1 + u_{i_k}^{k,x}) \right) = \mathbb{E}_{x, i \sim q} \log(1 + u_{i_k}^{k,x}) \leq \mathbb{E}_{x, i \sim q} u_{i_k}^{k,x} = \mathbb{E}_{x, i \sim q_{\leq k}} u_{i_k}^{k,x} \tag{6.30}$$

$$= \mathbb{E}_{x, i \sim q_{\leq k-1}} \sum_{i_k} \frac{\lambda_{i_k}^k}{d} (1 + u_{i_k}^{k,x}) u_{i_k}^{k,x} = \mathbb{E}_{x, i \sim q_{\leq k-1}} \sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,x})^2 \tag{6.31}$$

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 $-\log(1+x) \leq -x + x^2$ verified for all $x \in (-1/2, +\infty)$:

$$\mathbb{E}\left(-\sum_y q(y|i_{\leq k-1}) \log(1+u_{i_k}^{k,y})\right) = -\sum_y \mathbb{E}_{x,i\sim q} q(y|i_{\leq k-1}) \log(1+u_{i_k}^{k,y}) \tag{6.32}$$

$$= -\sum_y \mathbb{E}_{x,i\sim q_{\leq k-1}} q(y|i_{\leq k-1}) \sum_{i_k} \frac{\lambda_{i_k}^k}{d}(1+u_{i_k}^{k,x}) \log(1+u_{i_k}^{k,y}) \tag{6.33}$$

$$\leq \sum_y \mathbb{E}_{x,i\sim q_{\leq k-1}} q(y|i_{\leq k-1}) \sum_{i_k} \frac{\lambda_{i_k}^k}{d}(1+u_{i_k}^{k,x})(-u_{i_k}^{k,y} + (u_{i_k}^{k,y})^2) \tag{6.34}$$

$$= \sum_y \mathbb{E}_{x,i\sim q_{\leq k-1}} q(y|i_{\leq k-1}) \sum_{i_k} \frac{\lambda_{i_k}^k}{d}(-u_{i_k}^{k,y} - u_{i_k}^{k,y}u_{i_k}^{k,x} + (u_{i_k}^{k,y})^2 + u_{i_k}^{k,x}(u_{i_k}^{k,y})^2) \tag{6.35}$$

$$\leq \sum_y \mathbb{E}_{x,i\sim q_{\leq k-1}} q(y|i_{\leq k-1}) \sum_{i_k} \frac{\lambda_{i_k}^k}{d}(2(u_{i_k}^{k,x})^2 + 2(u_{i_k}^{k,y})^2) \tag{6.36}$$

$$= 4\sum_y \mathbb{E}_{x,i\sim q_{\leq k-1}} q(y|i_{\leq k-1}) \sum_{i_k} \frac{\lambda_{i_k}^k}{d}(u_{i_k}^{k,x})^2 = 4\mathbb{E}_{x,i\sim q_{\leq k-1}} \sum_{i_k} \frac{\lambda_{i_k}^k}{d}(u_{i_k}^{k,x})^2. \tag{6.37}$$

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 6.3.4. *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 $|\phi\rangle \in \mathbf{S}^d$:*

$$|\langle\phi| d\mathcal{P}\mathcal{N}_{m-1}\mathcal{P}\ldots\mathcal{P}\mathcal{N}_1\mathcal{P}(\rho)|\phi\rangle - 1| \leq (4\varepsilon)^m. \tag{6.38}$$

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

$$\mathcal{M}_x(\rho) = \mathcal{P}_x(\rho) - \mathrm{Tr}(\rho)\frac{\mathbb{I}}{d} = \sum_{P\in\{\mathbb{I},X,Y,Z\}^{\otimes n}} \frac{4\alpha_x(P)\varepsilon}{d^2} P\rho P, \tag{6.39}$$

where we have used the fact (see Lemma 6.6.2) that for all ρ:

$$\sum_{P\in\{\mathbb{I},X,Y,Z\}^{\otimes n}} P\rho P = d\mathrm{Tr}(\rho)\mathbb{I}. \tag{6.40}$$

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 :

$$\mathcal{P}_x\mathcal{N}\mathcal{P}_x(\rho) = \mathcal{P}_x\mathcal{N}\left(\mathrm{Tr}(\rho)\frac{\mathbb{I}}{d} + \mathcal{M}_x(\rho)\right) = \mathcal{P}_x\left(\mathrm{Tr}(\rho)\frac{\mathbb{I}}{d} + \mathcal{N}\mathcal{M}_x(\rho)\right) \tag{6.41}$$

$$= \mathrm{Tr}(\rho)\frac{\mathbb{I}}{d} + \mathcal{M}_x(\frac{\mathbb{I}}{d} + \mathcal{N}\mathcal{M}_x(\rho)) = \mathrm{Tr}(\rho)\frac{\mathbb{I}}{d} + \mathcal{M}_x\mathcal{N}\mathcal{M}_x(\rho) \tag{6.42}$$

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 $\mathcal{P}_x$:

$$\underbrace{\mathcal{P}_x\mathcal{N}_{m-1}\mathcal{P}_x\ldots\mathcal{P}_x\mathcal{N}_1\mathcal{P}_x}_{m\text{ times}}(\rho)=\mathrm{Tr}(\rho)\frac{\mathbb{I}}{d}+\underbrace{\mathcal{M}_x\mathcal{N}_{m-1}\mathcal{M}_x\ldots\mathcal{M}_x\mathcal{N}_1\mathcal{M}_x}_{m\text{ times}}(\rho) \tag{6.43}$$

Therefore

$$\begin{aligned}\langle\phi|\, d\mathcal{P}\mathcal{N}_{m-1}\mathcal{P}\ldots\mathcal{P}\mathcal{N}_1\mathcal{P}(\rho)\,|\phi\rangle &= \langle\phi|\, I+d\mathcal{M}\mathcal{N}_{m-1}\mathcal{M}\ldots\mathcal{M}\mathcal{N}_1\mathcal{M}(\rho)\,|\phi\rangle &(6.44)\\ &= 1+d\,\langle\phi|\,\mathcal{M}\mathcal{N}_{m-1}\mathcal{M}\ldots\mathcal{M}\mathcal{N}_1\mathcal{M}(\rho)\,|\phi\rangle\,. &(6.45)\end{aligned}$$

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 6.6.2:

$$\begin{aligned}|\langle\phi|\,\mathcal{M}(X)\,|\phi\rangle\,| &= \left|\langle\phi|\sum_{P\in\mathbb{P}_n}\frac{4\alpha(P)\varepsilon}{d^2}PXP\,|\phi\rangle\right|\leq\frac{4\varepsilon}{d^2}\sum_{P\in\mathbb{P}_n}|\,\langle\phi|\,PXP\,|\phi\rangle\,| &(6.46)\\ &\leq\frac{4\varepsilon}{d^2}\sum_{P\in\mathbb{P}_n}\langle\phi|\,P|X|P\,|\phi\rangle=\frac{4\varepsilon}{d^2}\,\langle\phi|\,d\mathrm{Tr}|X|\mathbb{I}\,|\phi\rangle=\frac{4\varepsilon}{d}\mathrm{Tr}|X|, &(6.47)\end{aligned}$$

moreover we can also obtain:

$$\begin{aligned}\mathrm{Tr}|\mathcal{M}(X)| &= \left\|\sum_{P\in\mathbb{P}_n}\frac{4\alpha(P)\varepsilon}{d^2}PXP\right\|_1\leq\sum_{P\in\mathbb{P}_n}\frac{4\varepsilon}{d^2}\|PXP\|_1 &(6.48)\\ &=\sum_{P\in\mathbb{P}_n}\frac{4\varepsilon}{d^2}\mathrm{Tr}|X|=4\varepsilon\mathrm{Tr}|X|, &(6.49)\end{aligned}$$

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

$$\begin{aligned}\mathrm{Tr}|\mathcal{N}_j(X)| &= \|\mathcal{N}_j(X)\|_1=\left\|\sum_i\lambda_i\mathcal{N}_j(|\phi_i\rangle\langle\phi_i|)\right\|_1 &(6.50)\\ &\leq\sum_i\|\lambda_i\mathcal{N}_j(|\phi_i\rangle\langle\phi_i|)\|_1=\sum_i|\lambda_i|=\mathrm{Tr}|X|. &(6.51)\end{aligned}$$

Therefore by induction we can prove:

$$\begin{aligned}|\langle\phi|\,d\mathcal{P}\mathcal{N}_{m-1}\mathcal{P}\ldots\mathcal{P}\mathcal{N}_1\mathcal{P}(\rho)\,|\phi\rangle-1| &= d|\,\langle\phi|\,\mathcal{M}\mathcal{N}_{m-1}\mathcal{M}\ldots\mathcal{M}\mathcal{N}_1\mathcal{M}(\rho)\,|\phi\rangle\,|\\ &\leq d\frac{4\varepsilon}{d}\mathrm{Tr}|\mathcal{N}_{m-1}\mathcal{M}\ldots\mathcal{M}\mathcal{N}_1\mathcal{M}(\rho)|\\ &=4\varepsilon\mathrm{Tr}|\mathcal{M}\ldots\mathcal{M}\mathcal{N}_1\mathcal{M}(\rho)|\\ &\leq(4\varepsilon)^2\mathrm{Tr}|\mathcal{N}_{m-2}\ldots\mathcal{M}\mathcal{N}_1\mathcal{M}(\rho)|\\ &\leq(4\varepsilon)^m\,. &(6.52)\end{aligned}$$

□

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

Lemma 6.3.5. *The mutual information can be upper bounded as follows:*

$$\mathcal{I}(X:I_1,\ldots,I_N)=\mathcal{O}(N\varepsilon^2). \tag{6.53}$$

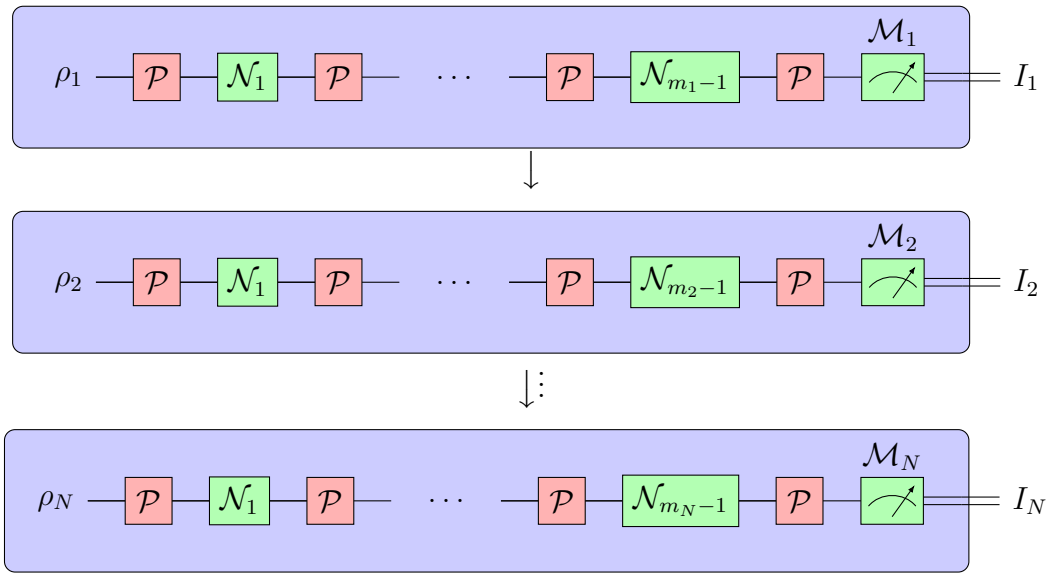

Figure 6.2: Illustration of a non-adaptive strategy for learning Pauli channel.

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 Lemmas 6.3.3 and 6.3.4:

$$\mathcal{I}(X : I_t | I_{\leq t-1}) \leq 5\mathbb{E}_{x,i\sim q_{\leq t-1}} \sum_{i_t} \frac{\lambda_{i_t}^t}{d} (u_{i_t}^{t,x})^2 \leq 5\mathbb{E}_{x,i\sim q_{\leq k-1}} \sum_{i_t} \frac{\lambda_{i_t}^t}{d} 16\varepsilon^2 = 80\varepsilon^2 \tag{6.54}$$

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

$$\mathcal{I}(X : I_1, \dots, I_N) = \sum_{t=1}^{N} \mathcal{I}(X : I_t | I_{\leq t-1}) = \mathcal{O}(N\varepsilon^2). \tag{6.55}$$

Using Lemmas 6.3.2 and 6.3.5 we obtain:

$$\Omega(d^2) \leq \mathcal{I} \leq \mathcal{O}(N\varepsilon^2), \tag{6.56}$$

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 [FW20] implies an upper bound of $\mathcal{O}\left(\frac{d^3 \log(d)}{\varepsilon^2}\right)$ (see Section 6.7 for a self contained proof), so there is a gap between our lower bound and this upper bound. However, note that the algorithm of [FW20] (and in fact most channel learning protocols we are aware of) use non-adaptive strategies. We will now show that indeed [FW20] is optimal if we restrict to non-adaptive protocols.

6.4 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:

$$\rho_t^{\text{output}} = \underbrace{\mathcal{P}\circ\mathcal{N}_{m_t-1}\circ\mathcal{P}\circ\cdots\circ\mathcal{P}\circ\mathcal{N}_1\circ\mathcal{P}(\rho_t)}_{m_t \text{ times}}. \tag{6.57}$$

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:

$$\text{Tr}\left(\lambda_{i_t}^t \left|\phi_{i_t}^t\right\rangle\left\langle\phi_{i_t}^t\right| \rho_t^{\text{output}}\right) = \lambda_{i_t}^t \left\langle\phi_{i_t}^t\right| \mathcal{P}\circ\mathcal{N}_{m_t-1}\circ\mathcal{P}\circ\cdots\circ\mathcal{P}\circ\mathcal{N}_1\circ\mathcal{P}(\rho_t) \left|\phi_{i_t}^t\right\rangle. \tag{6.58}$$

We refer to Figure 6.2 for an illustration. We prove the following lower bound on the total number of measurements and steps:

Theorem 6.4.1. *The problem of Pauli channel tomography using non-adaptive incoherent measurements requires a total number of channel uses verifying:*

$$\sum_{t=1}^{N} m_t = \Omega\left(\frac{d^4}{\varepsilon^6}\right) \tag{6.59}$$

or a total number of steps satisfying:

$$N = \Omega\left(\frac{d^3}{\varepsilon^2}\right). \tag{6.60}$$

At a first sight we can think that this Theorem is not comparable to Theorem 6.3.1 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 [FW20] 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 process tomography problem has an optimal copy complexity of $\Theta\left(\frac{d^6}{\varepsilon^2}\right)$ (see Chapter 5): 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 6.3.1. 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 6.3.3 that the mutual information can be upper bounded as follows:

$$\mathcal{I}(X : I_1, \ldots, I_N) = \sum_{t=1}^{N} \mathcal{I}(X : I_t | I_{\leq t-1}) \leq 5 \sum_{t=1}^{N} \mathbb{E}_{x,i\sim q_{\leq t-1}} \sum_{i_t} \frac{\lambda_{i_t}^t}{d} (u_{i_t}^{t,x})^2. \tag{6.61}$$

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

$$5\mathbb{E}_{x,i\sim q_{\leq t-1}} \sum_{i_t} \frac{\lambda_{i_t}^t}{d} (u_{i_t}^{t,x})^2 = 5\frac{1}{M} \sum_{x=1}^{M} \sum_{i_t \in \mathcal{I}_t} \frac{\lambda_{i_t}^t}{d} (u_{i_t}^{t,x})^2. \tag{6.62}$$

We remark that we only need to approximate

$$\frac{1}{M} \sum_{x=1}^{M} \sum_{t=1}^{N} \sum_{i \in \mathcal{I}_t} \frac{\lambda_i^t}{d} \left(\langle \phi_i^t | \, d\mathcal{P}_x^{m_t}(\rho_t) \, | \phi_i^t \rangle - 1 \right)^2. \tag{6.63}$$

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 (6.52)). Also, we have for all $t \in [N]$, $\sum_{i \in \mathcal{I}_t} \frac{\lambda_i^t}{d} = 1$ so

$$\sum_{t=1}^{N} \sum_{i \in \mathcal{I}_t} \frac{\lambda_i^t}{d} \left(\langle \phi_i^t | \, d\mathcal{P}_x^{m_t}(\rho_t) \, | \phi_i^t \rangle - 1 \right)^2 \in [0, 16N\varepsilon^2]. \tag{6.64}$$

Therefore by Hoeffding's inequality [Hoe63] for $s = \sqrt{\frac{(16N\varepsilon^2)^2 \log(10)}{2M}}$

$$\mathbb{P}\left(\left| \frac{1}{M} \sum_{x=1}^{M} \sum_{t=1}^{N} \sum_{i \in \mathcal{I}_t} \frac{\lambda_i^t}{d} \left(\langle \phi_i^t | \, d\mathcal{P}_x^{m_t}(\rho_t) \, | \phi_i^t \rangle - 1 \right)^2 - \mathbb{E}_\alpha \sum_{t=1}^{N} \sum_{i \in \mathcal{I}_t} \frac{\lambda_i^t}{d} \left(\langle \phi_i^t | \, d\mathcal{P}_\alpha^{m_t}(\rho_t) \, | \phi_i^t \rangle - 1 \right)^2 \right| > s \right) \tag{6.65}$$

$$\leq \exp\left(-\frac{2Ms^2}{(16N\varepsilon^2)^2} \right) = \exp\left(-\frac{2M \frac{(16N\varepsilon^2)^2 \log(10)}{2M}}{(16N\varepsilon^2)^2} \right) = \frac{1}{10}. \tag{6.66}$$

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:

$$\begin{aligned} \Omega(d^2) \leq \mathcal{I} &\leq \mathcal{I}(X : I_1, \ldots, I_N) \\ &\leq 5 \sum_{t} \sum_{i_t \in \mathcal{I}_t} \frac{\lambda_{i_t}^t}{d} \mathbb{E}_\alpha \left(\langle \phi_{i_t}^t | \, d\mathcal{P}_\alpha^{m_t}(\rho_t) \, | \phi_{i_t}^t \rangle - 1 \right)^2 + 60N\varepsilon^2 \exp(-\Omega(d^2)). \end{aligned} \tag{6.67}$$

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

Lemma 6.4.1. *For all $t \in [N]$, for all unit vectors $|\phi\rangle \in \mathbf{S}^d$:*

$$\mathbb{E}_\alpha \left(\langle \phi | \, d\mathcal{P}_\alpha(\rho_t) \, | \phi \rangle - 1 \right)^2 \leq \frac{32\varepsilon^2}{d}. \tag{6.68}$$

If the claim is true, the inequalities 6.67 imply using the fact that for all $t \leq N$, $\sum_{i_t \in \mathcal{I}_t} \lambda_{i_t}^t = d$:

$$\Omega(d^2) \leq \mathcal{I} \leq 5 \sum_{t=1}^{N} \sum_{i_t \in \mathcal{I}_t} \frac{\lambda_{i_t}^t}{d} \frac{\varepsilon^2}{d} + 60 N \varepsilon^2 \exp(-\Omega(d^2)) \leq \mathcal{O}\left(N \frac{\varepsilon^2}{d}\right) \tag{6.69}$$

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:

$$\mathbb{E}_\alpha(\langle\phi| \, d\mathcal{P}_\alpha(\rho_t) \, |\phi\rangle - 1)^2 = \mathbb{E}_\alpha \left(\sum_{P \in \mathbb{P}_n} \frac{4\alpha(P)\varepsilon}{d} \langle\phi| \, P\rho_t P^\dagger \, |\phi\rangle \right)^2 \tag{6.70}$$

$$= \mathbb{E}_\alpha \sum_{P,Q \in \mathbb{P}_n} \frac{16\alpha(P)\alpha(Q)\varepsilon^2}{d^2} \langle\phi| \, P\rho_t P^\dagger \, |\phi\rangle \langle\phi| \, Q\rho_t Q^\dagger \, |\phi\rangle \tag{6.71}$$

$$= \sum_{P \in \mathbb{P}_n} \frac{16\varepsilon^2}{d^2} (\langle\phi| \, P\rho_t P^\dagger \, |\phi\rangle \langle\phi| \, P\rho_t P^\dagger \, |\phi\rangle - \langle\phi| \, P\rho_t P^\dagger \, |\phi\rangle \langle\phi| \, \sigma(P)\rho_t \sigma(P)^\dagger \, |\phi\rangle) \tag{6.72}$$

$$\leq \sum_{P \in \mathbb{P}_n} \frac{32\varepsilon^2}{d^2} \langle\phi| \, P\rho_t P^\dagger \, |\phi\rangle^2 \leq \sum_{P \in \mathbb{P}_n} \frac{32\varepsilon^2}{d^2} \langle\phi| \, P\rho_t^2 P^\dagger \, |\phi\rangle = \frac{32\varepsilon^2}{d^2} \langle\phi| \, d\mathrm{Tr}(\rho_t^2)\mathbb{I} \, |\phi\rangle \leq \frac{32\varepsilon^2}{d}, \tag{6.73}$$

where we used the Cauchy-Schwarz 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 6.4.2. *For all $t \in [N]$, $m \geq 2$ and unit vectors $|\phi\rangle \in \mathbf{S}^d$:*

$$\mathbb{E}_\alpha \left(\langle\phi| \, d\mathcal{P}_\alpha^m(\rho_t) \, |\phi\rangle - 1\right)^2 \leq 4m \cdot \frac{(4\varepsilon)^{2m}}{d^{\min\{2, m-1\}}}. \tag{6.74}$$

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:

$$\underbrace{\mathcal{P}_\alpha \mathcal{N}_{m-1} \mathcal{P}_\alpha \ldots \mathcal{P}_\alpha \mathcal{N}_1 \mathcal{P}_\alpha(\rho)}_{m \text{ times}} = \mathrm{Tr}(\rho)\frac{\mathbb{I}}{d} + \underbrace{\mathcal{M}_\alpha \mathcal{N}_{m-1} \mathcal{M}_\alpha \ldots \mathcal{M}_\alpha \mathcal{N}_1 \mathcal{M}_\alpha(\rho)}_{m \text{ times}}. \tag{6.75}$$

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

$$\mathcal{M}_\alpha(\rho) = \mathcal{P}_\alpha(\rho) - \mathrm{Tr}(\rho)\frac{\mathbb{I}}{d} = \sum_{P \in \mathbb{P}_n} \frac{4\alpha(P)\varepsilon}{d^2} P\rho P = \sum_{P \in \mathbb{P}_n} \frac{4\alpha(P)\varepsilon}{d^2} \mathcal{N}_P(\rho) \tag{6.76}$$

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 $\mathcal{N}_{P,m} = \mathcal{N}_{P_m} \mathcal{N}_{m-1} \mathcal{N}_{P_{m-1}} \ldots \mathcal{N}_{P_2} \mathcal{N}_1 \mathcal{N}_{P_1}$, we can develop the

quantity we want to upper bound as follows:

$$\mathbb{E}_\alpha \left(\langle\phi| \, d\mathcal{P}_\alpha^m(\rho) \, |\phi\rangle - 1\right)^2 = d^2 \mathbb{E}_\alpha \left(\langle\phi| \, \mathcal{M}_\alpha \mathcal{N}_{m-1} \mathcal{M}_\alpha \dots \mathcal{M}_\alpha \mathcal{N}_1 \mathcal{M}_\alpha(\rho) \, |\phi\rangle\right)^2 \tag{6.77}$$

$$= d^2 \mathbb{E}_\alpha \left(\sum_{P_1,\dots,P_m} \frac{4\alpha(P_1)\varepsilon}{d^2} \cdots \frac{4\alpha(P_m)\varepsilon}{d^2} \langle\phi| \, \mathcal{N}_{P_m} \mathcal{N}_{m-1} \mathcal{N}_{P_{m-1}} \dots \mathcal{N}_{P_2} \mathcal{N}_1 \mathcal{N}_{P_1}(\rho) \, |\phi\rangle \right)^2 \tag{6.78}$$

$$= \frac{(4\varepsilon)^{2m}}{d^{4m-2}} \sum_{P,Q \in \mathbb{P}_n} \mathbb{E}_\alpha \left(\alpha(P_1) \cdots \alpha(P_m)\alpha(Q_1) \cdots \alpha(Q_m)\right) \langle\phi| \, \mathcal{N}_{P,m}(\rho) \, |\phi\rangle \, \langle\phi| \, \mathcal{N}_{Q,m}(\rho) \, |\phi\rangle \, . \tag{6.79}$$

If $Q_1, \sigma(Q_1) \notin \{P_1, \dots, P_m, Q_2, \dots, Q_m\}$ 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 $m \geq 3$:

$$\frac{(4\varepsilon)^{2m}}{d^{4m-2}} \sum_{P,Q : Q_1 = P_1} \mathbb{E}_\alpha \left(\alpha(P_1) \cdots \alpha(P_m)\alpha(Q_1) \cdots \alpha(Q_m)\right) \langle\phi| \, \mathcal{N}_{P,m}(\rho) \, |\phi\rangle \, \langle\phi| \, \mathcal{N}_{Q,m}(\rho) \, |\phi\rangle \tag{6.80}$$

$$\leq \frac{(4\varepsilon)^{2m}}{d^{4m-2}} \sum_{P,Q : Q_1 = P_1} \langle\phi| \, \mathcal{N}_{P,m}(\rho) \, |\phi\rangle \, \langle\phi| \, \mathcal{N}_{Q,m}(\rho) \, |\phi\rangle \tag{6.81}$$

$$= \frac{(4\varepsilon)^{2m}}{d^{4m-2}} \sum_{P,Q_{<m} : Q_1 = P_1} \langle\phi| \sum_{P_m} \mathcal{N}_{P,m}(\rho) \, |\phi\rangle \, \langle\phi| \sum_{Q_m} \mathcal{N}_{Q,m}(\rho) \, |\phi\rangle \tag{6.82}$$

$$= \frac{(4\varepsilon)^{2m}}{d^{4m-2}} \sum_{P,Q_{<m} : Q_1 = P_1} \langle\phi| \, d\mathrm{Tr}(\mathcal{N}_{m-1,\dots,1,P_1}(\rho))\mathbb{I} \, |\phi\rangle \, \langle\phi| \, d\mathrm{Tr}(\mathcal{N}_{m-1,\dots,1,Q_1}(\rho))\mathbb{I} \, |\phi\rangle \tag{6.83}$$

$$= \frac{(4\varepsilon)^{2m}}{d^{4m-2}} \sum_{P,Q_{<m} : Q_1 = P_1} d^2 = \frac{(4\varepsilon)^{2m}}{d^{4m-2}} (d^2)^{2m-3} d^2 = \frac{(4\varepsilon)^{2m}}{d^2}. \tag{6.84}$$

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

$$\mathbb{E}_\alpha \left(\langle\phi| \, d\mathcal{P}_\alpha^m(\rho) \, |\phi\rangle - 1\right)^2 \leq 2(2m-1)\frac{(4\varepsilon)^{2m}}{d^2} \leq 4m\frac{(4\varepsilon)^{2m}}{d^2}. \tag{6.85}$$

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

$$\mathbb{E}_\alpha \left(\langle\phi| \, d\mathcal{P}_\alpha^2(\rho) \, |\phi\rangle - 1\right)^2 \leq 6\frac{(4\varepsilon)^4}{d}. \tag{6.86}$$

□

Using the inequalities 6.67 and the fact that for all $t \in [N]$, $\sum_{i_t \in \mathcal{I}_t} \lambda_{i_t}^t = d$, we deduce:

$$2^{10} \sum_{t : m_t \leq 2} \frac{\varepsilon^2}{d} + 4 \sum_{t : m_t \geq 3} m_t \frac{(4\varepsilon)^{2m_t}}{d^2} = \Omega(d^2). \tag{6.87}$$

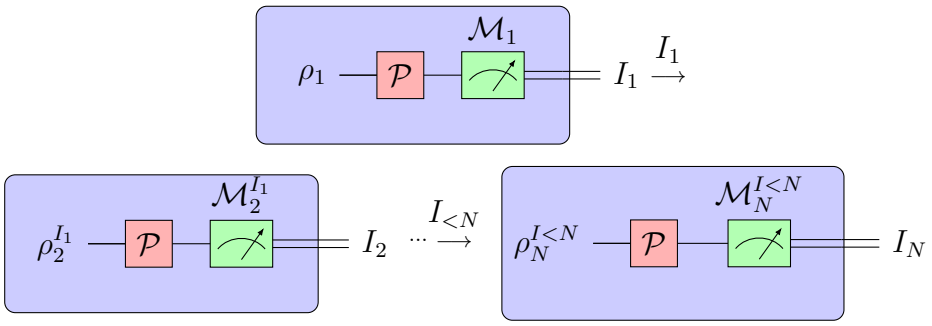

Figure 6.3: Illustration of an adaptive strategy for learning Pauli channel using one channel per step.

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 6.3.3. 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 6.3.1. In the next section, we explore how to overcome this difficulty in some regime of ε and improve the general lower bound $N=\Omega(d^2/\varepsilon^2)$.

6.5 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 6.3.1 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 Figure 6.3 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:

$$\mathrm{Tr}\left(\rho_{t+1}^{I_1,\ldots,I_t}\lambda_{i_{t+1}}^{I_1,\ldots,I_t}\left|\phi_{i_{t+1}}^{I_1,\ldots,I_t}\right\rangle\!\left\langle\phi_{i_{t+1}}^{I_1,\ldots,I_t}\right|\right)=\lambda_{i_{t+1}}^{I_1,\ldots,I_t}\left\langle\phi_{i_{t+1}}^{I_1,\ldots,I_t}\left|\rho_{t+1}^{I_1,\ldots,I_t}\right|\phi_{i_{t+1}}^{I_1,\ldots,I_t}\right\rangle. \tag{6.88}$$

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

$$N = \Omega\left(\frac{d^{5/2}}{\varepsilon^2}\right). \tag{6.89}$$

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

$$N = \Omega\left(\frac{d^3}{\varepsilon^2}\right). \tag{6.90}$$

In this Theorem, we show that we can improve on the general lower bound of Theorem 6.3.1 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 [FW20] 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 $x \in \mathcal{F} = [\![1, M]\!]$:

$$\mathcal{P}_x(\rho) = \sum_{P \in \mathbb{P}_n} \frac{1 + 2\tilde{\alpha}_x(P)\varepsilon d/\|\alpha_x\|_2}{d^2} P\rho P = \sum_{P \in \mathbb{P}_n} p_x(P) P\rho P \tag{6.91}$$

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

$$\mathbb{P}\left(\mathrm{TV}(p_\alpha, p_\beta) < \frac{\varepsilon}{5}\right) \leq \exp\left(-\Omega(d^2)\right). \tag{6.92}$$

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

$$\mathbb{E}\left(\mathrm{TV}(p_\alpha, p_\beta)\right) \geq \frac{7\varepsilon}{20}. \tag{6.93}$$

Proof. We start by writing:

$$\mathrm{TV}(p_\alpha, p_\beta) = \frac{\varepsilon}{d}\sum_{P\in\mathbb{P}_n}\left|\frac{\tilde{\alpha}(P)}{\|\alpha\|_2} - \frac{\tilde{\beta}(P)}{\|\beta\|_2}\right| \tag{6.94}$$

$$\geq \frac{\varepsilon}{d}\sum_{P\in\mathbb{P}_n}\left|\frac{\alpha(P)}{\|\alpha\|_2} - \frac{\beta(P)}{\|\beta\|_2}\right| - \frac{\varepsilon}{d}\left|\frac{\sum_{Q\in\mathbb{P}_n}\alpha(Q)}{\|\alpha\|_2} - \frac{\sum_{Q\in\mathbb{P}_n}\beta(Q)}{\|\beta\|_2}\right| \tag{6.95}$$

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 $\{\alpha(P)/\|\alpha\|_2\}_P$:

$$\mathbb{E}\left(\frac{\varepsilon}{d}\left|\frac{\sum_{Q\in\mathbb{P}_n}\alpha(Q)}{\|\alpha\|_2} - \frac{\sum_{Q\in\mathbb{P}_n}\beta(Q)}{\|\beta\|_2}\right|\right) \leq 2\mathbb{E}\left(\frac{\varepsilon}{d}\left|\frac{\sum_{P\in\mathbb{P}_n}\alpha(P)}{\|\alpha\|_2}\right|\right) \tag{6.96}$$

$$= \frac{2\varepsilon}{d}\frac{\mathbb{E}\left(\left|\sum_{P\in\mathbb{P}_n}\alpha(P)\right|\right)}{\mathbb{E}\left(\|\alpha\|_2\right)} \leq \frac{4\varepsilon}{d}. \tag{6.97}$$

Indeed, by the Cauchy-Schwarz inequality:

$$\mathbb{E}\left(\left|\sum_{P\in\mathbb{P}_n}\alpha(P)\right|\right) \leq \sqrt{\mathbb{E}\left(\left(\sum_{P\in\mathbb{P}_n}\alpha(P)\right)^2\right)} = d \tag{6.98}$$

and by the Hölder's inequality:

$$\mathbb{E}\left(\|\alpha\|_2\right) \geq \sqrt{\frac{\mathbb{E}\left(\|\alpha\|_2^2\right)^3}{\mathbb{E}\left(\|\alpha\|_2^4\right)}} = \sqrt{\frac{(d^2)^3}{d^2(d^2-1)+3d^2}} \geq \frac{d}{2}. \tag{6.99}$$

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

$$\mathbb{E}\left(\frac{\varepsilon}{d}\sum_{P\in\mathbb{P}_n}\left|\frac{\alpha(P)}{\|\alpha\|_2} - \frac{\beta(P)}{\|\beta\|_2}\right|\right) = \varepsilon d\,\mathbb{E}\left(\left|\frac{\alpha(P)}{\|\alpha\|_2} - \frac{\beta(P)}{\|\beta\|_2}\right|\right) \geq \varepsilon d\sqrt{\frac{\mathbb{E}\left(\left|\frac{\alpha(P)}{\|\alpha\|_2} - \frac{\beta(P)}{\|\beta\|_2}\right|^2\right)^3}{\mathbb{E}\left(\left|\frac{\alpha(P)}{\|\alpha\|_2} - \frac{\beta(P)}{\|\beta\|_2}\right|^4\right)}} \tag{6.100}$$

$$\geq \varepsilon d\sqrt{\frac{8/d^6}{48/d^4}} \geq \frac{2\varepsilon}{5}. \tag{6.101}$$

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

$$\mathbb{E}\left(\left|\frac{\alpha(P)}{\|\alpha\|_2} - \frac{\beta(P)}{\|\beta\|_2}\right|^2\right) = 2\mathbb{E}\left(\frac{\alpha(P)^2}{\|\alpha\|_2^2}\right) - 2\mathbb{E}\left(\frac{\alpha(P)\beta(P)}{\|\alpha\|_2\|\beta\|_2}\right) = 2\frac{\mathbb{E}\left(\alpha(P)^2\right)}{\mathbb{E}\left(\|\alpha\|_2^2\right)} = \frac{2}{d^2}. \tag{6.102}$$

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

$$\mathbb{E}\left(\left|\frac{\alpha(P)}{\|\alpha\|_2}-\frac{\beta(P)}{\|\beta\|_2}\right|^4\right)\leq 16\mathbb{E}\left(\frac{\alpha(P)^4}{\|\alpha\|_2^4}\right)=16\frac{\mathbb{E}\left(\alpha(P)^4\right)}{\mathbb{E}\left(\|\alpha\|_2^4\right)}=\frac{48}{d^2(d^2-1)+3d^2}\leq\frac{48}{d^4}. \tag{6.103}$$

Therefore the expected value of the TV-distance satisfies:

$$\mathbb{E}\left(\mathrm{TV}(p_\alpha,p_\beta)\right)\geq\frac{2\varepsilon}{5}-\frac{4\varepsilon}{d}\geq\frac{7\varepsilon}{20}. \tag{6.104}$$

□

Once we have a lower bound on the expected value of $\mathrm{TV}(p_\alpha,p_\beta)$, we can proceed to prove Lemma 6.5.1.

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:

$$|\,\mathrm{TV}(p_\alpha,p_\gamma)-\mathrm{TV}(p_\beta,p_\delta)|\leq|\,\mathrm{TV}(p_\alpha,p_\gamma)-\mathrm{TV}(p_\gamma,p_\beta)|+|\,\mathrm{TV}(p_\gamma,p_\beta)-\mathrm{TV}(p_\beta,p_\delta)| \tag{6.105}$$

$$\leq\mathrm{TV}(p_\alpha,p_\beta)+\mathrm{TV}(p_\gamma,p_\delta). \tag{6.106}$$

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

$$\mathrm{TV}(p_\alpha,p_\beta)\leq\frac{\varepsilon}{d}\sum_{P\in\mathbb{P}_n}\left|\frac{\alpha(P)}{\|\alpha\|_2}-\frac{\beta(P)}{\|\beta\|_2}\right|+\frac{\varepsilon}{d}\left|\sum_{P\in\mathbb{P}_n}\frac{\alpha(P)}{\|\alpha\|_2}-\frac{\beta(P)}{\|\beta\|_2}\right|\leq\frac{2\varepsilon}{d}\sum_{P\in\mathbb{P}_n}\left|\frac{\alpha(P)}{\|\alpha\|_2}-\frac{\beta(P)}{\|\beta\|_2}\right| \tag{6.107}$$

$$\leq\frac{2\varepsilon}{d}\sum_{P\in\mathbb{P}_n}\left|\frac{\alpha(P)}{\|\alpha\|_2}-\frac{\alpha(P)}{\|\beta\|_2}\right|+\frac{2\varepsilon}{d}\sum_{P\in\mathbb{P}_n}\left|\frac{\alpha(P)}{\|\beta\|_2}-\frac{\beta(P)}{\|\beta\|_2}\right| \tag{6.108}$$

$$\leq\frac{2\varepsilon}{d}\sum_{P\in\mathbb{P}_n}|\alpha(P)|\left|\frac{\|\alpha\|_2-\|\beta\|_2}{\|\alpha\|_2\|\beta\|_2}\right|+\frac{2\varepsilon}{d}\sum_{P\in\mathbb{P}_n}\left|\frac{\alpha(P)-\beta(P)}{\|\beta\|_2}\right| \tag{6.109}$$

$$=\frac{2\varepsilon}{d}\|\alpha\|_1\left|\frac{\|\alpha\|_2-\|\beta\|_2}{\|\alpha\|_2\|\beta\|_2}\right|+\frac{2\varepsilon}{d}\left|\frac{\|\alpha-\beta\|_1}{\|\beta\|_2}\right| \tag{6.110}$$

$$\leq 2\varepsilon\|\alpha\|_2\frac{\|\alpha-\beta\|_2}{\|\alpha\|_2\|\beta\|_2}+2\varepsilon\frac{\|\alpha-\beta\|_2}{\|\beta\|_2} \tag{6.111}$$

$$\leq 2\varepsilon\frac{4}{d}\|\alpha-\beta\|_2+2\varepsilon\|\alpha-\beta\|_2\frac{4}{d}=\frac{16\varepsilon}{d}\|\alpha-\beta\|_2. \tag{6.112}$$

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 $x+y\leq\sqrt{2}\sqrt{x^2+y^2}$:

$$|\,\mathrm{TV}(p_\alpha,p_\gamma)-\mathrm{TV}(p_\beta,p_\delta)|\leq\mathrm{TV}(p_\alpha,p_\beta)+\mathrm{TV}(p_\gamma,p_\delta)\leq\frac{16\varepsilon}{d}\|\alpha-\beta\|_2+\frac{16\varepsilon}{d}\|\gamma-\delta\|_2 \tag{6.113}$$

$$\leq\frac{16\sqrt{2}\varepsilon}{d}\sqrt{\|\alpha-\beta\|_2^2+\|\gamma-\delta\|_2^2}=:L\|(\alpha,\gamma)-(\beta,\delta)\|_2. \tag{6.114}$$

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. [Mat99]):

$$f(\alpha, \beta) = \inf_{(\gamma,\delta)\in E^2} \{\mathrm{TV}(p_\gamma, p_\delta) + L\|(\alpha, \beta) - (\gamma, \delta)\|_2\}. \tag{6.115}$$

We can control the expected value of f using the lower bound on the expected value of $\mathrm{TV}(p_\alpha, p_\beta)$ (Lemma 6.5.2) as follows:

$$\mathbb{E}\,(f) = \mathbb{E}\,(f\mathbf{1}_{E^2}) + \mathbb{E}\left(f\mathbf{1}_{(E^2)^c}\right) \geq \mathbb{E}\,(f\mathbf{1}_{E^2}) \geq \frac{7\varepsilon}{20} - 8\varepsilon \exp(-\Omega(d^2)) \geq \frac{3\varepsilon}{10} \tag{6.116}$$

because

$$|\mathbb{E}\,(f(\alpha, \beta)\mathbf{1}_{E^2}) - \mathbb{E}\,(\mathrm{TV}(p_\alpha, p_\beta))\,| = \mathbb{E}\left(\mathrm{TV}(p_\alpha, p_\beta)\mathbf{1}_{(E^2)^c}\right) \leq 8\varepsilon \mathbb{P}\,(E^c) \leq 8\varepsilon \exp(-\Omega(d^2)) \tag{6.117}$$

where we have used the fact that $\mathrm{TV}(p_\alpha, p_\beta) \leq 4\varepsilon$ and $\mathbb{P}\,(E^c) = \mathbb{P}\,(\|\alpha\|_2 \leq d/4) \leq \exp(-\Omega(d^2))$. Indeed, we can apply the concentration of Lipschitz functions of Gaussian random variables [Wai19, Theorem 2.26] for the function $\alpha \to \|\alpha\|_2$ which is 1-Lipschitz by the triangle inequality:

$$|\|\alpha\|_2 - \|\beta\|_2| \leq \|\alpha - \beta\|_2 \tag{6.118}$$

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

$$\mathbb{P}\,(E^c) = \mathbb{P}\,(\|\alpha\|_2 \leq d/4) = \mathbb{P}\,(\|\alpha\|_2 - \mathbb{E}\,(\|\alpha\|_2) \leq -d/4) \leq \exp(-\Omega(d^2)). \tag{6.119}$$

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 [Wai19, Theorem 2.26], we obtain for all $s \geq 0$:

$$\mathbb{P}\,(|f - \mathbb{E}\,(f)\,| > s) \leq \exp\left(-\frac{cd^2s^2}{\varepsilon^2}\right) \tag{6.120}$$

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

$$\begin{aligned}
&\mathbb{P}\,(\mathrm{TV}(p_\alpha, p_\beta) < \varepsilon/5) && (6.121)\\
&= \mathbb{P}\left(\mathrm{TV}(p_\alpha, p_\beta) < \varepsilon/5, (\alpha, \beta) \in E^2\right) + \mathbb{P}\left(\mathrm{TV}(p_\alpha, p_\beta) < \varepsilon/5, (\alpha, \beta) \notin E^2\right) && (6.122)\\
&\leq \mathbb{P}\left(f(\alpha, \beta) < \varepsilon/5, (\alpha, \beta) \in E^2\right) + \mathbb{P}\left((\alpha, \beta) \notin E^2\right) && (6.123)\\
&\leq \mathbb{P}\left(f(\alpha, \beta) - \mathbb{E}\,(f) < \varepsilon/5 - 3\varepsilon/10, (\alpha, \beta) \in E^2\right) + 2\mathbb{P}\,(\alpha \notin E) && (6.124)\\
&\leq \mathbb{P}\,(f(\alpha, \beta) - \mathbb{E}\,(f) < -\varepsilon/10) + 2\mathbb{P}\,(\alpha \notin E) && (6.125)\\
&\leq 3\exp(-\Omega(d^2)) \leq \exp(-\Omega(d^2)). && (6.126)
\end{aligned}$$

□

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 6.5.3 ([Fan61]). *The mutual information between the encoder and the decoder is at least*

$$\mathcal{I} \geq 2/3 \log(M) - \log(2) \geq \Omega(d^2). \tag{6.127}$$

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:

$$\mathcal{I} \leq \mathcal{I}(X : I_1, \ldots, I_N). \tag{6.128}$$

By upper bounding the mutual information between X and $I_1, \ldots, I_N$ and using a contradiction argument, we prove Theorem 6.5.1 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:*

$$N = \Omega\left(\frac{d^{5/2}}{\varepsilon^2}\right). \tag{6.129}$$

Proof. Recall that we can write the mutual information as: $\mathcal{I}(X : I_1, \ldots, I_N) = \sum_{k=1}^{N} \mathcal{I}(X : I_k | I_{\leq k-1})$. Fix $k \in [N]$, by Lemma 6.3.3, we can upper bound the conditional mutual information:

$$\mathcal{I}(X : I_k | I_{\leq k-1}) \leq 5 \mathbb{E}_x \mathbb{E}_{i \sim q_{\leq k-1}} \left[\sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,x})^2 \right] \tag{6.130}$$

where we use the notation:

$$u_{i_k}^{k,x} = \left\langle \phi_{i_k}^k \right| \left(d\mathcal{P}_x(\rho_k) - \mathbb{I} \right) \left| \phi_{i_k}^k \right\rangle = \left\langle \phi_{i_k}^k \right| \left(\sum_{P \in \mathbb{P}_n} \frac{2\tilde{\alpha}_x(P)\varepsilon}{\|\alpha_x\|_2} P\rho_k P \right) \left| \phi_{i_k}^k \right\rangle \tag{6.131}$$

$$= \sum_{P \in \mathbb{P}_n} \frac{2\alpha_x(P)\varepsilon}{\|\alpha_x\|_2} \left\langle \phi_{i_k}^k \right| P\rho_k P \left| \phi_{i_k}^k \right\rangle - \sum_{P,Q \in \mathbb{P}_n} \frac{2\alpha_x(Q)\varepsilon}{d^2 \|\alpha_x\|_2} \left\langle \phi_{i_k}^k \right| P\rho_k P \left| \phi_{i_k}^k \right\rangle \tag{6.132}$$

$$= \sum_{P \in \mathbb{P}_n} \frac{2\alpha_x(P)\varepsilon}{\|\alpha_x\|_2} \left\langle \phi_{i_k}^k \right| P\rho_k P \left| \phi_{i_k}^k \right\rangle - \sum_{P \in \mathbb{P}_n} \frac{2\alpha_x(P)\varepsilon}{d\|\alpha_x\|_2}. \tag{6.133}$$

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

$$u_{i_k}^{k,\alpha} = \sum_{P \in \mathbb{P}_n} \frac{2\alpha(P)\varepsilon}{\|\alpha\|_2} \left\langle \phi_{i_k}^k \right| P\rho_k P \left| \phi_{i_k}^k \right\rangle - \sum_{P \in \mathbb{P}_n} \frac{2\alpha(P)\varepsilon}{d\|\alpha\|_2} \tag{6.134}$$

$$= \frac{2}{\|\alpha\|_2} \sum_{P \in \mathbb{P}_n} \alpha(P)\varepsilon \left\langle \phi_{i_k}^k \right| P(\rho_k - \mathbb{I}/d)P \left| \phi_{i_k}^k \right\rangle. \tag{6.135}$$

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

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

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. Note that the error probability can be absorbed in the total error probability by adding these inequalities in the construction of the family $\mathcal{F}$.

Proposition 6.5.1. *There is a universal constant $C > 0$ such that with probability at least $9/10$ we have:*

$$\sum_{k=1}^{N} \frac{1}{M} \sum_{x=1}^{M} \sum_{i_1,\dots,i_{k-1}} \left(\prod_{t=1}^{k-1} \lambda_{i_t}^t \left(\frac{1+u_{i_t}^{t,x}}{d} \right) \right) \sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,x})^2 \tag{6.136}$$

$$\leq \sum_{k=1}^{N} \mathbb{E}_\alpha \left[\sum_{i_1,\dots,i_{k-1}} \left(\prod_{t=1}^{k-1} \lambda_{i_t}^t \left(\frac{1+u_{i_t}^{t,\alpha}}{d} \right) \right) \sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,\alpha})^2 \right] + N\varepsilon^2 \exp(-Cd^2). \tag{6.137}$$

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

$$f_k(x) = \sum_{i_1,\dots,i_{k-1}} \left(\prod_{t=1}^{k-1} \lambda_{i_t}^t \left(\frac{1+u_{i_t}^{t,x}}{d} \right) \right) \sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,x})^2. \tag{6.138}$$

Similarly we define:

$$f_k(\alpha) = \sum_{i_1,\dots,i_{k-1}} \left(\prod_{t=1}^{k-1} \lambda_{i_t}^t \left(\frac{1+u_{i_t}^{t,\alpha}}{d} \right) \right) \sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,\alpha})^2. \tag{6.139}$$

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 6.5.4) 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 [Hoe63] writes:

$$\mathbb{P}\left(\left| \sum_{k=1}^{N} \frac{1}{M} \sum_{x=1}^{M} f_k(x) - \mathbb{E}\left(\sum_{k=1}^{N} \frac{1}{M} \sum_{x=1}^{M} f_k(x) \right) \right| > s \right) \leq 2\exp\left(-\frac{2s^2}{\sum_{j=1}^{M} \left(\frac{32N\varepsilon^2}{M}\right)^2} \right). \tag{6.140}$$

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

$$\mathbb{P}\left(\left| \sum_{k=1}^{N} \frac{1}{M} \sum_{x=1}^{M} f_k(x) - \sum_{k=1}^{N} \mathbb{E}_\alpha(f_k(\alpha)) \right| > s \right) \leq 2\exp\left(-\frac{s^2 M}{512 N^2 \varepsilon^4} \right). \tag{6.141}$$

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

$$\sum_{k=1}^{N} \frac{1}{M} \sum_{x=1}^{M} \sum_{i_1,\dots,i_{k-1}} \left(\prod_{t=1}^{k-1} \lambda_{i_t}^t \left(\frac{1+u_{i_t}^{t,x}}{d} \right) \right) \sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,x})^2 \tag{6.142}$$

$$= \sum_{k=1}^{N} \frac{1}{M} \sum_{x=1}^{M} f_k(x) \leq \sum_{k=1}^{N} \mathbb{E}_\alpha(f_k(\alpha)) + 25N\varepsilon^2\sqrt{\frac{\log(20)}{M}} \tag{6.143}$$

$$\leq \sum_{k=1}^{N} \mathbb{E}_\alpha \left[\sum_{i_1,\dots,i_{k-1}} \left(\prod_{t=1}^{k-1} \lambda_{i_t}^t \left(\frac{1+u_{i_t}^{t,\alpha}}{d} \right) \right) \sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,\alpha})^2 \right] + N\varepsilon^2 \exp(-Cd^2) \tag{6.144}$$

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

□

Therefore we obtain the upper bound on the mutual information:

$$\begin{aligned}
&\sum_{k=1}^{N} \mathcal{I}(X : I_k | I_{\leq k-1}) \leq 5 \sum_{k=1}^{N} \mathbb{E}_{x, i \sim q_{\leq k-1}} \sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,x})^2 \\
&= 5 \sum_{k=1}^{N} \frac{1}{M} \sum_{x=1}^{M} \sum_{i_1,\ldots,i_{k-1}} \left(\prod_{t=1}^{k-1} \lambda_{i_t}^t \left(\frac{1 + u_{i_t}^{t,x}}{d} \right) \right) \sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,x})^2 \\
&\leq 5 \sum_{k=1}^{N} \mathbb{E}_{\alpha} \left[\sum_{i_1,\ldots,i_{k-1}} \left(\prod_{t=1}^{k-1} \lambda_{i_t}^t \left(\frac{1 + u_{i_t}^{t,\alpha}}{d} \right) \right) \sum_{i_k} \frac{\lambda_{i_k}^k}{d} (u_{i_k}^{k,\alpha})^2 \right] + 5N\varepsilon^2 \exp(-Cd^2) \qquad (6.145) \\
&= 5 \sum_{k=1}^{N} \mathbb{E}_{\leq k} \mathbb{E}_{\alpha} \left[\left(\prod_{t=1}^{k-1} \left(1 + u_{i_t}^{t,\alpha} \right) \right) (u_{i_k}^{k,\alpha})^2 \right] + 5N\varepsilon^2 \exp(-Cd^2)
\end{aligned}$$

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

$$\mathcal{I}(X : I_1, \ldots, I_N) \leq 5 \sum_{k=1}^{N} \mathbb{E}_k \mathbb{E}_{\alpha} \left[(u_{i_k}^{k,\alpha})^2 \right] + 5N\varepsilon^2 \exp(-Cd^2). \qquad (6.146)$$

For this expression, using methods similar to the proof of Theorem 6.4.1, 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 6.4.1.

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 $(i_1, \ldots, i_{k-1})$. For this reason, we use Gaussian integration by parts (Theorem 1.4.7) 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:

$$u_{i_t}^{t,\alpha} = \frac{2}{\|\alpha\|_2} \sum_{P \in \mathbb{P}_n} \alpha(P) \varepsilon \left\langle \phi_{i_t}^t \right| P \tilde{\rho}_t P \left| \phi_{i_t}^t \right\rangle . \qquad (6.147)$$

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

$$\begin{aligned}
&\mathbb{E} \left(\|\alpha\|_2^2 \right) \mathbb{E} \left(\left(\prod_{t=1}^{k-1} \left(1 + u_{i_t}^{t,\alpha} \right) \right) (u_{i_k}^{k,\alpha})^2 \right) \qquad (6.148) \\
&= 2\varepsilon \sum_{P \in \mathbb{P}_n} \left\langle \phi_{i_k}^k \right| P \tilde{\rho}_k P \left| \phi_{i_k}^k \right\rangle \mathbb{E} \left(\|\alpha\|_2^2 \right) \mathbb{E} \left(\frac{\alpha(P)}{\|\alpha\|_2} \left(u_{i_k}^{k,\alpha} \right) \prod_{t=1}^{k-1} \left(1 + u_{i_t}^{t,\alpha} \right) \right) \qquad (6.149) \\
&= 2\varepsilon \sum_{P \in \mathbb{P}_n} \left\langle \phi_{i_k}^k \right| P \tilde{\rho}_k P \left| \phi_{i_k}^k \right\rangle \mathbb{E} \left(\alpha(P) \left(\|\alpha\|_2 u_{i_k}^{k,\alpha} \right) \prod_{t=1}^{k-1} \left(1 + u_{i_t}^{t,\alpha} \right) \right) \qquad (6.150) \\
&= 2\varepsilon \sum_{P \in \mathbb{P}_n} \left\langle \phi_{i_k}^k \right| P \tilde{\rho}_k P \left| \phi_{i_k}^k \right\rangle \mathbb{E} \left(\alpha(P) F(\alpha) \right), \qquad (6.151)
\end{aligned}$$

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 (Theorem 1.4.7) implies:

$$\mathbb{E}\left(\alpha(P)F(\alpha)\right) = \mathbb{E}\left(\partial_P F(\alpha)\right) \tag{6.152}$$

$$= 2\varepsilon \left\langle \phi_{i_k}^k \right| P\tilde{\rho}_k P \left| \phi_{i_k}^k \right\rangle \mathbb{E}\left(\prod_{t=1}^{k-1}\left(1 + u_{i_t}^{t,\alpha}\right)\right) + \sum_{s=1}^{k-1} \mathbb{E}\left(\|\alpha\|_2 u_{i_k}^{k,\alpha} \partial_P u_{i_s}^{s,\alpha} \prod_{t\in[k-1]\setminus s}\left(1 + u_{i_t}^{t,\alpha}\right)\right) \tag{6.153}$$

Moreover, we have

$$\|\alpha\|_2 \partial_P u_{i_s}^{s,\alpha} = 2\frac{\left\langle \phi_{i_s}^s \right| P\tilde{\rho}_s P \left| \phi_{i_s}^s \right\rangle \varepsilon \|\alpha\|_2 - \partial_P \|\alpha\|_2 \sum_{P\in\mathbb{P}_n} \alpha(P) \left\langle \phi_{i_s}^s \right| P\tilde{\rho}_s P \left| \phi_{i_s}^s \right\rangle \varepsilon}{\|\alpha\|_2} \tag{6.154}$$

$$= 2\left\langle \phi_{i_s}^s \right| P\tilde{\rho}_s P \left| \phi_{i_s}^s \right\rangle \varepsilon - \frac{1}{\|\alpha\|_2}\alpha(P) u_{i_s}^{s,\alpha} \tag{6.155}$$

hence:

$$\mathbb{E}_{\leq k}\mathbb{E}\left(\left(\prod_{t=1}^{k-1}\left(1 + u_{i_t}^{t,\alpha}\right)\right)(u_{i_k}^{k,\alpha})^2\right) = \mathbb{E}_{\leq k}\frac{2\varepsilon}{d^2}\sum_{P\in\mathbb{P}_n}\left\langle \phi_{i_k}^k \right| P\tilde{\rho}_k P \left| \phi_{i_k}^k \right\rangle \mathbb{E}\left(\alpha(P)F(\alpha)\right) \tag{6.156}$$

$$= \mathbb{E}_{\leq k}\frac{4\varepsilon^2}{d^2}\sum_{P\in\mathbb{P}_n}\left\langle \phi_{i_k}^k \right| P\tilde{\rho}_k P \left| \phi_{i_k}^k \right\rangle^2 \mathbb{E}\left(\prod_{t=1}^{k-1}\left(1 + u_{i_t}^{t,\alpha}\right)\right) \tag{L1}$$

$$+ \mathbb{E}_{\leq k}\frac{4\varepsilon^2}{d^2}\sum_{P\in\mathbb{P}_n}\sum_{s=1}^{k-1}\left\langle \phi_{i_k}^k \right| P\tilde{\rho}_k P \left| \phi_{i_k}^k \right\rangle \left\langle \phi_{i_s}^s \right| P\tilde{\rho}_s P \left| \phi_{i_s}^s \right\rangle \mathbb{E}\left(u_{i_k}^{k,\alpha}\prod_{t\in[k-1]\setminus s}\left(1 + u_{i_t}^{t,\alpha}\right)\right) \tag{L2}$$

$$- \mathbb{E}_{\leq k}\frac{2\varepsilon}{d^2}\sum_{P\in\mathbb{P}_n}\left\langle \phi_{i_k}^k \right| P\tilde{\rho}_k P \left| \phi_{i_k}^k \right\rangle \sum_{s=1}^{k-1}\mathbb{E}\left(\frac{\alpha(P)}{\|\alpha\|_2}u_{i_k}^{k,\alpha}u_{i_s}^{s,\alpha}\prod_{t\in[k-1]\setminus s}\left(1 + u_{i_t}^{t,\alpha}\right)\right). \tag{L3}$$

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

$$\sum_{P\in\mathbb{P}_n}\frac{1}{d}\sum_{i_k}\lambda_{i_k}^k \left\langle \phi_{i_k}^k \right| P\tilde{\rho}_k P \left| \phi_{i_k}^k \right\rangle^2 \leq \sum_{P\in\mathbb{P}_n}\frac{1}{d}\cdot \mathrm{Tr}(P\tilde{\rho}_k^2 P) = \sum_{P\in\mathbb{P}_n}\frac{1}{d}\cdot\mathrm{Tr}(\tilde{\rho}_k^2) \leq d, \tag{6.157}$$

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:

$$\text{(L1)} \leq \mathbb{E}_{\leq k-1}\frac{4\varepsilon^2}{d}\mathbb{E}\left(\prod_{t=1}^{k-1}\left(1 + u_{i_t}^{t,\alpha}\right)\right) = \frac{4\varepsilon^2}{d}. \tag{6.158}$$

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

$$\frac{1}{d}\sum_{i_t}\lambda_{i_t}^t (u_{i_t}^{t,\alpha})^2 \leq 16\varepsilon^2. \tag{6.159}$$

Observe that if we apply this upper bound directly on the expression of the conditional mutual information (Lemma 6.3.3) 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 6.3.1. 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 6.6.1

$$\sum_{i_t} \lambda_{i_t}^t (u_{i_t}^{t,\alpha})^2 = \frac{4\varepsilon^2}{\|\alpha\|_2^2} \sum_{i_t} \lambda_{i_t}^t \left\langle \phi_{i_t}^t \right| \left(\sum_{P\in\mathbb{P}_n} \alpha(P) P\tilde{\rho}_t P \right) \left| \phi_{i_t}^t \right\rangle^2 \leq \frac{4\varepsilon^2}{\|\alpha\|_2^2} \mathrm{Tr} \left(\sum_{P\in\mathbb{P}_n} \alpha(P) P\tilde{\rho}_t P \right)^2 \tag{6.160}$$

$$= \frac{4\varepsilon^2}{\|\alpha\|_2^2} \mathrm{Tr} \left(\sum_{P\in\mathbb{P}_n} \alpha(P) \frac{1}{d} \sum_{R\in\mathbb{P}_n} \mathrm{Tr}(R\tilde{\rho}_t) PRP \right)^2 \tag{6.161}$$

$$= \frac{4\varepsilon^2}{\|\alpha\|_2^2} \mathrm{Tr} \left(\sum_{P\in\mathbb{P}_n} \alpha(P) \frac{1}{d} \sum_{R\in\mathbb{P}_n} \mathrm{Tr}(R\tilde{\rho}_t) (-1)^{R.P} R \right)^2 \tag{6.162}$$

$$= \frac{4\varepsilon^2}{\|\alpha\|_2^2} \sum_{P,P',R,R'\in\mathbb{P}_n} \alpha(P)\alpha(P') \frac{1}{d^2} \mathrm{Tr}(R\tilde{\rho}_t) \mathrm{Tr}(R'\tilde{\rho}_t) (-1)^{R.P} (-1)^{R'.P'} \mathrm{Tr}(RR') \tag{6.163}$$

$$= \frac{4\varepsilon^2}{\|\alpha\|_2^2} \sum_{P,P',R\in\mathbb{P}_n} \alpha(P)\alpha(P') \frac{1}{d} \cdot \mathrm{Tr}(R\tilde{\rho}_t)^2 (-1)^{R.P} (-1)^{R.P'} \tag{6.164}$$

$$= \frac{4\varepsilon^2}{\|\alpha\|_2^2} \sum_{R\in\mathbb{P}_n} \left(\sum_{P\in\mathbb{P}_n} \alpha(P) (-1)^{R.P} \right)^2 \frac{1}{d} \cdot \mathrm{Tr}(R\tilde{\rho}_t)^2 \tag{6.165}$$

$$\leq \frac{16\varepsilon^2}{\|\alpha\|_2^2} \sum_{P,P',R\in\mathbb{P}_n} \alpha(P)\alpha(P') \frac{1}{d} (-1)^{R.(PP')} \tag{6.166}$$

$$= \frac{16\varepsilon^2}{\|\alpha\|_2^2} \sum_{P,P'\in\mathbb{P}_n} \alpha(P)\alpha(P') \cdot d \cdot \mathbf{1}_{PP'=\mathbb{I}} = \frac{16d\varepsilon^2}{\|\alpha\|_2^2} \sum_{P=P'\in\mathbb{P}_n} \alpha(P)^2 = 16d\varepsilon^2. \tag{6.167}$$

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:

$$|\mathrm{Tr}(R\tilde{\rho})| = \left| \sum_i \lambda_i \mathrm{Tr}(|\phi_i\rangle\langle\phi_i| \tilde{\rho}) \right| \leq \sum_i |\lambda_i \mathrm{Tr}(|\phi_i\rangle\langle\phi_i| \tilde{\rho})| \leq \sum_i |\mathrm{Tr}(|\phi_i\rangle\langle\phi_i| (\rho - \mathbb{I}/d))| \tag{6.168}$$

$$\leq \sum_i |\mathrm{Tr}(|\phi_i\rangle\langle\phi_i| \rho)| + \sum_i |\mathrm{Tr}(|\phi_i\rangle\langle\phi_i| \mathbb{I}/d))| = \mathrm{Tr}(\rho) + \mathrm{Tr}(\mathbb{I}/d) = 2. \tag{6.169}$$

□

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, if we

denote $\Pi_k^\alpha = \prod_{t<k}\left(1+u_{i_t}^{t,\alpha}\right)$, (L3) can be controlled as follows:

$$(\text{L3}) = -\mathbb{E}_{\leq k}\frac{1}{d^2}\sum_{s=1}^{k-1}\mathbb{E}\left(\left(u_{i_k}^{k,\alpha}\right)^2 u_{i_s}^{s,\alpha}\prod_{t\in[k-1]\setminus s}\left(1+u_{i_t}^{t,\alpha}\right)\right) \tag{6.170}$$

$$= \frac{1}{d^2}\mathbb{E}_{\leq k}\left(-\mathbb{E}\left(\left(\sum_{s<k}u_{i_s}^{s,\alpha}(u_{i_k}^{k,\alpha})^2\right)\prod_{t\in[k-1]\setminus s}\left(1+u_{i_t}^{t,\alpha}\right)\right)\right) \tag{6.171}$$

$$\leq \frac{1}{d^2}\mathbb{E}_{\leq k}\left(\mathbb{E}\left(\left(\left|\sum_{s<k}u_{i_s}^{s,\alpha}\right|(u_{i_k}^{k,\alpha})^2\right)\prod_{t\in[k-1]\setminus s}\left(1+u_{i_t}^{t,\alpha}\right)\right)\right) \tag{6.172}$$

$$\leq \frac{16\varepsilon^2}{d^2}\mathbb{E}_{<k}\left(\mathbb{E}\left(\left(\left|\sum_{s<k}\frac{u_{i_s}^{s,\alpha}}{(1+u_{i_s}^{s,\alpha})}\right|\right)\prod_{t<k}\left(1+u_{i_t}^{t,\alpha}\right)\right)\right) \qquad \text{(Lemma 6.5.4)} \tag{6.173}$$

$$\leq \frac{16\varepsilon^2}{d^2}\sqrt{\mathbb{E}_\alpha\mathbb{E}_{<k}\left(\left|\sum_{s<k}\frac{u_{i_s}^{s,\alpha}}{(1+u_{i_s}^{s,\alpha})}\right|^2\Pi_k^\alpha\right)}\sqrt{\mathbb{E}_\alpha\mathbb{E}_{<k}\Pi_k^\alpha} \qquad \text{(Cauchy-Schwarz)} \tag{6.174}$$

$$\leq \frac{16\varepsilon^2}{d^2}\sqrt{\mathbb{E}_\alpha\mathbb{E}_{<k}\sum_{s,r<k}\frac{u_{i_s}^{s,\alpha}}{(1+u_{i_s}^{s,\alpha})}\cdot\frac{u_{i_r}^{r,\alpha}}{(1+u_{i_r}^{r,\alpha})}\prod_{t<k}\left(1+u_{i_t}^{t,\alpha}\right)} \qquad \left(\mathbb{E}_{<k}\prod_{t<k}\left(1+u_{i_t}^{t,\alpha}\right)=1\right) \tag{6.175}$$

$$\leq \frac{16\varepsilon^2}{d^2}\sqrt{\mathbb{E}_\alpha\mathbb{E}_{<k}\sum_{s<k}\frac{(u_{i_s}^{s,\alpha})^2}{(1+u_{i_s}^{s,\alpha})^2}\prod_{t<k}\left(1+u_{i_t}^{t,\alpha}\right)} \qquad \left(\mathbb{E}_{\leq\max\{s,r\}}\left(u_{i_s}^{s,\alpha}u_{i_r}^{r,\alpha}\right)=0 \text{ if } s\neq r\right) \tag{6.176}$$

$$\leq \frac{64\varepsilon^2}{d^2}\sqrt{\mathbb{E}_\alpha\mathbb{E}_{<k}\sum_{s<k}(u_{i_s}^{s,\alpha})^2\prod_{t\in[k-1]\setminus s}\left(1+u_{i_t}^{t,\alpha}\right)} \qquad \left(1+u_{i_t}^{t,\alpha}\geq 1/16 \Leftarrow \varepsilon\leq 1/4d\right) \tag{6.177}$$

$$\leq \sqrt{k}\frac{256\varepsilon^3}{d^2}, \tag{6.178}$$

where we use Lemma 6.5.4 for the last inequality. Indeed, we can simplify the expectation as follows

$$\mathbb{E}_{<k}\sum_{s<k}(u_{i_s}^{s,\alpha})^2\prod_{t\in[k-1]\setminus s}\left(1+u_{i_t}^{t,\alpha}\right) = \sum_{s<k}\mathbb{E}_{<k}(u_{i_s}^{s,\alpha})^2\prod_{t\in[k-1]\setminus s}\left(1+u_{i_t}^{t,\alpha}\right) \tag{6.179}$$

$$= \sum_{s<k}\mathbb{E}_{\leq s}(u_{i_s}^{s,\alpha})^2\prod_{t\in[s-1]}\left(1+u_{i_t}^{t,\alpha}\right) \tag{6.180}$$

$$\leq \sum_{s<k}\mathbb{E}_{\leq s-1}16\varepsilon^2\prod_{t\in[s-1]}\left(1+u_{i_t}^{t,\alpha}\right) \qquad \text{(Lemma 6.5.4)} \tag{6.181}$$

$$= \sum_{s<k}16\varepsilon^2 \leq 16\varepsilon^2 k. \tag{6.182}$$

Finally, we control the line (L2) which is more involved. Let us adopt the notation for

$s, k \in [N]$:

$$M_{s,k} = \sum_{P\in\mathbb{P}_n} P\tilde{\rho}_k P \left\langle \phi^s_{i_s} \right| P\tilde{\rho}_s P \left| \phi^s_{i_s} \right\rangle \tag{6.183}$$

$$= \frac{1}{d^2} \sum_{P,Q,R\in\mathbb{P}_n} \mathrm{Tr}(\tilde{\rho}_k Q)\mathrm{Tr}(\tilde{\rho}_s R) PQP \left\langle \phi^s_{i_s} \right| PRP \left| \phi^s_{i_s} \right\rangle \tag{6.184}$$

$$= \frac{1}{d^2} \sum_{P,Q,R\in\mathbb{P}_n} \mathrm{Tr}(\tilde{\rho}_k Q)\mathrm{Tr}(\tilde{\rho}_s R)(-1)^{P.(QR)} Q \left\langle \phi^s_{i_s} \right| R \left| \phi^s_{i_s} \right\rangle \tag{6.185}$$

$$= \sum_{Q\in\mathbb{P}_n} \mathrm{Tr}(\tilde{\rho}_k Q)\mathrm{Tr}(\tilde{\rho}_s Q) \left\langle \phi^s_{i_s} \right| Q \left| \phi^s_{i_s} \right\rangle Q \qquad \text{(Lemma 6.6.1)} \tag{6.186}$$

so that we can write $\sum_{P\in\mathbb{P}_n} \left\langle \phi^k_{i_k} \right| P\tilde{\rho}_k P \left| \phi^k_{i_k} \right\rangle \left\langle \phi^s_{i_s} \right| P\tilde{\rho}_s P \left| \phi^s_{i_s} \right\rangle = \mathrm{Tr}(\left| \phi^k_{i_k} \right\rangle\!\left\langle \phi^k_{i_k} \right| M_{s,k})$. Also we use the notation $\Psi_k = \mathbb{E}_{\leq k}\mathbb{E}\left(\left(u^{k,\alpha}_{i_k} \right)^2 \prod_{t<k} \left(1 + u^{t,\alpha}_{i_t}\right) \right)$ so that we have the (in)equalities:

$$\text{(L2)} = \frac{4\varepsilon^2}{d^2}\mathbb{E}_{\leq k}\mathbb{E}\left(\sum_{s=1}^{k-1} \mathrm{Tr}(\left| \phi^k_{i_k} \right\rangle\!\left\langle \phi^k_{i_k} \right| M_{s,k}) u^{k,\alpha}_{i_k} \prod_{t\in[k-1]\setminus s} \left(1 + u^{t,\alpha}_{i_t}\right) \right) \tag{6.187}$$

$$= \frac{4\varepsilon^2}{d^2}\mathbb{E}_{\leq k}\mathbb{E}\left(\mathrm{Tr}\left(\left| \phi^k_{i_k} \right\rangle\!\left\langle \phi^k_{i_k} \right| \sum_{s=1}^{k-1} \frac{M_{s,k}}{\left(1 + u^{s,\alpha}_{i_s}\right)} \right) u^{k,\alpha}_{i_k} \prod_{t\leq k-1} \left(1 + u^{t,\alpha}_{i_t}\right) \right) \tag{6.188}$$

$$\leq \frac{4\varepsilon^2}{d^2} \sqrt{ \mathbb{E}_{\leq k}\mathbb{E}\left(\left(\mathrm{Tr}\left(\left| \phi^k_{i_k} \right\rangle\!\left\langle \phi^k_{i_k} \right| \sum_{s=1}^{k-1} \frac{M_{s,k}}{\left(1 + u^{s,\alpha}_{i_s}\right)} \right) \right)^2 \prod_{t\leq k-1} \left(1 + u^{t,\alpha}_{i_t}\right) \right) } \sqrt{\Psi_k} \tag{6.189}$$

$$\leq \frac{4\varepsilon^2}{d^2} \sqrt{ \mathbb{E}_{\leq k}\mathbb{E}\left(\mathrm{Tr}\left(\left| \phi^k_{i_k} \right\rangle\!\left\langle \phi^k_{i_k} \right| \left(\sum_{s=1}^{k-1} \frac{M_{s,k}}{\left(1 + u^{s,\alpha}_{i_s}\right)} \right)^2 \right) \prod_{t\leq k-1} \left(1 + u^{t,\alpha}_{i_t}\right) \right) } \sqrt{\Psi_k} \tag{6.190}$$

$$= \frac{4\varepsilon^2}{d^2} \sqrt{ \frac{1}{d}\mathbb{E}_{\leq k-1}\mathbb{E}\left(\mathrm{Tr}\left(\left(\sum_{s=1}^{k-1} \frac{M_{s,k}}{\left(1 + u^{s,\alpha}_{i_s}\right)} \right)^2 \right) \prod_{t\leq k-1} \left(1 + u^{t,\alpha}_{i_t}\right) \right) } \sqrt{\Psi_k} \tag{6.191}$$

where we use the Cauchy-Schwarz inequality. From the definition of $M_{s,k}$ we can write that for $s, t < k$

$$\mathrm{Tr}(M_{s,k}M_{t,k}) = d \sum_{Q\in\mathbb{P}_n} \mathrm{Tr}(\tilde{\rho}_k Q)^2 \mathrm{Tr}(\tilde{\rho}_s Q)\mathrm{Tr}(\tilde{\rho}_t Q) \left\langle \phi^s_{i_s} \right| Q \left| \phi^s_{i_s} \right\rangle \left\langle \phi^t_{i_t} \right| Q \left| \phi^t_{i_t} \right\rangle . \tag{6.192}$$

Hence

$$\mathrm{Tr}\left(\sum_{s=1}^{k-1}\frac{M_{s,k}}{(1+u_{i_s}^{s,\alpha})}\right)^2 = \sum_{s,t<k} d\sum_{Q\in\mathbb{P}_n}\mathrm{Tr}(\tilde{\rho}_k Q)^2\frac{\mathrm{Tr}(\tilde{\rho}_s Q)\mathrm{Tr}(\tilde{\rho}_t Q)\left\langle\phi_{i_s}^s\right|Q\left|\phi_{i_s}^s\right\rangle\left\langle\phi_{i_t}^t\right|Q\left|\phi_{i_t}^t\right\rangle}{(1+u_{i_s}^{s,\alpha})(1+u_{i_t}^{t,\alpha})} \tag{6.193}$$

$$= d\sum_{Q\in\mathbb{P}_n}\mathrm{Tr}(\tilde{\rho}_k Q)^2\left(\sum_{s<k}\frac{\mathrm{Tr}(\tilde{\rho}_s Q)\left\langle\phi_{i_s}^s\right|Q\left|\phi_{i_s}^s\right\rangle}{(1+u_{i_s}^{s,\alpha})}\right)^2 \tag{6.194}$$

$$\leq 4d\sum_{Q\in\mathbb{P}_n}\left(\sum_{s<k}\frac{\mathrm{Tr}(\tilde{\rho}_s Q)\left\langle\phi_{i_s}^s\right|Q\left|\phi_{i_s}^s\right\rangle}{(1+u_{i_s}^{s,\alpha})}\right)^2 \tag{6.195}$$

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 $\mathbb{E}_{\leq k-1}\mathbb{E}\left(\mathrm{Tr}\left(\left(\sum_{s=1}^{k-1}\frac{M_{s,k}}{(1+u_{i_s}^{s,\alpha})}\right)^2\right)\prod_{t\leq k-1}\left(1+u_{i_t}^{t,\alpha}\right)\right)$ 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 $\mathbb{E}_{s_2}$:

$$\mathbb{E}_{s_2}\mathrm{Tr}(\tilde{\rho}_{s_2}Q)\left\langle\phi_{i_{s_2}}^{s_2}\right|Q\left|\phi_{i_{s_2}}^{s_2}\right\rangle = \frac{1}{d}\sum_{i_{s_2}}\mathrm{Tr}(\tilde{\rho}_{s_2}Q)\lambda_{i_{s_2}}^{s_2}\left\langle\phi_{i_{s_2}}^{s_2}\right|Q\left|\phi_{i_{s_2}}^{s_2}\right\rangle = \mathrm{Tr}(\tilde{\rho}_{s_2}Q)\mathrm{Tr}(Q) = 0 \tag{6.196}$$

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 using the notation $\Pi_k^\alpha=\prod_{t<k}\left(1+u_{i_t}^{t,\alpha}\right)$:

$$(\text{L2}) \leq \frac{4\varepsilon^2}{d^2}\sqrt{\frac{1}{d}\mathbb{E}_{\leq k-1}\mathbb{E}\left(\mathrm{Tr}\left(\sum_{s=1}^{k-1}\frac{M_{s,k}}{(1+u_{i_s}^{s,\alpha})}\right)^2\Pi_k^\alpha\right)}\sqrt{\Psi_k} \tag{6.197}$$

$$\leq \frac{4\varepsilon^2}{d^2}\sqrt{\frac{1}{d}\mathbb{E}_{\leq k-1}\mathbb{E}\left(4d\sum_{Q\in\mathbb{P}_n}\left(\sum_{s<k}\frac{\mathrm{Tr}(\tilde{\rho}_s Q)\left\langle\phi_{i_s}^s\right|Q\left|\phi_{i_s}^s\right\rangle}{(1+u_{i_s}^{s,\alpha})}\right)^2\Pi_k^\alpha\right)}\sqrt{\Psi_k} \tag{6.198}$$

$$\leq \frac{4\varepsilon^2}{d^2}\sqrt{\sum_{s<k}16\sum_{Q\in\mathbb{P}_n}\mathbb{E}_{\leq s}\mathrm{Tr}(\tilde{\rho}_s Q)^2\left\langle\phi_{i_s}^s\right|Q\left|\phi_{i_s}^s\right\rangle^2\mathbb{E}\left(\Pi_s^\alpha\right)}\sqrt{\Psi_k} \qquad \left((1+u_{i_s}^{s,\alpha})\geq\frac{1}{16}\right) \tag{6.199}$$

$$\leq \frac{4\varepsilon^2}{d^2}\sqrt{\sum_{s<k}64\mathbb{E}_{\leq s}\sum_{Q\in\mathbb{P}_n}\left\langle\phi_{i_s}^s\right|Q\left|\phi_{i_s}^s\right\rangle\left\langle\phi_{i_s}^s\right|Q\left|\phi_{i_s}^s\right\rangle\mathbb{E}\left(\Pi_s^\alpha\right)}\sqrt{\Psi_k} \qquad (|\mathrm{Tr}(\tilde{\rho}_s Q)|\leq 2) \tag{6.200}$$

$$= \frac{4\varepsilon^2}{d^2}\sqrt{\sum_{s<k}64\mathbb{E}_{\leq s}\left\langle\phi_{i_s}^s\right|d\mathrm{Tr}(\left|\phi_{i_s}^s\right\rangle\left\langle\phi_{i_s}^s\right|)\mathbb{I}\left|\phi_{i_s}^s\right\rangle\mathbb{E}\left(\Pi_s^\alpha\right)}\sqrt{\Psi_k} \qquad (\text{Lemma 6.6.2}) \tag{6.201}$$

$$\leq \frac{32\varepsilon^2}{d\sqrt{d}}\sqrt{k}\sqrt{\Psi_k}. \tag{6.202}$$

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

$$\Psi_k \leq \frac{4\varepsilon^2}{d} + 256\sqrt{k}\frac{\varepsilon^3}{d^2} + \frac{32\varepsilon^2}{d\sqrt{d}}\sqrt{k}\sqrt{\Psi_k}. \tag{6.203}$$

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 $\sum_{k=1}^{N}\mathcal{I}(X : I_k|I_{<k}) \leq 5\sum_{k=1}^{N}\Psi_k + 5N\varepsilon^2\exp(-Cd^2)$. We suppose that $N \leq c\frac{d^{5/2}}{\varepsilon^2}$ for sufficiently small $c > 0$. We know that from Lemmas 6.3.3 and 6.5.3

$$c_0 d^2 \leq \mathcal{I}(X : Y) \leq 5\sum_{k\leq N}\Psi_k + 5N\varepsilon^2\exp(-Cd^2). \tag{6.204}$$

So $\sum_{k\leq N}\Psi_k \geq c'd^2$ (for example $c' = c_0/4$), on the other hand the inequality (6.203) implies:

$$\begin{aligned}
\sum_k \Psi_k &\leq \sum_k \frac{4\varepsilon^2}{d} + 256\frac{\varepsilon^3}{d^2}\sqrt{k} + 32\frac{\varepsilon^2\sqrt{k}}{d\sqrt{d}}\sqrt{\Psi_k} && (6.205)\\
&\leq 4\frac{N\varepsilon^2}{d} + 256\frac{N\varepsilon^3}{d^2}\sqrt{N} + 32\sum_k \frac{\varepsilon^2\sqrt{k}}{d\sqrt{d}}\sqrt{\Psi_k} && (6.206)\\
&\leq 4\frac{N\varepsilon^2}{\sqrt{c'}d^2}\sqrt{\sum_{k\leq N}\Psi_k} + 256\frac{N\varepsilon^2}{d^2}\sqrt{cd^{5/2}} + 32\frac{\varepsilon^2}{d\sqrt{d}}\sqrt{\sum_k k}\sqrt{\sum_k \Psi_k} \quad \text{(Cauchy-Schwarz)} && (6.207)\\
&\leq \left(\frac{8}{\sqrt{c'}d} + \frac{512}{\sqrt{c'}d^{1/4}} + 32\right)\frac{\varepsilon^2}{d\sqrt{d}}\sqrt{\sum_k k}\sqrt{\sum_k \Psi_k} && (6.208)\\
&\leq C'\frac{\varepsilon^2}{d\sqrt{d}}N\sqrt{\sum_k \Psi_k} && (6.209)
\end{aligned}$$

where C' is a universal constant. Therefore:

$$\sum_k \Psi_k \leq C'^2\left(\frac{N^2\varepsilon^4}{d^3}\right) \leq C'^2c^2d^2 \tag{6.210}$$

Hence

$$c_0 d^2 \leq \mathcal{I}(X : Y) \leq \sum_{k\leq N} 5\Psi_k + 5N\varepsilon^2\exp(-Cd^2) \leq 10C'^2c^2d^2 \tag{6.211}$$

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 6.5.1.

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

$$\sum_k \Psi_k \leq \sum_k \frac{4\varepsilon^2}{d} + 256\sqrt{H}\frac{\varepsilon^2}{d^2} + \frac{32\varepsilon}{d\sqrt{d}}\sqrt{H}\sqrt{\Psi_k} \tag{6.212}$$

$$\leq \sum_k \frac{4\varepsilon^2}{d} + 256\sqrt{H}\frac{\varepsilon^2}{d^2} + \frac{(32\varepsilon)^2 H}{d^3} + \frac{\Psi_k}{2} \tag{6.213}$$

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

$$c_0 d^2 \leq \mathcal{I}(X:Y) \leq \sum_k 5\Psi_k + 5N\varepsilon^2 \exp(-Cd^2) \tag{6.214}$$

$$\leq \frac{40\varepsilon^2 N}{d} + 10 \cdot 256\sqrt{H}\frac{\varepsilon^2 N}{d^2} + \frac{10 \cdot (32\varepsilon)^2 HN}{d^3} \tag{6.215}$$

and finally we obtain:

$$N = \Omega\left(\min\left\{\frac{d^4}{\sqrt{H}\varepsilon^2}, \frac{d^5}{H\varepsilon^2}, \frac{d^3}{\varepsilon^2}\right\}\right). \tag{6.216}$$

For $H = \mathcal{O}(d^2)$, this gives Equation (6.90). □

6.6 Properties of Pauli operators

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

Lemma 6.6.1. *We have for all $Q \in \{\mathbb{I}, X, Y, Z\}^{\otimes n}$:*

$$\sum_{P \in \mathbb{P}_n} (-1)^{P.Q} = d^2 \cdot \mathbf{1}_{Q=\mathbb{I}}. \tag{6.217}$$

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:

$$\sum_{P \in \mathbb{P}_n} (-1)^{P.Q} = \sum_{P_1, \ldots, \mathbb{P}_n \in \{\mathbb{I}, X, Y, Z\}} (-1)^{P_1.Q_1 + 2 \cdots + 2\mathbb{P}_n.Q_n} \tag{6.218}$$

$$= \prod_{i=1}^n \sum_{P_i \in \{\mathbb{I}, X, Y, Z\}} (-1)^{P_i.Q_i} \tag{6.219}$$

$$= \prod_{i=1}^n 4\mathbf{1}_{Q_i=\mathbb{I}_2} \tag{6.220}$$

$$= d^2 \mathbf{1}_{Q=\mathbb{I}_d} \tag{6.221}$$

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 6.6.2. *We have for all matrices ρ:*

$$\sum_{P\in\{\mathbb{I},X,Y,Z\}^{\otimes n}} P\rho P = d\mathrm{Tr}(\rho)\mathbb{I}. \tag{6.222}$$

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:

$$\rho = \sum_{P\in\{\mathbb{I},X,Y,Z\}^{\otimes n}} \mathrm{Tr}\left(\frac{P}{\sqrt{d}}\rho\right)\frac{P}{\sqrt{d}} = \frac{1}{d}\sum_{P\in\{\mathbb{I},X,Y,Z\}^{\otimes n}} \mathrm{Tr}\left(P\rho\right)P. \tag{6.223}$$

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

$$\begin{aligned}\sum_{P\in\mathbb{P}_n} P\rho P &= \frac{1}{d}\sum_{P,Q\in\mathbb{P}_n}\mathrm{Tr}(Q\rho)PQP = \frac{1}{d}\sum_{P,Q\in\mathbb{P}_n}\mathrm{Tr}(Q\rho)(-1)^{P.Q}QPP &(6.224)\\ &= \sum_{Q\in\mathbb{P}_n}\mathrm{Tr}(Q\rho)Q\frac{1}{d}\sum_{P\in\mathbb{P}_n}(-1)^{P.Q} = \sum_{Q\in\mathbb{P}_n}\mathrm{Tr}(Q\rho)Q\cdot d\cdot\mathbf{1}_{Q=\mathbb{I}} &(6.225)\\ &= d\mathrm{Tr}(\rho)\mathbb{I}, &(6.226)\end{aligned}$$

where we have used Lemma 6.6.1 to obtain the fourth equality. □

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

In this section, we simplify the algorithm of [FW20], and consider only the learning algorithm. In particular, we show that, if we do not 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 6.4.1 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 do not 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 6.7.1. *$\mathbb{P}_n$ 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 [BBRV02] and [WF89].

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

$$\mathbb{I} \mapsto (0,0) \in \mathbb{Z}_2^2, \tag{6.227}$$
$$X \mapsto (1,0) \in \mathbb{Z}_2^2, \tag{6.228}$$
$$Y \mapsto (1,1) \in \mathbb{Z}_2^2, \tag{6.229}$$
$$Z \mapsto (0,1) \in \mathbb{Z}_2^2 \tag{6.230}$$

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:

$$a \times b = a_1 b_2 + a_2 b_1 \mod (2). \tag{6.231}$$

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

$$a \times b = a_1 \times b_1 + \cdots + a_n \times b_n \mod (2). \tag{6.232}$$

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 $(\mathbb{Z}_2^2)^n \simeq (\mathbb{Z}_2^n)^2 : a \mapsto (\alpha|\beta) = (((a_1)_1, \ldots (a_n)_1)|((a_1)_2, \ldots (a_n)_2)$. 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 $\{G_i\}_{i=1}^{d+1} = \{\{\mathbb{1}\} \cup \mathcal{C}_i\}_{i=1}^{d+1}$ [BBRV02] where $\{\mathcal{C}_i\}_{i=1}^{d+1}$ are the sets of non identity Pauli operators corresponding to matrices of the form:

$$(0_n|\mathbb{I}_n), (\mathbb{I}_n|A_1), \ldots, (\mathbb{I}_n|A_d). \tag{6.233}$$

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 [BBRV02]. 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 $B_1, \ldots, B_n \in \mathbb{Z}_2^{n \times n}$ be n symmetric matrices. If we have for all $\alpha \in \mathbb{Z}_2^n \setminus \{0\}$:

$$\det \left(\sum_{i=1}^n \alpha_i B_i \right) \neq 0 \tag{6.234}$$

then we can choose $A_i = \sum_{k=1}^n i_k B_k$ where $i = (i_1, \ldots, i_n)$ 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 [WF89]. Let $f_1, \ldots, f_n$ be a basis of $\mathbb{F}_{2^n}$ as a vector space over $\mathbb{Z}_2$. Then we can write for all $i, j \in [n]$:

$$f_i f_j = \sum_{k=1}^n B_{i,j}^{(k)} f_k. \tag{6.235}$$

Then $B_k = \left(B_{i,j}^{(k)}\right)_{i,j\in[n]}$ satisfy the wanted condition [WF89]. We can verify this, let $\alpha \in \mathbb{Z}_2^n \setminus \{0\}$, we have:

$$\sum_{k=1}^{n} \alpha_k B_k = \left(\sum_k \alpha_k B_{i,j}^{(k)}\right)_{i,j\in[n]}. \tag{6.236}$$

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:

$$\sum_{k,i,j} y_i \alpha_k B_{i,j}^{(k)} x_j = 0. \tag{6.237}$$

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 6.7.2. *Let $G \in \{G_1, \ldots, G_{d+1}\}$. 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*

$$\frac{1}{|G|}\sum_{P\in G}(-1)^{P.Q} = \mathbf{1}\{Q \in C_G\}. \tag{6.238}$$

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:

$$\sum_{P\in G}(-1)^{P.Q} = \sum_{P\in C}(-1)^0 + \sum_{P\in A}(-1)^1 = |C| - |A| = 0. \tag{6.239}$$

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 6.7.3. *Let $G \in \{G_1, \ldots, G_{d+1}\}$. 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 $|G| = d$:

$$\rho_G^2 = \left(\frac{1}{d}\sum_{P\in G} P\right)^2 = \frac{1}{d^2}\sum_{P,Q\in G} PQ = \frac{1}{d^2}\sum_{R\in G}\sum_{P\in G:PR\in G} R = \frac{1}{d}\sum_{R\in G} R = \rho_G \tag{6.240}$$

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

$$(M_G^R)^2 = \frac{1}{d^2} \sum_{P,Q \in G} (-1)^{(P+Q).R} PQ = \frac{1}{d^2} \sum_{P \in G} \sum_{Q \in G : QP \in G} (-1)^{P.R} P \tag{6.241}$$

$$= \frac{1}{d} \sum_{P \in G} (-1)^{P.R} P = M_G^R \tag{6.242}$$

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

$$\sum_{Q \in A_G} M_G^Q = \sum_{Q \in A_G} \frac{1}{d} \sum_{P \in G} (-1)^{P.Q} P = \sum_{P \in G} \frac{1}{d} \sum_{Q \in A_G} (-1)^{P.Q} P \tag{6.243}$$

$$= \sum_{P \in G} \mathbf{1}\{P \in C_{A_G}\} P = \sum_{P \in G \cap \mathbb{P}_n/G} P = \mathbb{I}. \tag{6.244}$$

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

Now, we can state a simplified version of the algorithm proposed by [FW20].

Algorithm 10 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 $\|\mathcal{P} - \mathcal{R}\|_\diamond \leq \varepsilon$.

for $G \in \{G_1, \ldots, G_{d+1}\}$ **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 $Q_1, \ldots, Q_{N_G} \in A_G$.

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 6.7.1. *Algorithm 10 performs $\mathcal{O}(d^3 \log(d)/\varepsilon^2)$ measurements to learn a Pauli channel to within ε in diamond norm with at least a probability $2/3$.*

The proof is taken from [FW20].

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:

$$p_G = \{\mathrm{Tr}(M_G^Q \mathcal{P}(\rho_G))\}_{Q \in A_G} = \left\{ \sum_{P \in G} p(P + Q) \right\}_{Q \in A_G}, \tag{6.245}$$

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

$$\operatorname{Tr}(M_G^Q \mathcal{P}(\rho_G)) = \frac{1}{d^2} \sum_{P_1,P_3 \in G, P_2} (-1)^{P_1.Q} p(P_2) \operatorname{Tr}(P_1 P_2 P_3 P_2) \tag{6.246}$$

$$= \frac{1}{d^2} \sum_{P_1,P_3 \in G, P_2} p(P_2)(-1)^{P_1.Q+P_2.P_3} \operatorname{Tr}(P_1 P_3) \tag{6.247}$$

$$= \frac{1}{d} \sum_{P_1 \in G, P_2} p(P_2)(-1)^{P_1.(Q+P_2)} \tag{6.248}$$

$$= \sum_{P_2} p(P_2) \mathbf{1}(Q + P_2 \in C_G) \tag{6.249}$$

$$= \sum_{P \in C_G} p(P + Q). \tag{6.250}$$

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

$$\mathbb{E}\left((-1)^{Q.P}\right) = \sum_{Q \in A_G} p_G(Q)(-1)^{Q.P} = \sum_{Q \in A_G} \sum_{R \in C_G} p(R+Q)(-1)^{(Q+R).P} \tag{6.251}$$

$$= \sum_{S \in \mathbb{P}_n} p(S)(-1)^{S.P} = \hat{p}(P) \tag{6.252}$$

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 [Hoe63], 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:

$$q(P) = \frac{1}{d^2} \sum_{Q \in \mathbb{P}_n} (-1)^{Q.P} \hat{q}(Q) \tag{6.253}$$

so that we have by the Parseval–Plancherel identity:

$$d\|q - p\|_2 = \|\hat{q} - \hat{p}\|_2 = \sqrt{\sum_{P \in \mathbb{P}_n} (\hat{q} - \hat{p})^2} \leq \sqrt{\sum_{P \in \mathbb{P}_n} (\varepsilon/d)^2} = \varepsilon. \tag{6.254}$$

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:

$$\|r - p\|_2 \leq \|q - p\|_2 \leq \varepsilon/d. \tag{6.255}$$

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

$$\|r - p\|_1 \leq d\|r - p\|_2 \leq \varepsilon. \tag{6.256}$$

□

6.8 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 [CZSJ22] 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 chapter 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.

Bibliography

[AABB+19] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell, B. Burkett, Y. Chen, Z. Chen, B. Chiaro, R. Collins, W. Courtney, A. Dunsworth, E. Farhi, B. Foxen, A. Fowler, C. Gidney, M. Giustina, R. Graff, K. Guerin, S. Habegger, M. P. Harrigan, M. J. Hartmann, A. Ho, M. Hoffmann, T. Huang, Travis S. Humble, S. V. Isakov, E. Jeffrey, Z. Jiang, D. Kafri, K. Kechedzhi, J. Kelly, P. V. Klimov, S. Knysh, A. Korotkov, F. Kostritsa, D. Landhuis, M. Lindmark, E. Lucero, D. Lyakh, S. Mandrà, J. R. McClean, M. McEwen, A. Megrant, X. Mi, K. Michielsen, M. Mohseni, J. Mutus, O. Naaman, M. Neeley, C. Neill, M. Y. Niu, E. Ostby, A. Petukhov, J. C. Platt, C. Quintana, E. G. Rieffel, P. Roushan, N. C. Rubin, D. Sank, K. J. Satzinger, V. Smelyanskiy, K. J. Sung, M. D. Trevithick, A. Vainsencher, B. Villalonga, T. White, Z. J. Yao, P. Yeh, A. Zalcman, H. Neven, and J. M. Martinis. "Quantum supremacy using a programmable superconducting processor". In: *Nature* 574.7779 (2019), pp. 505–510. DOI: 10.1038/s41586-019-1666-5. URL: https://doi.org/10.1038/s41586-019-1666-5.

[Aar19] Scott Aaronson. "Shadow tomography of quantum states". In: *SIAM Journal on Computing* 49.5 (2019), STOC18–368.

[AB09] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

[ACMTBMAV07] Koenraad MR Audenaert, John Calsamiglia, Ramón Munoz-Tapia, Emilio Bagan, Ll Masanes, Antonio Acin, and Frank Verstraete. "Discriminating states: The quantum Chernoff bound". In: *Physical review letters* 98.16 (2007), p. 160501.

[AS17] Guillaume Aubrun and Stanisław J Szarek. *Alice and Bob meet Banach*. Vol. 223. American Mathematical Soc., 2017.

[ASWCC16] David R Anderson, Dennis J Sweeney, Thomas A Williams, Jeffrey D Camm, and James J Cochran. *Statistics for business & economics*. Cengage Learning, 2016.

[Aud14] Koenraad MR Audenaert. "Quantum skew divergence". In: *Journal of Mathematical Physics* 55.11 (2014), p. 112202.

[BAHL+10] Radoslaw C Bialczak, Markus Ansmann, Max Hofheinz, Erik Lucero, Matthew Neeley, Aaron D O'Connell, Daniel Sank, Haohua Wang, James Wenner, Matthias Steffen, et al. "Quantum process tomography of a universal entangling gate implemented with Josephson phase qubits". In: *Nature Physics* 6.6 (2010), pp. 409–413.

[BBRV02] Somshubhro Bandyopadhyay, P Oscar Boykin, Vwani Roychowdhury, and Farrokh Vatan. "A new proof for the existence of mutually unbiased bases". In: *Algorithmica* 34.4 (2002), pp. 512–528.

[BCL20] Sebastien Bubeck, Sitan Chen, and Jerry Li. "Entanglement is necessary for optimal quantum property testing". In: *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2020, pp. 692–703.

[BDKSSSS05] Igor Bjelaković, Jean-Dominique Deuschel, Tyll Krüger, Ruedi Seiler, Rainer Siegmund-Schultze, and Arleta Szkoła. "A quantum version of Sanov's theorem". In: *Communications in mathematical physics* 260.3 (2005), pp. 659–671.

[BH02] Richard J Bolton and David J Hand. "Statistical fraud detection: A review". In: *Statistical science* 17.3 (2002), pp. 235–255.

[BKGNMSM13] Robin Blume-Kohout, John King Gamble, Erik Nielsen, Jonathan Mizrahi, Jonathan D. Sterk, and Peter Maunz. *Robust, self-consistent, closed-form tomography of quantum logic gates on a trapped ion qubit*. 2013. DOI: 10.48550/ARXIV.1310.4492. URL: https://arxiv.org/abs/1310.4492.

[BLT23] Shrigyan Brahmachari, Josep Lumbreras, and Marco Tomamichel. "Quantum contextual bandits and recommender systems for quantum data". In: *arXiv preprint arXiv:2301.13524* (2023).

[BO21] Costin Bădescu and Ryan O'Donnell. "Improved Quantum Data Analysis". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. New York, NY, USA: Association for Computing Machinery, 2021, 1398–1411. ISBN: 9781450380539. URL: https://doi.org/10.1145/3406325.3451109.

[BOW19] Costin Bădescu, Ryan O'Donnell, and John Wright. "Quantum state certification". In: *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. 2019, pp. 503–514.

[BR15] Akshay Balsubramani and Aaditya Ramdas. "Sequential nonparametric testing with the law of the iterated logarithm". In: *arXiv preprint arXiv:1506.03486* (2015).

[BV94] Carl W Baum and Venugopal V Veeravalli. "A sequential procedure for multihypothesis testing". In: *IEEE Transactions on Information Theory* 40.6 (1994).

[Can20] Clément L Canonne. "A survey on distribution testing: Your data is big. but is it blue?" In: *Theory of Computing* (2020), pp. 1–100.

[CCHL22] Sitan Chen, Jordan Cotler, Hsin-Yuan Huang, and Jerry Li. "Exponential separations between learning with and without quantum memory". In: *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2022, pp. 574–585.

[CDVV14] Siu-On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. "Optimal algorithms for testing closeness of discrete distributions". In: *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*. SIAM. 2014, pp. 1193–1203.

[CHLL22] Sitan Chen, Brice Huang, Jerry Li, and Allen Liu. "Tight Bounds for Quantum State Certification with Incoherent Measurements". In: *arXiv preprint arXiv:2204.07155* (2022).

[CHLLS22] Sitan Chen, Brice Huang, Jerry Li, Allen Liu, and Mark Sellke. "Tight bounds for state tomography with incoherent measurements". In: *arXiv preprint arXiv:2206.05265* (2022).

[Cho75] Man-Duen Choi. "Completely positive linear maps on complex matrices". In: *Linear algebra and its applications* 10.3 (1975), pp. 285–290.

[CLO22] Sitan Chen, Jerry Li, and Ryan O'Donnell. "Toward instance-optimal state certification with incoherent measurements". In: *Conference on Learning Theory*. PMLR. 2022, pp. 2541–2596.

[CMS12] Benoit Collins, Sho Matsumoto, and Nadia Saad. "Integration of invariant matrices and application to statistics". In: *arXiv preprint arXiv:1205.0956* (2012).

[CN97] Isaac L Chuang and Michael A Nielsen. "Prescription for experimental determination of the dynamics of a quantum black box". In: *Journal of Modern Optics* 44.11-12 (1997), pp. 2455–2467.

[CŚ06] Benoît Collins and Piotr Śniady. "Integration with respect to the Haar measure on unitary, orthogonal and symplectic group". In: *Communications in Mathematical Physics* 264.3 (2006), pp. 773–795.

[CWLY22] Kean Chen, Qisheng Wang, Peixun Long, and Mingsheng Ying. "Unitarity estimation for quantum channels". In: *arXiv preprint arXiv:2212.09319* (2022).

[CZSJ22] Senrui Chen, Sisi Zhou, Alireza Seif, and Liang Jiang. "Quantum advantages for pauli channel estimation". In: *Physical Review A* 105.3 (2022), p. 032435.

[DDDSLWBW09] Uwe Dorner, Rafal Demkowicz-Dobrzanski, Brian J Smith, Jeff S Lundeen, Wojciech Wasilewski, Konrad Banaszek, and Ian A Walmsley. "Optimal quantum phase estimation". In: *Physical review letters* 102.4 (2009), p. 040403.

[DGKPP20] Ilias Diakonikolas, Themis Gouleakis, Daniel M Kane, John Peebles, and Eric Price. "Optimal Testing of Discrete Distributions with High Probability". In: *arXiv preprint arXiv:2009.06540* (2020).

[DGPP17] Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. "Optimal Identity Testing with High Probability". In: *arXiv preprint arXiv:1708.02728* (2017).

[DK16] Ilias Diakonikolas and Daniel M Kane. "A new approach for testing properties of discrete distributions". In: *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2016, pp. 685–694.

[DK17] Constantinos Daskalakis and Yasushi Kawase. "Optimal stopping rules for sequential hypothesis testing". In: *25th Annual European Symposium on Algorithms (ESA)*. 2017.

[DKW18] Constantinos Daskalakis, Gautam Kamath, and John Wright. "Which distribution distances are sublinearly testable?" In: *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, Philadelphia, PA, 2018, pp. 2747–2764. DOI: 10.1137/1.9781611975031.175. URL: https://doi.org/10.1137/1.9781611975031.175.

[DP01] GM D'Ariano and P Lo Presti. "Quantum tomography for measuring experimentally the matrix elements of an arbitrary quantum operation". In: *Physical review letters* 86.19 (2001), p. 4195.

[EFHKPVZ23] Andreas Elben, Steven T Flammia, Hsin-Yuan Huang, Richard Kueng, John Preskill, Benoît Vermersch, and Peter Zoller. "The randomized measurement toolbox". In: *Nature Reviews Physics* 5.1 (2023), pp. 9–24.

[EHWRMPCK20] J. Eisert, D. Hangleiter, N. Walk, I. Roth, D. Markham, R. Parekh, U. Chabaud, and E. Kashefi. "Quantum certification and benchmarking". In: *Nat. Rev. Phys.* 2.7 (2020), pp. 382–390. DOI: 10.1038/s42254-020-0186-4.

[EWLK+21] S. Ebadi, T. T. Wang, H. Levine, A. Keesling, G. Semeghini, A. Omran, D. Bluvstein, R. Samajdar, H. Pichler, W. W. Ho, S. Choi, S. Sachdev, M. Greiner, V. Vuletić, and M. D. Lukin. "Quantum phases of matter on a 256-atom programmable quantum simulator". In: *Nature* 595.7866 (2021), pp. 227–232. DOI: 10.1038/s41586-021-03582-4.

[Fan61] Robert M Fano. "Transmission of information: A statistical theory of communications". In: *American Journal of Physics* 29.11 (1961), pp. 793–794.

[FFDRG15] Lawrence M Friedman, Curt D Furberg, David L DeMets, David M Reboussin, and Christopher B Granger. *Fundamentals of clinical trials*. Springer, 2015.

[FFGO21] Omar Fawzi, Nicolas Flammarion, Aurélien Garivier, and Aadil Oufkir. "Sequential Algorithms for Testing Closeness of Distributions". In: *Advances in Neural Information Processing Systems* 34 (2021).

[FFGO22] Omar Fawzi, Nicolas Flammarion, Aurélien Garivier, and Aadil Oufkir. *Sequential algorithms for testing identity and closeness of distributions*. 2022. DOI: 10.48550/ARXIV.2205.06069. URL: https://arxiv.org/abs/2205.06069.

[FFGO23a] Omar Fawzi, Nicolas Flammarion, Aurélien Garivier, and Aadil Oufkir. "On Adaptivity in Quantum Testing". working paper or preprint. May 2023. URL: https://hal.science/hal-04107265.

[FFGO23b] Omar Fawzi, Nicolas Flammarion, Aurélien Garivier, and Aadil Oufkir. "Quantum Channel Certification with Incoherent Strategies". In: *arXiv preprint arXiv:2303.01188* (2023).

[FGLE12] Steven T Flammia, David Gross, Yi-Kai Liu, and Jens Eisert. "Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators". In: *New Journal of Physics* 14.9 (2012), p. 095022.

[FH18] Daniel Stilck França and AK Hashagen. "Approximate randomized benchmarking for finite groups". In: *Journal of Physics A: Mathematical and Theoretical* 51.39 (2018), p. 395302.

[FHT03] Alexei A Fedotov, Peter Harremoës, and Flemming Topsoe. "Refinements of Pinsker's inequality". In: *IEEE Transactions on Information Theory* 49.6 (2003), pp. 1491–1498.

[FO21] Steven T Flammia and Ryan O'Donnell. "Pauli error estimation via Population Recovery". In: *Quantum* 5 (2021), p. 549.

[FOF23] Omar Fawzi, Aadil Oufkir, and Daniel Stilck França. "Lower Bounds on Learning Pauli Channels". In: *arXiv preprint arXiv:2301.09192* (2023).

[FVDG99] Christopher A Fuchs and Jeroen Van De Graaf. "Cryptographic distinguishability measures for quantum-mechanical states". In: *IEEE Transactions on Information Theory* 45.4 (1999), pp. 1216–1227.

[FW20] Steven T Flammia and Joel J Wallman. "Efficient estimation of Pauli channels". In: *ACM Transactions on Quantum Computing* 1.1 (2020), pp. 1–32.

[GK19] Aurélien Garivier and Emilie Kaufmann. "Non-Asymptotic Sequential Tests for Overlapping Hypotheses and application to near optimal arm identification in bandit models". In: *arXiv preprint arXiv:1905.03495* (2019).

[GKKT20] Madalin Guţă, Jonas Kahn, Richard Kueng, and Joel A Tropp. "Fast state tomography with optimal error bounds". In: *Journal of Physics A: Mathematical and Theoretical* 53.20 (2020), p. 204001.

[Gro96] Lov K Grover. "A fast quantum mechanical algorithm for database search". In: *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*. 1996, pp. 212–219.

[GSLW19] András Gilyén, Yuan Su, Guang Hao Low, and Nathan Wiebe. "Quantum singular value transformation and beyond: exponential improvements for quantum matrix arithmetics". In: *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. 2019, pp. 193–204.

[Gu13] Yinzheng Gu. "Moments of random matrices and weingarten functions". PhD thesis. 2013.

[Haa33] Alfred Haar. "Der Massbegriff in der Theorie der kontinuierlichen Gruppen". In: *Annals of mathematics* (1933), pp. 147–169.

[Hay09] Masahito Hayashi. "Discrimination of two channels by adaptive methods and its application to quantum system". In: *IEEE transactions on information theory* 55.8 (2009), pp. 3807–3820.

[Hel69] Carl W Helstrom. "Quantum detection and estimation theory". In: *Journal of Statistical Physics* 1.2 (1969), pp. 231–252.

[HFW20] Robin Harper, Steven T Flammia, and Joel J Wallman. "Efficient learning of quantum noise". In: *Nature Physics* 16.12 (2020), pp. 1184–1188.

[HHJWY16] Jeongwan Haah, Aram W. Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. "Sample-optimal tomography of quantum states". In: *STOC'16—Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, 2016, pp. 913–925.

[HHLW10] Aram W Harrow, Avinatan Hassidim, Debbie W Leung, and John Watrous. "Adaptive versus nonadaptive strategies for quantum channel discrimination". In: *Physical Review A* 81.3 (2010), p. 032339.

[HKP20] Hsin-Yuan Huang, Richard Kueng, and John Preskill. "Predicting many properties of a quantum system from very few measurements". In: *Nature Physics* 16.10 (2020), pp. 1050–1057.

[HMRAR13] Michel Habib, Colin McDiarmid, Jorge Ramirez-Alfonsin, and Bruce Reed. *Probabilistic methods for algorithmic discrete mathematics*. Vol. 16. Springer Science & Business Media, 2013.

[Hoe63] Wassily Hoeffding. "Probability inequalities for sums of bounded random variables". In: *J. Amer. Statist. Assoc.* 58 (1963), pp. 13–30. ISSN: 0162-1459. URL: http://links.jstor.org/sici?sici=0162-1459(196303)58:301<13:PIFSOB>2.0.CO;2-D&origin=MSN.

[Hol73] Alexander S Holevo. "Statistical decision theory for quantum systems". In: *Journal of multivariate analysis* 3.4 (1973), pp. 337–394.

[HP91] Fumio Hiai and Dénes Petz. "The proper formula for relative entropy and its asymptotics in quantum probability". In: *Communications in mathematical physics* 143.1 (1991), pp. 99–114.

[HRMS18] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. "Uniform, nonparametric, non-asymptotic confidence sequences". In: *arXiv preprint arXiv:1810.08240* (2018).

[HRMS20] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. "Time-uniform Chernoff bounds via nonnegative supermartingales". In: *Probability Surveys* 17 (2020), pp. 257–317.

[HROWE22] Jonas Helsen, Ingo Roth, Emilio Onorati, Albert H Werner, and Jens Eisert. "General framework for randomized benchmarking". In: *PRX Quantum* 3.2 (2022), p. 020357.

[HXVW19] Jonas Helsen, Xiao Xue, Lieven MK Vandersypen, and Stephanie Wehner. "A new class of efficient randomized benchmarking protocols". In: *npj Quantum Information* 5.1 (2019), pp. 1–9.

[Iss18] Leon Isserlis. "On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables". In: *Biometrika* 12.1/2 (1918), pp. 134–139.

[Jam72] Andrzej Jamiołkowski. "Linear transformations which preserve trace and positive semidefiniteness of operators". In: *Reports on Mathematical Physics* 3.4 (1972), pp. 275–278.

[Jan97] Svante Janson. *Gaussian hilbert spaces*. 129. Cambridge university press, 1997.

[JDM00] Anil K Jain, Robert P. W. Duin, and Jianchang Mao. "Statistical pattern recognition: A review". In: *IEEE Transactions on pattern analysis and machine intelligence* 22.1 (2000), pp. 4–37.

[JM15] Michael I Jordan and Tom M Mitchell. "Machine learning: Trends, perspectives, and prospects". In: *Science* 349.6245 (2015), pp. 255–260.

[JP16] Anna Jenčová and Martin Plávala. "Conditions for optimal input states for discrimination of quantum channels". In: *Journal of Mathematical Physics* 57.12 (2016), p. 122203.

[KF15] Richard Kueng and Christopher Ferrie. "Near-optimal quantum tomography: estimators and bounds". In: *New Journal of Physics* 17.12 (2015), p. 123013.

[KK07] Richard M Karp and Robert Kleinberg. "Noisy binary search and its applications". In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. 2007, pp. 881–890.

[KKEG19] Martin Kliesch, Richard Kueng, Jens Eisert, and David Gross. "Guaranteed recovery of quantum processes from few measurements". In: *Quantum* 3 (2019), p. 171.

[KTCT21] Jonathan Kunjummen, Minh C Tran, Daniel Carney, and Jacob M Taylor. "Shadow process tomography of quantum channels". In: *arXiv preprint arXiv:2110.03629* (2021).

[LeC73] Lucien LeCam. "Convergence of estimates under dimensionality restrictions". In: *The Annals of Statistics* (1973), pp. 38–53.

[Leu00] Debbie Wun Chi Leung. *Towards robust quantum computation*. stanford university, 2000.

[Leu85] C Leubner. "Generalised Stirling approximations to N!" In: *European Journal of Physics* 6.4 (1985), p. 299.

[LHSS06] Nan-Ying Liang, Guang-Bin Huang, Paramasivan Saratchandran, and Narasimhan Sundararajan. "A fast and accurate online sequential learning algorithm for feedforward networks". In: *IEEE Transactions on neural networks* 17.6 (2006), pp. 1411–1423.

[LHT22a] Yonglong Li, Christoph Hirche, and Marco Tomamichel. "Sequential Quantum Channel Discrimination". In: *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2022, pp. 270–275.

[LHT22b] Josep Lumbreras, Erkka Haapasalo, and Marco Tomamichel. "Multi-armed quantum bandits: Exploration versus exploitation when learning properties of quantum states". In: *Quantum* 6 (2022), p. 749.

[Lid19] Daniel A Lidar. "Lecture notes on the theory of open quantum systems". In: *arXiv preprint arXiv:1902.00967* (2019).

[LN22] Angus Lowe and Ashwin Nayak. "Lower bounds for learning quantum states with single-copy measurements". In: *arXiv preprint arXiv:2207.14438* (2022).

[LTT22] Yonglong Li, Vincent YF Tan, and Marco Tomamichel. "Optimal adaptive strategies for sequential quantum hypothesis testing". In: *Communications in Mathematical Physics* (2022), pp. 1–35.

[LZK08] Daniel A Lidar, Paolo Zanardi, and Kaveh Khodjasteh. "Distance bounds on quantum dynamics". In: *Physical Review A* 78.1 (2008), p. 012308.

[Mat99] Pertti Mattila. *Geometry of sets and measures in Euclidean spaces: fractals and rectifiability*. 44. Cambridge university press, 1999.

[Mec19] Elizabeth S Meckes. *The random matrix theory of the classical compact groups*. Vol. 218. Cambridge University Press, 2019.

[MGE12] Easwar Magesan, Jay M Gambetta, and Joseph Emerson. "Characterizing quantum gates via randomized benchmarking". In: *Physical Review A* 85.4 (2012), p. 042311.

[MM13] Elizabeth Meckes and Mark Meckes. "Spectral measures of powers of random matrices". In: *Electronic communications in probability* 18 (2013).

[MRL08] Masoud Mohseni, Ali T Rezakhani, and Daniel A Lidar. "Quantum-process tomography: Resource analysis of different strategies". In: *Physical Review A* 77.3 (2008), p. 032322.

[MS86] Vitali D Milman and Gideon Schechtman. *Asymptotic theory of finite dimensional normed spaces*. Vol. 1200. Springer Berlin, 1986.

[MW13] Ashley Montanaro and Ronald de Wolf. "A survey of quantum property testing". In: *arXiv preprint arXiv:1310.2035* (2013).

[MWW09] William Matthews, Stephanie Wehner, and Andreas Winter. "Distinguishability of quantum states under restricted families of measurements with an application to quantum data hiding". In: *Communications in Mathematical Physics* 291.3 (2009), pp. 813–843.

[NC02] Michael A Nielsen and Isaac Chuang. *Quantum computation and quantum information*. 2002.

[NJ13] Mohammad Naghshvar and Tara Javidi. "Sequentiality and adaptivity gains in active hypothesis testing". In: *IEEE Journal of Selected Topics in Signal Processing* 7.5 (2013), pp. 768–782.

[NPPŻ18] Ion Nechita, Zbigniew Puchała, Łukasz Pawela, and Karol Życzkowski. "Almost all quantum channels are equidistant". In: *Journal of Mathematical Physics* 59.5 (2018), p. 052201.

[NS09] Michael Nussbaum and Arleta Szkoła. "The Chernoff lower bound for symmetric quantum hypothesis testing". In: *The Annals of Statistics* 37.2 (2009), pp. 1040–1057.

[OPGJLRW04] Jeremy L O'Brien, Geoff J Pryde, Alexei Gilchrist, Daniel FV James, Nathan K Langford, Timothy C Ralph, and Andrew G White. "Quantum process tomography of a controlled-NOT gate". In: *Physical review letters* 93.8 (2004), p. 080502.

[OT22] Yingkai Ouyang and Marco Tomamichel. "Learning quantum graph states with product measurements". In: *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2022, pp. 2963–2968.

[Ouf23] Aadil Oufkir. "Sample-Optimal Quantum Process Tomography with Non-Adaptive Incoherent Measurements". In: *arXiv preprint arXiv:2301.12925* (2023).

[OW15] Ryan O'Donnell and John Wright. "Quantum spectrum testing". In: *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. 2015, pp. 529–538.

[Pan08] Liam Paninski. "A coincidence-based test for uniformity given very sparsely sampled discrete data". In: *IEEE Transactions on Information Theory* 54.10 (2008), pp. 4750–4755.

[PCZ97] JF Poyatos, J Ignacio Cirac, and Peter Zoller. "Complete characterization of a quantum process: the two-bit quantum gate". In: *Physical Review Letters* 78.2 (1997), p. 390.

[PLLP19] Stefano Pirandola, Riccardo Laurenza, Cosmo Lupo, and Jason L Pereira. "Fundamental limits to quantum channel discrimination". In: *npj Quantum Information* 5.1 (2019), p. 50.

[RKKLGEK18] I. Roth, R. Kueng, S. Kimmel, Y.-K. Liu, D. Gross, J. Eisert, and M. Kliesch. "Recovering Quantum Gates from Few Average Gate Fidelities". In: *Phys. Rev. Lett.* 121 (17 2018), p. 170502. DOI: 10.1103/PhysRevLett.121.170502. URL: https://link.aps.org/doi/10.1103/PhysRevLett.121.170502.

[RRSRR73] Calyampudi Radhakrishna Rao, Calyampudi Radhakrishna Rao, Mathematischer Statistiker, Calyampudi Radhakrishna Rao, and Calyampudi Radhakrishna Rao. *Linear statistical inference and its applications*. Vol. 2. Wiley New York, 1973.

[Rup04] David Ruppert. *Statistics and finance: An introduction*. Vol. 27. Springer, 2004.

[RW09] Mark D Reid and Robert C Williamson. "Generalised pinsker inequalities". In: *arXiv preprint arXiv:0906.1244* (2009).

[RYTR22] Markus Rambach, Akram Youssry, Marco Tomamichel, and Jacquiline Romero. "Efficient quantum state tracking in noisy environments". In: *Quantum Science and Technology* 8.1 (2022), p. 015010.

[SHW22] Farzin Salek, Masahito Hayashi, and Andreas Winter. "Usefulness of adaptive strategies in asymptotic quantum channel discrimination". In: *Physical Review A* 105.2 (2022), p. 022419.

[SMPE+22] Roman Stricker, Michael Meth, Lukas Postler, Claire Edmunds, Chris Ferrie, Rainer Blatt, Philipp Schindler, Thomas Monz, Richard Kueng, and Martin Ringbauer. "Experimental single-setting quantum state tomography". In: *PRX Quantum* 3.4 (2022), p. 040310.

[SSKKG22] Trystan Surawy-Stepney, Jonas Kahn, Richard Kueng, and Madalin Guta. "Projected least-squares quantum process tomography". In: *Quantum* 6 (2022), p. 844.

[SSWE+21] P. Scholl, M. Schuler, H. J. Williams, A. A. Eberharter, D. Barredo, K. N. Schymik, V. Lienhard, L. P. Henry, T. C. Lang, T. Lahaye, A. M. Lauchli, and A. Browaeys. "Quantum simulation of 2D antiferromagnets with hundreds of Rydberg atoms". In: *Nature* 595.7866 (2021), pp. 233–238. DOI: 10.1038/s41586-021-03585-1.

[Tro12] Joel A Tropp. "User-friendly tail bounds for sums of random matrices". In: *Foundations of computational mathematics* 12.4 (2012), pp. 389–434.

[Ünl04] Tonguç Ünlüyurt. "Sequential testing of complex systems: a review". In: *Discrete Applied Mathematics* 142.1-3 (2004), pp. 189–205.

[VH14] Ramon Van Handel. *Probability in high dimension*. Tech. rep. PRINCETON UNIV NJ, 2014.

[VV16] Gregory Valiant and Paul Valiant. "Instance optimal learning of discrete distributions". In: *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. 2016, pp. 142–155.

[VV17] Gregory Valiant and Paul Valiant. "An automatic inequality prover and instance optimal identity testing". In: *SIAM Journal on Computing* 46.1 (2017), pp. 429–455.

[Wai19] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press, 2019.

[Wal44] Abraham Wald. "On cumulative sums of random variables". In: *The Annals of Mathematical Statistics* 15.3 (1944), pp. 283–296.

[Wal45] Abraham Wald. "Sequential tests of statistical hypotheses". In: *The annals of mathematical statistics* 16.2 (1945), pp. 117–186.

[Wan11] Guoming Wang. "Property testing of unitary operators". In: *Physical Review A* 84.5 (2011), p. 052328.

[Wat18] John Watrous. *The theory of quantum information*. Cambridge university press, 2018.

[WE16] Joel J Wallman and Joseph Emerson. "Noise tailoring for scalable quantum computation via randomized compiling". In: *Physical Review A* 94.5 (2016), p. 052325.

[WF89] William K Wootters and Brian D Fields. "Optimal state-determination by mutually unbiased measurements". In: *Annals of Physics* 191.2 (1989), pp. 363–381.

[WW70] AA Walters and AA Walters. *Statistical inference*. Springer, 1970.

[YFT19] Akram Youssry, Christopher Ferrie, and Marco Tomamichel. "Efficient online quantum state estimation using a matrix-exponentiated gradient method". In: *New Journal of Physics* 21.3 (2019), p. 033006.

[Yu20] Nengkun Yu. "Sample optimal Quantum identity testing via Pauli Measurements". In: *arXiv preprint arXiv:2009.11518* (2020).

[YW10] Neeley M. Lucero E. Bialczak R.C. Kelly J. Lenander M. Mariantoni M. O'Connell A.D. Sank D. Wang H. Yamamoto T. and M. Weides. "Quantum process tomography of two-qubit controlled-Z and controlled-NOT gates using superconducting phase qubits". In: *Physical Review B* 82.18 (2010), p. 184515.

[ZWDC+20] H. S. Zhong, H. Wang, Y. H. Deng, M. C. Chen, L. C. Peng, Y. H. Luo, J. Qin, D. Wu, X. Ding, Y. Hu, P. Hu, X. Y. Yang, W. J. Zhang, H. Li, Y. Li, X. Jiang, L. Gan, G. Yang, L. You, Z. Wang, L. Li, N. L. Liu, C. Y. Lu, and J. W. Pan. "Quantum computational advantage using photons". In: *Science* 370.6523 (2020), pp. 1460–1463. DOI: 10.1126/science.abe8770.

[ZZSE16] Shengjia Zhao, Enze Zhou, Ashish Sabharwal, and Stefano Ermon. "Adaptive concentration inequalities for sequential decision problems". In: *Advances in Neural Information Processing Systems* 29 (2016), pp. 1343–1351.