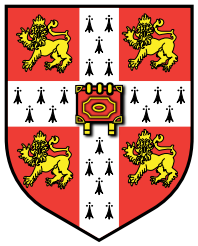
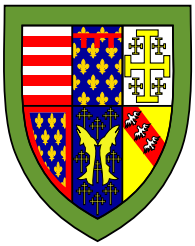
Learning in Quantum Mechanics

Wilfred Alan Salmon

Department of Applied Mathematics and Theoretical Physics
University of Cambridge

This thesis is submitted for the degree of Doctor of Philosophy

Queens' College

Aug 2024

Declarations

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted, or, is being concurrently submitted, for any degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

The background Sections of chapters 2, 3 and 4 contain a review of existing, relevant work. The work in Section 2.2, based off Ref. [1], was produced in collaboration with Dr. David Arvidsson-Shukur and Dr. Sergii Strelchuk, who provided helpful discussions. The work in Section 2.3.1, based off Ref. [2], was produced in collaboration with Dr. David Arvidsson-Shukur, Prof. Crispsin Barnes, and Flavio Silvati. The parts of the manuscript presented in this thesis were part of my contributions to the project. The work in Section 2.3.4, based off Ref. [3], was produced in collaboration with Dr. David Arvidsson-Shukur and Dr. Sergii Strelchuk, who provided helpful discussions. The work in Section 3.2, based off Ref. [4], was produced in collaboration with Dr. Sergii Strelchuk and Prof. Tom Gur, who provided helpful discussions. The work in Sections 4.2-4.4, based off Ref. [5], was produced in collaboration with Dr. Nadish de Silva and Minghua Yin. The algorithms were jointly created with Dr. de Silva, and Minghua, who also helped with the implementations.

Wilfred Salmon
Aug 2024

Abstract

This thesis explores the interactions of learning and quantum mechanics. At its heart, learning consists of extracting information from data. We will consider two types of data; random and deterministic. When data is random, one usually tries to learn an approximation to a desired object, when it is deterministic, one usually tries to learn an exact description of an object. Chapters 2 and 3 of this thesis focus on approximate learning of classical data embedded in quantum systems. The quantum mechanical nature of the physical systems leads to inherent randomness, even if the classical data is embedded in a deterministic way. In the final chapter we turn to exact learning of quantum objects, where we explore learning succinct characterisation of quantum objects from verbose descriptions.

Chapter 1 gives a brief motivation for quantum information, and gives a review of notation used throughout the thesis.

Chapter 2 introduces quantum metrology, the quantum generalisation of parameter estimation. Parameter estimation is a foundational area of modern statistics. In parameter estimation, one considers data generated from an instance of a parameterised model, with the aim of inferring the values of the parameters that generated the dataset. Parameter estimation is ubiquitous across modern science and thus has a rich and well-developed theory. Quantum metrology is a comparatively modern theory of parameter estimation, where classical parameters are encoded in quantum systems.

Existing literature in quantum metrology focuses on the so-called asymptotic regime, corresponding to understanding the limit of attainable precision as experimental resources (and time) tend to infinity. We consider the non-asymptotic regime, of particular relevance when experimental resources are limited; often the case for quantum systems. We generalise the classical framework of non-asymptotic parameter estimation to quantum metrology. This generalisation prescribes a non-asymptotic, operational method to determine if one measurement is better than another. We say that a measurement is admissible if there is no measurement that is strictly better than it; a measurement is optimal if it is as least as good as any other candidate measurement. Note that optimality is a much stronger condition than admissibility - there may be many *incomparable* admissible measurements, none of which are optimal. We prove several results within our non-asymptotic framework. First, we give a complete characterisation of when an optimal measurement exists - if and only if a parameterised state lies in a very restrictive class, called *classical parameterised states*. Second, we give three sufficient conditions for when an approximately optimal measurement exists, giving explicit bounds on the level of optimality. Finally, we give several necessary conditions, and one sufficient condition, for when a measurement is admissible. Given the fundamental nature of admissibility within classical parameter estimation, our results serve as a foundation for non-asymptotic quantum metrology.

Second, we explore an emerging technique in quantum metrology, known as *quantum filtering* (or postselection). Quantum filtering considers how one can distil information (on classical parameters) from many copies of a parameterised quantum state into few copies of a quantum state. Recently, a filter was discovered that can losslessly compress information into arbitrarily few quantum states, as long as one has a suitable initial guess of the parameter. First, we completely characterise which filters allow for lossless compression, showing that the known lossless filter lies within a more general family. Our discovery of the family of lossless filters allows for flexibility in the design of experiments. Second, we show that filters that are optimal for noiseless systems may be sub-optimal in the presence of noise. This counter-intuitive result shows the subtlety of filter design. Third, we show that lossless compression is not achievable in classical experiments (except in trivial cases), showing that lossless compression is a genuine quantum effect. The quantum nature of lossless compression has already been explored, by appealing to established indicators of classicality, rather than *operational* quantities. Our proof, however, is purely operational in nature, completely characterising the limited cases where classical lossless compression is possible. Finally, we give the first practical, iterative quantum filtering algorithm. All existing literature on filtering assumes the asymptotic regime, and does not give a concrete way to realise a practical advantage. Our algorithm is the first result in non-asymptotic quantum filtering. In our explicit scheme, we show that the quantum effect of filtering can have a subtle interplay with the classical theory of admissibility of estimators.

Chapter 3 considers probably approximately correct (PAC) learning, which underpins of all modern machine learning. We consider quantum machine learning of classical data, where classical objects are encoded in quantum states. Existing literature has demonstrated a wide range of quantum speedups in PAC learning, ranging from polynomial to exponential. However, all known improvements apply to special cases - not to the general theory. Indeed, it was recently shown that established access models do not admit a generic asymptotic quantum speedup. We consider a natural extension to the most widespread quantum access model, and argue it is applicable to most quantum scenarios of interest. We show that in this new access model, there is a generic quadratic reduction in the quantity of resources required for PAC learning. Furthermore, we prove that this quadratic improvement is optimal. Given the success of modern machine learning algorithms, our results pave the way for generic improvements in quantum machine learning.

Chapter 4 considers classical, exact learning of elements of the stabiliser formalism. The stabiliser formalism is a cornerstone of modern quantum computation (see Section 4.1.1 for a more thorough review). We explore common algorithmic primitives that are used for learning efficient classical descriptions of elements of the stabiliser formalism from inefficient ones. Such primitives have seen widespread use in classical simulation of quantum computers, and in increasing the efficiency of quantum circuit design. We give several new algorithms for these primitives, that have asymptotic and realisable speedups over existing implementations. Given the ubiquity of our primitives, these speedups will be widely useful in a range of problems.

"Do you guys just put the word quantum in front of everything?"
Scott Lang, Ant-Man and the Wasp, 2018

Acknowledgements

First, I must give thanks to both of my supervisors. Sergii, thank you for fighting for so many opportunities for me, for always being understanding, and always having your students' best interests at heart. David, thank you for everything you have done for me academically, for the hours of fun, and for your endless generosity. I am also greatly indebted to Normann Mertig, for his incredible commitment to the students and staff of the Hitachi Cambridge Laboratory.

I am forever grateful to my family. To my parents, Christopher and Gina, thank you for being the most loving and supporting parents I could ask for, and for pushing me throughout my childhood. I owe all my success to you. To my sisters, Clarissa and Josie, thank you for for being the best sisters in the world - you both inspire me. Finally, to my grandparents, Jeanine, Jean and Tony, thank you for making every moment with you special, and your unwavering support.

My time in Cambridge has been greatly enriched by the people I have met here. To Campbell, Mitchell and Josh - thank you for all of the fun-filled procrastination, the conference adventures and, from the Aussie pair, an enforced diet of Auntie Donna. Thanks to Florian, Toby and Miren for the countless silly lunchtime discussions, and other CMS activities. James - you inspired me as a teacher, and you are one of the kindest people I have ever met, thank you for being such a good friend. Will, you have been the best flatmate. Thank you for all of the biscuits, the late night chats and the Rick and Morty viewing parties. You have been a constant rock I could rely on whenever I have experienced difficulty. I am forever grateful that you chose to walk down the CC staircase to meet me!

During my undergraduate at Cambridge, I met the most amazing group of people. Thank you Hugh, George, Kay, Tomos, James, Carly, Ollie, Patrick, Guy, Beth, Zeb, Becca, Alice and Ed for all of the endless happy memories - I look forward to making many more.

Whenever I have had a difficult day at the office, I have always found "fun" online - thank you to Seb, Finley, Charlie, Matt and Bjarne for putting up with my terrible Dota gameplay.

Thank you to Nadish, for giving me the opportunity to spend four months in Vancouver - one the best experiences of my life. Thanks to Aggelos, Alex, Rebekah, Leo, and Archit for making my time there so fun, and to Michael for all the outlandish activities, spanning multiple continents. Finally, thank you to Ming, you are the bubbliest, golden-retrieverest person I have ever met (and super smart!).

Johanna, I don't have the words to describe how grateful I am that you are in my life. Thank you for everything - ich liebe dich.

Contents

Chapter 1

Introduction

1.1 Quantum Mechanics and Quantum Information

The theory of quantum mechanics runs contrary to our intuitive understanding of the universe. Niels Bohr, one of the founding fathers of the field, once said "Those who are not shocked when they first come across quantum theory cannot possibly have understood it" [6]. Moreover, the theory has remained fundamentally, metaphysically incomplete since its inception in the early 20^{th} century. For example, the theory predicts that observing a system changes is behaviour, but provides no satisfactory explanation of how a system should know it is being observed, or indeed a precise definition of "observation" [7]. Despite this, Quantum mechanics offers astonishingly accurate predictions, matching experimental observations in up to 13 decimal places of precision [8]. Undoubtedly, our universe is quantum.

Quantum mechanical effects have become increasingly important in microelectronics and computing. The fundamental building block of a computer is a transistor, a gate that can be toggled to allow electrons to, or not to, flow. As we make smaller and better transistors, we can make smaller and faster computers. Since the 1970s, we have observed roughly exponential scaling in the size of transistors, known as Moore's law [9], leading to roughly exponential growth in computing power. Yet, quantum mechanics gives a fundamental limit to how small transistors can be made (via. quantum tunneling [10]). Thus, if there is any hope to continue Moore's law, we must understand how to build computers out of quantum particles, that obey quantum effects.

The idea of utilising quantum systems for building computers long predates issues with transistors. Feynman first proposed the idea in a 1981 talk [11], famously concluding "Nature isn't classical, dammit, and if you want to make a simulation of Nature, you'd better make it quantum mechanical". Such observations have spawned an entire field of research: quantum information and computation.

Since its inception, the field of quantum information and computation has exploded. At its core, the field studies the utility of quantum mechanical systems for information theoretic (e.g. transmitting information securely between two parties) and computational (e.g. computing the product of two numbers) tasks. One considers whether the quantum system is "better" at the task than a system described by classical physics, or whether quantum effects are a hindrance. This thesis explores a variety of such tasks, finding quantum advantages and hindrances.

1.2 Notation

1.2.1 General

Let V, W be vector spaces. Unless otherwise stated, we work over the complex numbers. Denote the set of (bounded) linear maps from V to W by $\mathcal{B}(V, W)$. We write $\mathcal{B}(V, V) = \mathcal{B}(V)$. Let $\langle \cdot, \cdot \rangle$ be a choice of inner product on V. We let $\mathrm{Herm}(V)$ denote the set of Hermitian operators on V, and

$$S(V) = \{v \in V \mid \|v\| = 1\}. \tag{1.1}$$

For $A, B \in \mathcal{B}(V)$, we write $A \geq 0$ [resp. $A > 0$] to mean that A is positive semi-definite [resp. positive definite], with respect to $\langle \cdot, \cdot \rangle$. Similarly, we say that $A \geq B$ [resp. $A > B$] if $A - B \geq 0$ [resp. $A - B > 0$]. For $A \in \mathrm{Herm}(V)$, we call $\mathrm{Ker}(A)^{\perp}$ the support of A, denoted $\mathrm{supp}(A)$. We occasionally make use of the following properties of positive semi-definite matrices

Proposition 1.1: Suppose that V is an inner product space, $A \in \mathcal{B}(V)$ is positive semi-definite, $A \geq 0$. Then

(i) For $v \in V$, $Av = 0$ iff. $\langle v, Av \rangle = 0$.

(ii) For $v \in V$, $Av = 0$ iff. every eigenvector w of A with non-zero eigenvalue has $\langle v, w \rangle = 0$.

(iii) If w is an eigenvector of A with non-zero eigenvalue, then $w \in \mathrm{supp}(A)$.

Proof: (i) follows by noting $A = B^{\dagger}B$ for some $B \geq 0$. (ii) follows by writing A in its spectral decomposition, and using (i). (iii) follows from (ii). □

If V, W are dimension d, d' over a field $\mathbb{F}$ respectively, then $\mathcal{B}(V, W)$ is in bijection with the set of $d \times d'$ matrices over $\mathbb{F}$, denoted $\mathcal{M}_{d,d'}(\mathbb{F})$. We let $\mathcal{M}_d(\mathbb{F}) = \mathcal{M}_{d,d}(\mathbb{F})$.

Given two linear maps $A, B \in \mathcal{B}(V)$, we denote their commutator by $[A, B] = AB - BA$, and their anticommutator by $\{A, B\} = AB + BA$.

Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be some differentiable function. We denote its derivative by $Df : \mathbb{R}^m \to \mathcal{B}(\mathbb{R}^m, \mathbb{R}^n)$. In the case that $n = 1$ then $\mathcal{B}(\mathbb{R}^m, \mathbb{R})$ is in bijection with $\mathbb{R}^m$, and we let by $\nabla f : \mathbb{R}^m \to \mathbb{R}^m$, $\nabla f(x)_i = Df(x)(e_i)$.

For a natural number $K \in \mathbb{N} = \{1, 2, \dots\}$, we let $[K] = \{1, 2, \dots, K\}$. Given a set X with subset $A \subseteq X$, we denote the indicator function of A by $\mathbb{1}_A$:

$$\mathbb{1}_A : X \to \{0, 1\},\ \mathbb{1}_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases} \tag{1.2}$$

We let $A^c = X \setminus A$ denote the complement of A (in X).

Let $X \subseteq \mathbb{R}$, let $f, g : X \to \mathbb{R}$ and $y \in X$. We write

(i) $f = O_{x\to y}(g)$ if there exist constants $C, \delta > 0$ such that if $|x - y| \leq \delta$, then $f(x) \leq Cg(x)$.

(ii) $f = \Omega_{x\to y}(g)$ if there exist constants $C, \delta > 0$ such that if $|x - y| \leq \delta$, then $f(x) \geq Cg(x)$.

(iii) $f = \Theta_{x\to y}(g)$ if $f = O_{x\to y}(g)$ and $f = \Omega_{x\to y}(g)$.

We may also take $y = \infty$, in which case e.g. $f = O_{x\to\infty}(g)$ if there exist constants $C, L > 0$ such that if $|x| \geq L$, then $f(x) \leq Cg(x)$. It is common to write e.g. $f = O(g)$, with the precise limit y implicit. Most often f will be a function of a "small" parameter (e.g. ϵ) in which case $y = 0$, or an unbounded parameter (e.g. $n \in \mathbb{N}$), in which case $y = \infty$. These definitions can also be extended to $X \subseteq \mathbb{R}^k$, e.g. for $k = 2$, $f = O_{\epsilon\to 0,\, d\to\infty}(g)$ if there exist constants C, δ, L such that if $|\epsilon| \leq \delta$ and $d \geq L$, then $f(\epsilon, d) \leq Cg(\epsilon, d)$.

1.2.2 Quantum Theory

Let $\mathcal{H}$ be a Hilbert space. Unless otherwise stated, we will always take our Hilbert spaces to be finite dimensional. We denote the set of (quantum) states on $\mathcal{H}$ by $\mathcal{D}(\mathcal{H})$:

$$\mathcal{D}(\mathcal{H}) = \{\rho \in \mathcal{B}(\mathcal{H}) \mid \rho \geq 0, \mathrm{Tr}(\rho) = 1\}. \tag{1.3}$$

For two Hilbert spaces $\mathcal{H}, \mathcal{H}'$ we denote the set of completely-positive trace-preserving (CPTP) maps between them by $\mathcal{T}(\mathcal{H}, \mathcal{H}')$:

$$\mathcal{T}(\mathcal{H}, \mathcal{H}') = \{\Lambda \in \mathcal{B}(\mathcal{B}(\mathcal{H}), \mathcal{B}(\mathcal{H}')) \mid \Lambda \text{ is CPTP}\}. \tag{1.4}$$

For $K \in \mathbb{N}$ we let $\mathbb{M}_K(\mathcal{H})$ denote the set of K-outcome positive operator valued measures (POVMs) on $\mathcal{H}$:

$$\mathbb{M}_K(\mathcal{H}) = \left\{ M = (M_1, \ldots, M_K) \in \mathcal{B}(\mathcal{H})^K \;\middle|\; M_i \geq 0,\ \sum_{i=1}^{K} M_i = \mathbb{1} \right\}. \tag{1.5}$$

We let $\mathbb{M}(\mathcal{H})$ denote the disjoint union of the $\mathbb{M}_K(\mathcal{H})$, i.e. the set of POVMs with any number of allowed outcomes:

$$\mathbb{M}(\mathcal{H}) = \bigsqcup_{K\in\mathbb{N}} \mathbb{M}_K(\mathcal{H}) = \{(K, M) \mid K \in \mathbb{N}, M \in \mathbb{M}_K(\mathcal{H})\}. \tag{1.6}$$

When the number of outcomes of a POVM is unimportant, given $N = (K, M) \in \mathbb{M}(\mathcal{H})$, we will write N_i to mean M_i, i.e. directly index the elements of the POVM.

1.2.3 Statistics

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let X_n be a sequence of random variables on the probability space. We write $X_n \xrightarrow{a.s.} X$ to mean X_n converges almost surely to X, $X_n \xrightarrow{p} X$ to mean X_n converges in probability to X, and $X_n \xrightarrow{d} X$ to mean X_n converges in distribution to X.

For Ω finite, and $\mathcal{F}$ the power set of Ω, probability measures are in bijection with probability mass functions, i.e. functions $p : \Omega \to [0, 1]$ such that $\sum_{\omega\in\Omega} p(\omega) = 1$. We denote the set of probability mass functions on Ω by $\Delta(\Omega)$.

Chapter 2

Metrology

2.1 Background

2.1.1 Classical Parameter Estimation

In this Section, we introduce classical parameter estimation. Quantum metrology generalises parameter estimation, and thus the core ideas within parameter estimation are essential to quantum metrology. Indeed, all of our results will rely on the language introduced in this Section. The exposition is self-contained, and aims to motivate many of its core concepts. Naturally, it is far from exhaustive, and well-removed from the frontier of statistical research, but it covers the prerequisite background for understanding our results. We begin with some motivational material.

In physics, we are used to describing systems in terms of parameters: position, momentum, energy, temperature, pressure etc.. A finite collection of these parameters uniquely characterises a system; given knowledge of the parameters, we can predict the behaviour of the system in the future. Often, nature provides a system with unknown parameters, which we would like to determine. For example, "determine the temperature of this liquid" or "find the pressure of this gas". These are questions of "parameter estimation" [12], a foundational area of statistics. Parameter estimation problems can arise from such physical questions, but this is by no means exhaustive, as shown by the following example:

> **Example 2.1: Proportion estimation**
> Suppose there is a large population, with N members. One aims to estimate the proportion p that have a given property. For example p could be the fraction of an electorate that intend to vote for a given party, or the proportion of roads in the UK with a pothole. It is usually infeasible to sample the whole population, instead one takes a sample of $n \ll N$ members of the population, and determines the number of members X in the sample that have the property. If the sample is appropriately chosen, then X is binomially distributed: $X \sim B(n, p)$. Most commonly, the proportion p is estimated as $\hat{p} = X/n$.

We use this example to understand the more abstract theory of parameter estimation. A parameter θ (p in the example) is an element of a known (measurable) set $\Theta \subseteq \mathbb{R}^k$ ($[0, 1]$ in the example), called a parameter space. Often the case $k = 1$, called "single parameter", is more simple than the $k > 1$ "multiparameter" regime. One receives stochastic data depending on the parameter, corresponding to a random variable X, taking values in a measurable space χ ($[n]$ in the example), whose law P_θ

is unknown, but depends on the parameter in a known way (i.e. we know the map $\theta \mapsto P_\theta$, like the parameterised binomial distribution in the example). The map $\theta \mapsto P_\theta$ is called a model - it describes how the value of the parameter affects the data that we see. The model may arise from a physical theory (e.g. predicting how a gas at a given temperature and pressure will respond to an interaction), or other modelling assumptions (such as an appropriate sample being drawn in example 2.1). We say that a model P_θ is *identifiable* if $P_\theta = P_{\theta'} \Rightarrow \theta = \theta'$. Identifiablity appears as a weak assumption; without it, there are at least two values of the parameter that cannot be discriminated by observing X. However, we shall see that non-identifiable models naturally arise in quantum mechanics.

In the case that the random variable X takes value $x \in \chi$, one estimates the parameter, we denote the by estimate $\hat{\theta}(x)$ ($\hat{p}$ in the example). The (measurable) function $\hat{\theta} : \chi \to \mathbb{R}^k$ is called an estimator[1]. We consider the estimator $\hat{p}$ from example 2.1 in more detail. We find that $\mathbb{E}[\hat{p}(X)] = p$ and $\mathrm{Var}[\hat{p}(X)] = p(1-p)/n$. Thus, on average, $\hat{p}$ will give the correct answer for p, subject to random fluctuations, characterised by $\mathrm{Var}[\hat{p}(X)]$. We see that if p is close to 0 or 1, then $\hat{p}$ has small random fluctuations around the true value of p, whereas if $p \approx 1/2$, the random fluctuations of $\hat{p}$ are comparatively much larger. Moreover, the size of the fluctuations scales like $1/\sqrt{n}$ in the number of observations n. This illustrates two important points in parameter estimation

(i) The performance of an estimator $\hat{\theta}$ depends on the underlying parameter θ.

(ii) As more observations are taken, we can estimate θ more accurately.

We explore point (i) now, point (ii) is discussed in Section 2.1.3. Given an estimator $\hat{\theta}$ and a value of the parameter θ, $\hat{\theta}(X)$ is random variable over $\mathbb{R}^k$. We wish to quantify how "good" this random variable is, i.e. how "close" is it to the true parameter θ. First, we must define a quantative notion of distance. This is given by a function $L : \mathbb{R}^k \times \Theta \to [0, \infty]$, called a *loss function*. $L(y, \phi)$ quantifies how "bad" of a guess y is, if the underlying parameter were ϕ. We give several examples:

(i) Least-squares loss: $L(y, z) = \|y - z\|^2 := \sum_{i=1}^k (y_i - z_i)^2$.

(ii) Absolute loss: $L(y, z) = \|y - z\|$.

(iii) Discrete loss: $L(y, z) = 1 - \mathbb{1}_{\{y\}}(z)$.

(iv) Kullback-Leibler (KL) divergence [13]: If the parameters are probability distributions (i.e. parameters have non-negative entries and their entries sum to unity), and y is also a probability distribution, then $L(y, z) = \sum_{i=1}^k y_i \log z_i/y_i$. If y is not a probability distribution, $L(y, z) = \infty$.

Note that (i)-(iii) are metrics, whereas (iv) is not (it is not symmetric, and does not obey the triangle inequality). The choice of loss function is somewhat arbitrary; it should be motivated by the particular application of parameter estimation. Least-squares loss is the most common choice of loss function, given its fundamental connection to variance. Thus, unless stated otherwise, we will always take least-squares loss as our loss function.

Given a choice of loss function, the performance of the random variable $\hat{\theta}(X)$ for parameter value θ is characterised by the random variable $L(\hat{\theta}(X), \theta)$, which takes values in $[0, \infty]$. In order to compare

[1] It is sometimes helpful to allow estimators to take values outside of Θ. For example, it may be sensible to estimate the temperature of a system as zero. An estimator is called proper if it only takes values in Θ, otherwise it is called improper.

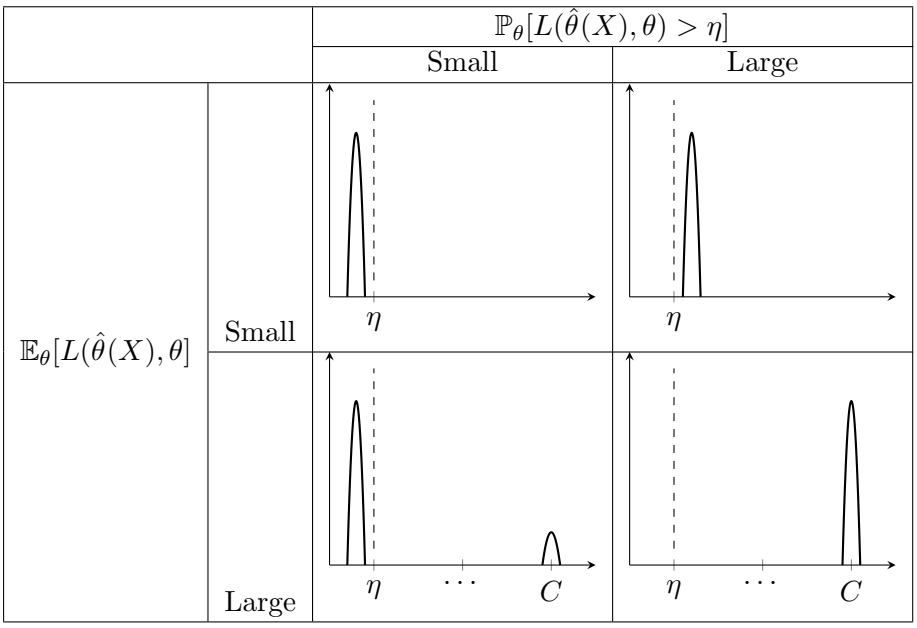


Figure 2.1: Contrasting conclusions that can be drawn from thresholding versus expectation. The plots show possible (schematic) pdfs of the random variable $L(\hat{\theta}(X),\theta)$. C represents a very large real number. Note that Markov's inequality implies that $\mathbb{P}_\theta[L(\hat{\theta}(X),\theta) > \eta] \leq \mathbb{E}_\theta[L(\hat{\theta}(X),\theta)]/\eta$, so that $\mathbb{P}_\theta[L(\hat{\theta}(X),\theta) > \eta]$ cannot be too large if $\mathbb{E}_\theta[L(\hat{\theta}(X),\theta)]$ is small. Conversely, as $C \to \infty$, $\mathbb{E}_\theta[L(\hat{\theta}(X),\theta)]$ can be made arbitrarily large whilst $\mathbb{P}_\theta[L(\hat{\theta}(X),\theta) > \eta]$ is simultaneously arbitrarily small.

estimators, we wish to quantify the performance of this random variable with a single number. There are two common methods:

(i) Expectation: The expected value $\mathbb{E}_\theta[L(\hat{\theta}(X),\theta)]$.

(ii) Thresholding: For a fixed (small) parameter value η, the probability $\mathbb{P}_\theta[L(\hat{\theta}(X),\theta) > \eta]$.

The subscript θ on $\mathbb{P}, \mathbb{E}$ is used to indicate that the random variable X has law P_θ. Taking the expectation and using thresholding can give contrasting conclusions on an estimator's performance, see figure 2.1 for a discussion. Thresholding is particularly common in theoretical computer science [14], most statistical works use least-squares loss. By introducing a new loss function $L'(y,\theta) = \mathbb{1}_{\{L(y,\theta)>\eta\}}(y)$, we note that $\mathbb{P}_\theta[L(\hat{\theta}(X),\theta) > \eta] = \mathbb{E}_\theta[L'(\hat{\theta}(X),\theta)]$ may also be written as an expectation of a loss function. However, in general, L' will not inherit any "nice" properties of L, such as continuity or convexity. In summary, the performance of an estimator for a given value of the parameter θ is quantified by $\mathbb{E}_\theta[L(\hat{\theta}(X),\theta)]$, where L is a choice of loss function.

Given an estimator $\hat{\theta}$, its performance across the whole parameter space Θ is measured by its risk function R, defined by

$$R(\hat{\theta},\cdot) : \Theta \to [0,\infty],\ R(\hat{\theta},\theta) = \mathbb{E}_\theta[L(\hat{\theta}(X),\theta)]. \tag{2.1}$$

The risk function gives us a way to compare estimators: we say that $\hat{\theta}_1 \leq \hat{\theta}_2$ if

$$\forall\theta \in \Theta,\ R(\hat{\theta}_1,\theta) \leq R(\hat{\theta}_2,\theta). \tag{2.2}$$

Note that $\leq$ is not generally a partial order, as we may have $R(\hat{\theta}_1,\cdot) = R(\hat{\theta}_2,\cdot)$ for distinct $\hat{\theta}_1 \neq \hat{\theta}_2$. It is important that two estimators are compared at all values of θ; before an experiment we do not

know the value of the parameter.

We write $\hat{\theta}_1 < \hat{\theta}_2$ if $\hat{\theta}_1 \leq \hat{\theta}_2$, but $\hat{\theta}_2 \not\leq \hat{\theta}_1$. If $\hat{\theta}_1 < \hat{\theta}_2$, then $\hat{\theta}_1$ is theoretically a better estimator than $\hat{\theta}_2$ - whatever the value of the parameter, one expects $\hat{\theta}_1$ to perform better[2]. An estimator $\hat{\theta}_2$ is said to be inadmissible if there exists another estimator $\hat{\theta}_1$ such that $\hat{\theta}_1 < \hat{\theta}_2$, otherwise it is said to be admissible. Most reasonable estimators will be incomparable via $\leq$; which is better will depend on the value of the parameter (see figure 2.2 for an illustration). If an estimator $\hat{\theta}$ were a minimum of $\leq$, then it would be a best estimator. However, in almost all problems of interest, no such minimal estimator exists.

Lemma 2.1: Suppose L is a loss function such that $L(\theta, \theta) = 0$ for all $\theta \in \Theta$. Then, if $\leq$ admits a minimal element $\hat{\theta}$, $R(\hat{\theta}, \cdot) \equiv 0$.

Proof: Fix $\theta_0 \in \Theta$. Define $\hat{\theta}^{\theta_0} \equiv \theta_0$, called a constant estimator, so that $R(\hat{\theta}^{\theta_0}, \theta_0) = 0$. But $\hat{\theta} \leq \hat{\theta}^{\theta_0}$, which implies that $R(\hat{\theta}, \theta_0) = 0$. As θ_0 was arbitrary, the result follows. □

If $R(\hat{\theta}, \cdot) \equiv 0$, the parameter estimation problem can always be solved exactly (from the perspective of the loss function L), in which case the underlying randomness is not important. Thus, in most cases of interest, there is no "best" estimator.

The worst-case risk of an estimator $R_{\max}(\hat{\theta})$, is defined by

$$R_{\max}(\hat{\theta}) = \sup_{\theta \in \Theta} R(\hat{\theta}, \theta). \tag{2.3}$$

One can also choose to compare estimators by their worst-case risk, giving parameter-agnostic guarantees. However, this can lead to sub-optimal performance across most of the parameter space, this is further discussed in figure 2.2. An estimator $\hat{\theta}^{\mathrm{m}}$ is called minimax if it minimises the worst case risk, i.e.

$$R_{\max}(\hat{\theta}^{\mathrm{m}}) = \inf_{\hat{\theta}} R_{\max}(\hat{\theta}). \tag{2.4}$$

We reiterate that $\hat{\theta}^{\mathrm{m}}$ is not (usually) a minimal element of $\leq$, other estimators will perform better depending on the value of the parameter.

The most common choice of estimator, studied for over 200 years [15] is the maximum-likelihood estimator (MLE). It can be used when P_θ admits a probability density function (pdf)[3] $f(x|\theta)$, with respect to some reference measure μ on χ. Usually, $\chi \subseteq \mathbb{R}^m$, and μ is the Lebesgue measure. The maxinum-likelihood estimator is defined by

$$\hat{\theta}^{\mathrm{MLE}}(x) = \arg\max_{\theta \in \Theta} f(x|\theta), \tag{2.5}$$

with ties broken arbitrarily. The optimisation is often written in terms of the (log-)likelihood function ℓ:

$$\ell(\cdot|x) : \Theta \to \mathbb{R},\ \ell(\theta|x) = \log f(x|\theta). \tag{2.6}$$

[2] In practice, there may be other considerations, such as computational cost or stability to noise.

[3] If χ is finite, then note that a probability density function (with respect to the counting measure) is a probability mass function.

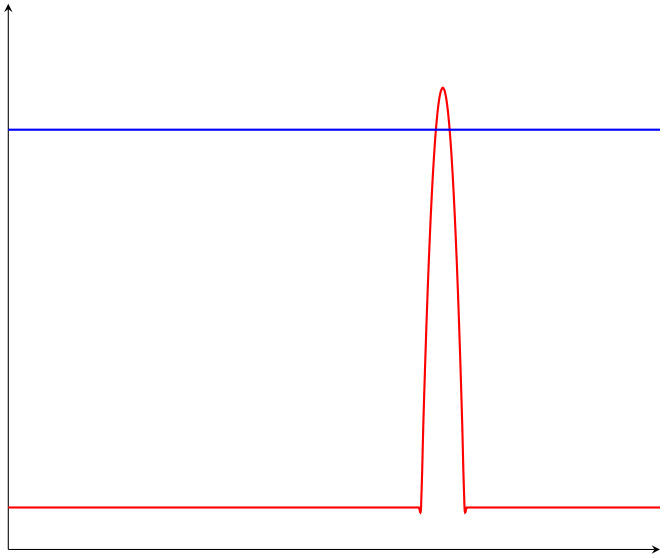


Figure 2.2: Schematic plots of two risk functions $R(\hat{\theta}, \cdot)$ for two different estimators, given by the red and blue curves. The blue estimator has smaller minimax risk than the red estimator, but has significantly worse performance across most of the parameter space. Moreoever, neither estimator dominates the other - they are incomparable via $\leq$.

Clearly, maximising ℓ is the same as maximising f. The MLE captures our intuition in parameter estimation: the parameter "probably" has a value that induces the highest possible probability of generating the data we have seen. Note that the estimator in example 2.1 is the MLE. Indeed, there are some situations in which the MLE is provably optimal (see Section 2.1.3). However, surprisingly, it is not always admissible, as shown by the following example.

Example 2.2: The James-Stein estimator [16–18]
Suppose that we are trying to estimate the mean $\theta \in \mathbb{R}^k$, where $k \geq 3$, of a Guassian (multivariate normal) distribution, with known covariance matrix $\Sigma > 0$, under least-squares loss. That is to say, $X \sim \mathcal{N}(\theta, \Sigma)$. The MLE $\hat{\theta}^{\mathrm{MLE}}(x) = x$ has risk

$$R(\hat{\theta}^{\mathrm{MLE}}, \theta) = \mathbb{E}_\theta[\|X - \theta\|^2] = \mathrm{Tr}(\Sigma). \tag{2.7}$$

The James-Stein estimator (JSE), $\hat{\theta}^{\mathrm{JS}}$ is defined by

$$\hat{\theta}^{\mathrm{JS}}(x) = x - \frac{(k-2)}{\|\Sigma^{-1}x\|^2}\Sigma^{-1}x, \tag{2.8}$$

with risk

$$R(\hat{\theta}^{\mathrm{JS}}, \theta) = \mathrm{Tr}(\Sigma) - (k-2)^2\mathbb{E}_\theta\left[\frac{1}{\|\Sigma^{-1}X\|^2}\right]. \tag{2.9}$$

Since the second term is clearly negative, we deduce that $\hat{\theta}^{\mathrm{JS}} < \hat{\theta}^{\mathrm{MLE}}$, i.e. the JSE dominates the MLE, and the MLE is inadmissible.

Lemma 2.1 illustrates the difficulty of statistics. In order to quantify the performance of estimator, one must choose an arbitrary loss function. Once the loss function is chosen, there is (usually) no subsequent best choice of estimator. Thus, extreme care must be taken in choosing a particular estimator, requiring extensive justification. Finally, example 2.2 shows that it is difficult to define general classes of (sensible) estimators; the MLE, one of the most pervasive and seemingly sensible

estimators, can be inadmissible.

2.1.2 Bayesian Statistics

In this Section we introduce Bayesian statistics. The two most common interpretations of probability are frequentism [19] and Bayesianism [20]. Frequentism, formally introduced in the mid 19th century [21], interprets probabilities as externally fixed, discoverable by averaging the results of many repetitions of the same experiment. Bayesianism, introduced in the late 18th century [22], however, views a probability as the quantification of one's belief, or expectation, that an event will occur. As one receives new information, one should update their belief, in accordance with Bayes' rule [23]. This thesis adopts a frequentest view of probability, but we will use techniques from Bayesian statistics to prove many of our results. Thus, we briefly introduce the mathematical machinery that we will require, particularly in Section 2.2.5.

In parameter estimation, Bayesianism amounts to a probability measure π on Θ (with σ-algebra given by the restriction of the standard σ-algebra of $\mathbb{R}^k$), representing one's belief about the value of the parameter, called the prior distribution. We will always assume that π admits a density (with respect to the Lebesgue measure). For notional brevity, we will also call this density π. Given a prior, we define the Bayes risk $R_\pi(\hat{\theta})$ of an estimator by

$$R_\pi(\hat{\theta}) = \mathbb{E}_{\vartheta\sim\pi}[R(\hat{\theta},\vartheta)] = \int d\theta\, \pi(\theta) R(\hat{\theta},\theta). \tag{2.10}$$

The Bayes risk quantifies the expected performance of $\hat{\theta}$ - if ones draws θ at random according to π, how well does one expect $\hat{\theta}$ to perform?

An estimator $\hat{\theta}^{\mathrm{B}}$ that minimises the Bayes risk is called Bayesian:

$$R_\pi(\hat{\theta}^{\mathrm{B}}) = \inf_{\hat{\theta}} R_\pi(\hat{\theta}). \tag{2.11}$$

It represents an optimal choice of estimator, *with respect to the prior* π.

Suppose X takes value x, then our belief about the parameter should be updated (via Bayes' rule) to give a new probability distribution Π over Θ, called the posterior distribution:

$$\Pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int d\theta\, f(x|\theta)\pi(\theta)}. \tag{2.12}$$

The posterior risk of an estimator, if x is observed, is given by

$$\int d\theta\, \Pi(\theta|x) L(\hat{\theta}(x),\theta). \tag{2.13}$$

An estimator $\hat{\theta}$ is Bayesian iff. it minimises the posterior risk [20]. Bayesian estimators are often admissible, as shown by the following result

Lemma 2.2: Let π be a prior such that the Bayes estimator $\hat{\theta}^{\mathrm{B}}$ is unique. Then $\hat{\theta}^{\mathrm{B}}$ is admissible

Proof: Suppose $\hat{\theta}^{\mathrm{B}}$ is inadmissible, then there exists an estimator $\hat{\theta} < \hat{\theta}^{\mathrm{B}}$. But then $R_\pi(\hat{\theta}) \leq R_\pi(\hat{\theta}^{\mathrm{B}})$ and $\hat{\theta}$ is also Bayes, contradicting the uniqueness of $\hat{\theta}^{\mathrm{B}}$. □

There are various generalisations of this result (see Ref. [24] theorems 7.13 and 8.7), which relax the condition of uniqueness.

The introduction of a prior leads to an (often unique and admissible) optimal estimator. However, one must have good reason to pick a particular prior π. If the parameter estimation problem has been repeated many times, then one can build up a suitable prior over time. However, if a parameter estimation problem is to be performed once, then the choice of prior π is somewhat arbitrary. If one tries to be prior-agnostic, by choosing the worst-case prior for a fixed estimator $\hat{\theta}$, one finds Bayes risk equal to the worst-case risk of $\hat{\theta}$:

$$\sup_{\pi} R_\pi(\hat{\theta}) = R_{\max}(\hat{\theta}), \tag{2.14}$$

yielding the same issues as discussed in figure 2.2.

2.1.3 The Asymptotic Limit

The (historically) most studied regime of parameter estimation is the asymptotic limit - where infinite observations are made. Correspondingly, the most commonly studied area of quantum metrology is also the asymptotic limit. Our results in Section 2.2 are decidedly non-asymptotic, but Section 2.3 relies on some asymptotic quantum theory. Thus, in the next two Sections we give a self-contained description of classical asymptotic statistical theory. We give more exposition than is strictly necessary for our results, in order to give the reader a better appreciation of the context of the theory. Of course, our review is far from exhaustive; see Ref. [25] for a thorough approach.

In the asymptotic limit, one usually takes $X = (X_1, \ldots, X_n)$ as a vector of independent, identically distributed (iid) random variables, each of whose law depends on the parameter in the usual way. Instead of a single estimator $\hat{\theta}$, one considers a family of estimators $\{\hat{\theta}_n : \chi^n \to \mathbb{R}^k\}$, one for each $n \in \mathbb{N}$. Often, the family $\{\hat{\theta}_n\}$ is also referred to as an estimator. The asymptotic limit is when $n \to \infty$, i.e. one is given infinite data.

The simplest desirable property of estimators is that they should get the "right answer" as $n \to \infty$, known as consistency. An estimator $\hat{\theta}_n$ is said to be strongly consistent at the point $\theta \in \Theta$ if

$$\hat{\theta}_n \xrightarrow{a.s.} \theta \text{ as } n \to \infty. \tag{2.15}$$

If the convergence is in probability, rather than a.s., it is said to be weakly consistent. An estimator is said to be strongly (weakly) consistent if it is strongly (weakly) consistent at all $\theta \in \Theta$.

In most reasonable scenarios, the MLE can be shown to be consistent. Since the MLE is defined for all n in a uniform sense, it is common to refer to the family $\hat{\theta}_n^{\mathrm{MLE}}$ by just $\hat{\theta}^{\mathrm{MLE}}$.

Lemma 2.3: Suppose that Θ is compact, the model P_θ is identifiable, and that $f(x|\theta)$ is "sufficiently regular", then $\hat{\theta}^{\mathrm{MLE}}$ is is weakly consistent

For a proof, and a discussion of possible regularity assumptions, see Ref. [25], Section 5.2. Under stronger assumptions, one can also show that the MLE is strongly consistent (see Ref. [26], theorem 2.5). Additionally, for a fixed prior π, one can show a version of consistency for Bayesian estimators.

Lemma 2.4: (*Doob's theorem*) Suppose that $\chi \subseteq \mathbb{R}^m$ for some m, with its Borel σ-algebra, and that the model P_θ is identifiable. Let π be some prior on Θ, then the Bayesian estimator $\hat{\theta}^{\mathrm{B}}$ is strongly consistent on a (measurable) subset $A \subseteq \Theta$, with $\pi(A) = 1$.

If there is a neighbourhood U of the true parameter θ such that $\pi(U) = 0$, then for all posteriors, $\Pi_n(U|x_1, \ldots, x_n) = 0$, and thus $\hat{\theta}^{\mathrm{B}}$ will not be consistent at θ. This shows the necessity of the π-a.s. part of Lemma 2.4. Aside from issues of null-sets of priors, Doob's theorem shows that given enough information one (almost always) recovers the correct parameter, regardless of one's initial belief. Thus, in the asymptotic limit, the arbitrary choice of prior is unimportant.

Consider the behaviour of $R(\hat{\theta}_n, \theta)$ as $n \to \infty$. It is certainly desirable that $R(\hat{\theta}_n, \theta) \to 0$ as $n \to \infty$, strongly related to consistency[4]. Moreover, inspired by the variance of sample means (for example, in example 2.1), we expect that $R(\hat{\theta}_n, \theta)$ should scale as $\Theta(1/n)$. For a fixed value of the parameter $\theta_0 \in \Theta$, We would like to determine the optimal prefactor, $O(\theta_0)$, in the $1/n$ scaling - giving a notion of an asymptotically optimal estimator. Given that the constant estimator $\hat{\theta}^{\theta_0}$ has $R(\hat{\theta}^{\theta_0}, \theta_0) = 0$, it is not useful to consider $\inf_{\hat{\theta}_n} nR(\hat{\theta}_n, \theta_0) \equiv 0$. Instead, $O(\theta_0)$ is commonly defined as a "local" minimax:

$$O : \Theta \to [0, \infty],\ O(\theta_0) = \inf_{\hat{\theta}_n} \lim_{\delta \to 0} \limsup_{n \to \infty} \sup_{\|\theta - \theta_0\| \leq \delta} nR(\hat{\theta}_n, \theta). \tag{2.16}$$

The limit $\delta \to 0$ makes O a local property of the model. Remarkably, this fearsome limit has an elegant characterisation for most models of interest, as discussed in the next Section.

2.1.4 The Fisher Information

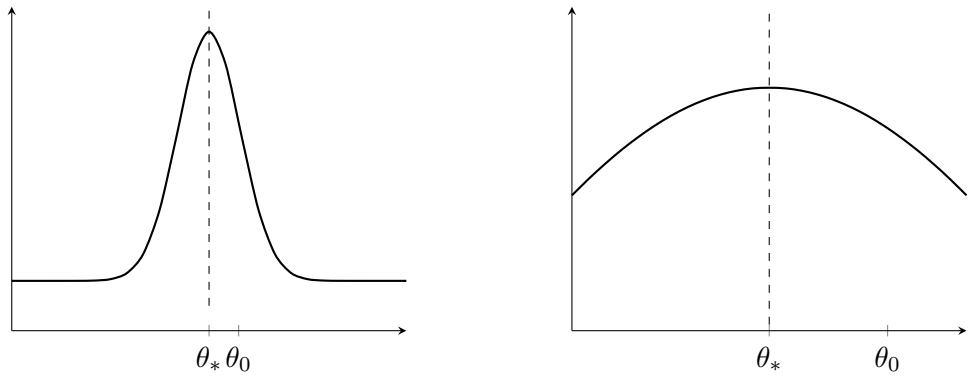
In this Section we introduce the Fisher information, the most important object in the classical asymptotic limit, finishing our exposition of classical statistics. We briefly motivate the Fisher information in a non-asymptotic sense, before relating it to the optimal prefactor O, introduced in the previous Section. We give several different operational interpretations of the Fisher information, with the aim of demonstrating its importance.

Consider a model P_θ that admits a pdf $f(x|\theta)$. Upon observing X, one intuitively expects the true parameter to be a maximum of the likelihood function ℓ (the motivation behind the MLE). Assuming that ℓ is differentiable, we define s, the score function

$$s(\cdot|x) : \Theta \to \mathbb{R}^k,\ s(\theta|x) = \nabla \ell(\theta|x). \tag{2.17}$$

One can calculate $\mathbb{E}_\theta[s(\theta|X)] = 0$, indeed showing that the true parameter is expected to be stationary point of the likelihood function. If ℓ has a very sharp peak, we can be confident that the true parameter is close to the maximum of ℓ. If ℓ is very flat, we can have little confidence in our estimate of the parameter. See figure 2.3 for an illustration of the two cases. Assuming ℓ is C^2, the "degree of

[4] In fact, if L is bounded, then this condition is implied by consistency.



(a) (b)

Figure 2.3: Possible (schematic) shapes of ℓ, the likelihood function. θ_* denotes the argmax of ℓ (i.e. the maximum likelihood estimator), the true parameter is θ_0. When the likelihood function is more sharply peaked (as in (a)), we can be more confident that θ_* is close to θ (than in (b)).

sharpness" of the peak is quantified by (minus) the second derivative (or curvature) of ℓ. This quantity is called the Fisher information (matrix):

$$I(\cdot|P) : \Theta \to \mathcal{M}_k(\mathbb{R}),\ I(\theta_0|P)_{ij} = -\mathbb{E}_{\theta_0}\left[\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\ell(\theta_0|X)\right]. \tag{2.18}$$

One can also show that $I(\theta|P) = \mathrm{Cov}_\theta[s(\theta|X)]$. When the model P is unambigous, we will write $I(\theta)$ instead of $I(\theta|P)$. $I(\theta|P)$ can be thought of as characterising the distinguishability of the law P_θ from $P_{\theta'}$, for θ' in an infinitesimal neighbourhood of θ.

The Fisher information elegantly captures many operational quantities in asymptotic statistics. Indeed, under suitable regularity assumptions, the Fisher information characterises the optimal prefactor O introduced in the previous Section:

Lemma 2.5: Suppose that Θ is compact, the model P_θ is identifiable, $f(x|\theta)$ is "sufficiently regular", θ_0 is an interior point of Θ, and that $I(\theta_0|P)$ is invertible. Then, $O(\theta_0) = \mathrm{Tr}\big[I(\theta_0|P)^{-1}\big]$.

For a proof, and discussion of regularity assumptions, see Ref. [25], theorem 8.11 and Section 8.9. Fewer regularity assumptions are needed to show that $O(\theta_0) \geq \mathrm{Tr}\big[I(\theta_0|P)^{-1}\big]$. The Fisher information has an additional asymptotic property:

Lemma 2.6: Suppose that that P_θ is "sufficiently regular". Further suppose, that $\hat{\theta}_n$ is an estimator such that for all $\theta \in \Theta$, $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in distribution. Then the set

$$\{\theta \mid \liminf_{n\to\infty} nR(\hat{\theta}_n, \theta) < \mathrm{Tr}\big[I(\theta|P)^{-1}\big]\}, \tag{2.19}$$

has Lebesgue measure 0.

See Ref. [25], theorem 8.9 for a proof. This shows that for sufficiently "nice" estimators, an improvement over $I(\theta|P)^{-1}$ can only be made on a null set.

Consider example 2.2, i.e. $X \sim \mathcal{N}(\theta, \Sigma)$. One can show that $I(\theta) = \Sigma^{-1}$ is independent of θ. If we take n samples, the sample mean has distribution $\overline{X}_n \sim \mathcal{N}(\theta, \Sigma/n)$, and $R(\hat{\theta}_n^{\mathrm{MLE}}, \theta) = \mathrm{Tr}(\Sigma)/n$. Aplying

the James-Stein estimator to the sample mean, we find that

$$R(\hat{\theta}_n^{\mathrm{JS}}, \theta) = \mathrm{Tr}(\Sigma)/n - \frac{(k-2)^2}{n^2}\mathbb{E}_\theta\left[\frac{1}{\left\|\Sigma^{-1}\overline{X}_n\right\|^2}\right], \tag{2.20}$$

For $\theta \neq 0$, the second term is $O(1/n^2)$, and thus $R(\hat{\theta}_n^{\mathrm{JS}}, \theta) = \mathrm{Tr}\big[I(\theta)^{-1}\big]/n + O(1/n^2)$. We consider the simplifying case $\Sigma = \sigma^2 \mathbb{1}_k$, for some $\sigma > 0$. We find that $R(\hat{\theta}_n^{\mathrm{JS}}, 0) = 2\sigma^2/n = (2/k)\,\mathrm{Tr}\big[I(\theta)^{-1}\big]/n$. This is an explicit example of an estimator whose risk scales no worse than the inverse Fisher information, yet scales better than the inverse Fisher information on a null set: $\{0\}$.

We give a final application of the Fisher information to non-asymptotic statistics. For an estimator $\hat{\theta}$, we define its bias $B_{\hat{\theta}}$ by

$$B_{\hat{\theta}} : \Theta \to \mathbb{R}^k,\ \theta \mapsto \mathbb{E}_\theta[\hat{\theta}(X)] - \theta. \tag{2.21}$$

One can bound $R(\hat{\theta}, \theta)$ in terms of $I(\theta)$ without reference to the asymptotic limit. In fact, one can bound the matrix $E(\hat{\theta}, \theta) := \mathbb{E}_\theta[(\hat{\theta}(X) - \theta)(\hat{\theta}(X) - \theta)^T]$, whose trace is equal to $R(\hat{\theta}, \theta)$:

Lemma 2.7: (Cramér-Rao (CR) bound)

$$E(\hat{\theta}, \theta) \geq B_{\hat{\theta}}(\theta)B_{\hat{\theta}}(\theta)^T + [\mathbb{1} + DB_{\hat{\theta}}(\theta)]I(\theta)^{-1}[\mathbb{1} + DB_{\hat{\theta}}(\theta)]^T. \tag{2.22}$$

Taking the trace of equation (2.22) gives a lower bound on $R(\hat{\theta}, \theta)$. An estimator is said to be locally unbiased at a point $\theta \in \Theta$ if $B_{\hat{\theta}}(\theta) = 0$ and $DB_{\hat{\theta}}(\theta) = 0$. In this case, the CR bound simplifies to $E(\hat{\theta}, \theta) \geq I(\theta)^{-1}$. The generic dependence of the CR bound on the bias of an estimator limits its practical applications.

We conclude the Section with the observation that under appropriate regularity assumptions, the MLE has optimal asymptotic scaling.

Lemma 2.8: Suppose that the model P_θ is identifiable, $f(x|\theta)$ is "sufficiently regular", θ_0 is an interior point of Θ and that $I(\theta_0|P)$ is invertible. Then

$$\sqrt{n}(\hat{\theta}_n^{\mathrm{MLE}} - \theta) \xrightarrow{d} \mathcal{N}(0, I(\theta_0|P)^{-1}). \tag{2.23}$$

Moreover, $nR(\hat{\theta}_n^{\mathrm{MLE}}, \theta) \to \mathrm{Tr}\big[I(\theta|P)^{-1}\big]$ for all $\theta \in \Theta$.

For a proof, see Ref. [25], Section 8.9.

2.1.5 Quantum Metrology

In this Section we introduce quantum metrology, the central field of this chapter. There are two main motivations for studying quantum metrology. Firstly, if one wishes build technology from sufficiently small systems, the systems will be described by the laws of quantum mechanics. Thus, if one wishes to measure any property of these systems (and thereby extract useful information), it is essential to understand the interaction of quantum mechanics and parameter estimation. Secondly, there have been theoretical and experimental demonstrations that quantum effects can lead to enhanced precision in parameter estimation [27]. We do not explicitly deal with these motivations, instead directly considering the extant, well-established theory. We will heavily rely on the language of classical statistics,

as introduced in the preceding Sections.

In quantum metrology, we explicitly describe how quantum systems depend on parameters [27, 28]. Instead of specifying a map $\theta \mapsto P_\theta$, we fix some Hilbert space $\mathcal{H}$, and consider a map

$$\rho : \Theta \to \mathcal{D}(\mathcal{H}). \tag{2.24}$$

The map ρ is called a parameterised (quantum) state. If the state ρ is always pure, it is usually described by a map

$$|\psi(\cdot)\rangle : \Theta \to S(\mathcal{H}), \tag{2.25}$$

which induces a corresponding $\rho(\cdot) = |\psi(\cdot)\rangle\langle\psi(\cdot)|$. Such a $|\psi(\cdot)\rangle$ is called a parameterised pure state. We give two examples of parameterised states.

Example 2.3: Gibbs state
Fix some Hamiltonian $H \in \mathrm{Herm}(\mathcal{H})$. We take

$$\rho_H : [0, \infty) \to \mathcal{D}(\mathcal{H}),\ \rho_H(\beta) = \frac{e^{-\beta H}}{\mathrm{Tr}(e^{-\beta H})}. \tag{2.26}$$

ρ_H (called a Gibbs state) corresponds to the thermal equilibrium of a system $\mathcal{D}(\mathcal{H})$ with Hamiltonian H at inverse temperature β.

Example 2.4: Phase-Encoded state
Fix $\mathcal{H} = \mathbb{C}^2$, the Hilbert space of a qubit. We take

$$|\psi(\cdot)\rangle : S^1 \to S(\mathcal{H}),\ e^{i\theta} \mapsto \frac{1}{\sqrt{2}}\left(|0\rangle + e^{i\theta}\,|1\rangle\right). \tag{2.27}$$

Such states arise in optics when using 2-arm interferometers. They may also be produced by the interaction of a two-level spin system with a magnetic field.

Given an unknown quantum state $\rho(\theta)$, we are tasked with estimating θ, as in the classical case. The most general quantum process one could use corresponds to a POVM $(K, M) \in \mathbb{M}(\mathcal{H})$. The choice of measurement induces a model $P_\theta^{(\rho;\, M)}$, which has a probability mass function on $[K]$:

$$p_\theta^{(\rho;\, M)} : [K] \to [0, 1],\ p_\theta^{(\rho;\, M)}(i) = \mathrm{Tr}[M_i \rho(\theta)]. \tag{2.28}$$

Thus, after the choice of measurement, quantum metrology reduces to classical parameter estimation. As before, one picks an estimator $\hat{\theta}^M : [K] \to \mathbb{R}^k$. After a choice of loss function, this estimator has a risk function $R_\rho(\hat{\theta}^M, \cdot)$. Usually, the parameterised state ρ is obvious from context, and we write $R(\hat{\theta}^M, \cdot)$. In summary, one specifies an approach to estimating the parameter by a measurement-estimator pair $(M, \hat{\theta}^M)$, where $M = (K, N) \in \mathbb{M}(\mathcal{H})$ and $\hat{\theta}^M : [K] \to \mathbb{R}^k$.

The relation $\leq$ on estimators immediately induces a relation $\leq$ on measurement-estimator pairs: $(M, \hat{\theta}^M) \leq (F, \hat{\theta}^F)$ iff. $\hat{\theta}^M \leq \hat{\theta}^F$. We emphasise that estimators that use different measurements can still be compared. As in the classical case (Lemma 2.1), we find that there is usually no optimal measurement-estimator pair:

Lemma 2.9: Suppose L is a loss function such that $L(\theta, \theta) = 0$ for all $\theta \in \Theta$. Then, if $\leq$ admits a minimal element $(M, \hat{\theta}^M)$, $R(\hat{\theta}^M, \cdot) \equiv 0$.

Proof: As in Lemma 2.1. □

If L is positive definite (i.e $L(y, z) = 0 \Rightarrow y = z$), then the existence of a minimal element of $\leq$ would imply that the states $\{\rho(\theta) \mid \theta \in \Theta\}$ can be discriminated with certainty, i.e. they are all orthogonal. This is an extremely restrictive condition, for example it is is impossible to satisfy if Θ is infinite and $\mathcal{H}$ is finite dimensional, or if ρ is continuous. As before, Lemma 2.9 demonstrates the difficulty of quantum metrology.

One can also consider Bayesian quantum metrology: given a prior π on the parameter space Θ, then, as in the classical case, a measurement-estimator pair $(M, \hat{\theta}^M)$ has some Bayes risk $R_\pi(\hat{\theta}^M)$. For a fixed measurement M, its Bayes risk is defined as the minimum possible Bayes-risk over choice of corresponding estimators $\hat{\theta}^M$:

$$R_\pi(M) = \inf_{\hat{\theta}^M} R_\pi(\hat{\theta}^M). \tag{2.29}$$

A measurement M [measurment-estimator pair $(M, \hat{\theta}^M)$] is called Bayes if it has minimal Bayes risk. In the single parameter case (and for least squares loss), it is easy to find Bayesian measurement-estimator pairs [29]. We discuss this further in proposition 2.7. We conclude with a remark that Ref. [30] has recently considered an interesting approach to thresholding in the quantum Bayesian regime. They show that optimisation of the measurement-estimator pair can be formulated as a semi-definite program, allowing for an efficient numerical solution. This approach leads to an intriguing new theory of both asymptotic and non-asymptotic quantum metrology, with the usual Bayesian drawback of an arbitrary choice of prior, or an oversimplification in terms of minimax risk.

2.1.6 Quantum Asymptotic Limit and the Quantum Fisher Information

As in the classical case, the most commonly considered regime of quantum metrology is the the asymptotic limit of infinite resources. As mentioned, our results do not heavily rely on this framework (though we do use it as motivation), but given its central importance to the field, we give a brief exposition of the main results of the theory. Asymptotic quantum metrology is significantly more complicated than its classical counterpart, and not fully understood.

In the quantum asymptotic limit, one considers a sequence of parameterised states $\rho_n : \Theta \to \mathcal{D}(\mathcal{H}^{\otimes n})$. We will only consider the case where there is some parameterised state σ, and $\rho_n = \sigma^{\otimes n}$, i.e. ρ_n is n copies of the state σ. We let $\rho = \rho_1$. As in the classical case, one considers a sequence of measurement-estimator pairs, $(M_n, \hat{\theta}^{M_n})$. One usually fixes θ and considers the behaviour of $R(\hat{\theta}^{M_n}, \theta)$ as $n \to \infty$. We will often refer to $\hat{\theta}^{M_n}$ as $\hat{\theta}_n$, leaving the choice of measurement implicit.

Given the classical characterisation of the asymptotic limit in terms of the Fisher information, initial attempts [31] at characterising the optimal behaviour of $R(\hat{\theta}^{M_n}, \theta)$ focused optimising the fisher information $I(\theta|P^{(\rho;\, M)})$ over choices of POVM M. As I is a positive semi-definite matrix, one cannot "maximise" the Fisher information directly (since most positive semi-definite matrices are incomparable). Instead, motivated by Lemma 2.5, one minimises $\mathrm{Tr}\big[I(\theta|P^{(\rho;\, M)})^{-1}\big]$. This gives rise to the most

informative bound [32]:

$$C^{\mathrm{MIB}}(\cdot|\rho):\Theta\to[0,\infty],\ C^{\mathrm{MIB}}(\theta|\rho)=\inf_{M\in\mathbb{M}(\mathcal{H})}\mathrm{Tr}\Big(I\left[\theta\Big|P^{(\rho;\,M)}\right]^{-1}\Big). \tag{2.30}$$

If $I\left[\theta|P^{(\rho;\,M)}\right]$ is not invertible, we adopt the convention $\mathrm{Tr}\Big(I\left[\theta\Big|P^{(\rho;\,M)}\right]^{-1}\Big)=\infty$.

The optimisation in C^{MIB} is, in general, difficult to perform. Thus, instead of calculating C^{MIB} directly, proxy functions are defined which are easier to calculate, and that lower bound C^{MIB}. We will discuss 3 such lower bounds, starting with the symmetric logarithmic derivative (SLD) quantum Fisher information (QFI) [31, 33].

Let $\partial_i\rho:\Theta\to\mathrm{Herm}(\mathcal{H})$ denote the i-th partial derivative of ρ. The i-th SLD is the map $L_i^{\mathrm{SLD}}:\Theta\to\mathrm{Herm}(\mathcal{H})$ implicitly defined by the equation

$$\frac{1}{2}\left\{\rho(\theta),L_i^{\mathrm{SLD}}(\theta)\right\}=\partial_i\rho(\theta). \tag{2.31}$$

The SLD QFI is defined as

$$J_{\mathrm{SLD}}(\cdot|\rho):\Theta\to\mathcal{M}_k(\mathbb{R}),\ J_{\mathrm{SLD}}(\theta|\rho)_{ij}=\mathrm{Re}\,\mathrm{Tr}\big[\rho(\theta)L_i^{\mathrm{SLD}}(\theta)L_j^{\mathrm{SLD}}(\theta)\big]. \tag{2.32}$$

We make two remarks:

1. Equation (2.31) may not have a solution, in which case we think of L_i^{SLD} as being "infinite".
2. The solution to (2.31) is, in general, non-unique. Let the orthogonal projection onto $\mathrm{Ker}\rho(\theta)$ be Π_θ. Let $\Pi_\theta^\perp=\mathbb{1}-\Pi_\theta$. Then $\Pi_\theta L_i^{\mathrm{SLD}}(\theta)\Pi_\theta$, $\Pi_\theta^\perp L_i^{\mathrm{SLD}}(\theta)\Pi_\theta$ and $\Pi_\theta L_i^{\mathrm{SLD}}(\theta)\Pi_\theta^\perp$ are uniquely determined by equation (2.31), whereas $\Pi_\theta^\perp L_i^{\mathrm{SLD}}(\theta)\Pi_\theta^\perp$ is arbitrary. One can remove this non-uniqueness by insisting that $\Pi_\theta^\perp L_i^{\mathrm{SLD}}(\theta)\Pi_\theta^\perp$ vanishes. Note that J_{SLD} does not depend on the choice of SLDs.

The right-logarithmic derivative (RLD) QFI is defined similarly [34, 35]. The i-th RLD is the map $L_i^{\mathrm{RLD}}:\Theta\to\mathrm{Herm}(\mathcal{H})$ implicitly defined by the equation

$$\partial_i\rho(\theta)=\rho(\theta)L_i^{\mathrm{RLD}}(\theta), \tag{2.33}$$

and the RLD QFI is defined as

$$J_{\mathrm{RLD}}(\cdot|\rho):\Theta\to\mathcal{M}_k(\mathbb{C}),\ J_{\mathrm{RLD}}(\theta|\rho)_{ij}=\mathrm{Tr}\Big[\rho(\theta)L_j^{\mathrm{RLD}}(\theta)L_i^{\mathrm{RLD}}(\theta)^\dagger\Big]. \tag{2.34}$$

Similar caveats to the definition of L_i^{SLD} also apply to L_i^{RLD}: a solution to equation (2.33) may not exist (we think of it as infinite) and $\Pi_\theta L_i^{\mathrm{RLD}}(\theta)$ is arbitrary (yet J_{RLD} is uniquely defined). One can also define a left logarithmic derivative (LLD), but taking the Hermitian conjugate of equation (2.33), we see that the LLD is the Hermitian conjugate of the RLD. Due to this redundancy, $J_{\mathrm{LLD}}=J_{\mathrm{RLD}}$.

L_i^{SLD} and L_i^{RLD} both attempt to generalise $\partial_i\ell(\theta|x)=\partial_i f(x|\theta)/f(x|\theta)$; the derivative of a pdf divded by that pdf. In the quantum case, $\partial_i\rho(\theta)$ and $\rho(\theta)$ may not commute, so there is no unique way to take their "ratio". They both directly bound the Fisher information of any measurement of ρ:

Lemma 2.10: Let $\rho : \Theta \to \mathcal{D}(\mathcal{H})$ be a C^2 parameterised quantum state. Then, $\forall \theta \in \Theta$ and $\forall M \in \mathbb{M}(\mathcal{H})$, we find that

$$I\left[\theta \middle| P^{(\rho;\, M)}\right]^{-1} \geq J_{\mathrm{SLD}}(\theta|\rho)^{-1} \quad \text{and} \quad I\left[\theta \middle| P^{(\rho;\, M)}\right]^{-1} \geq J_{\mathrm{RLD}}(\theta|\rho)^{-1}. \tag{2.35}$$

Moreover,

$$C^{\mathrm{MIB}}(\theta|\rho) \geq C^{\mathrm{SLD}}(\theta|\rho) := \mathrm{Tr}\big[J_{\mathrm{SLD}}(\theta|\rho)^{-1}\big], \tag{2.36}$$

$$C^{\mathrm{MIB}}(\theta|\rho) \geq C^{\mathrm{RLD}}(\theta|\rho) := \inf\{\mathrm{Tr}[X] \mid X \in \mathcal{M}_k(\mathbb{R}), X \geq J_{\mathrm{RLD}}(\theta|\rho)^{-1}\}, \tag{2.37}$$

$$= \mathrm{Tr}\big[\mathrm{Re}(J_{\mathrm{RLD}}(\theta|\rho)^{-1})\big] + \big\|\mathrm{Im}(J_{\mathrm{RLD}}(\theta|\rho)^{-1})\big\|_1. \tag{2.38}$$

See Ref. [32] for more detail. The bounds on C^{MIB} follow from taking the trace of equation (2.35). We consider one final bound on C^{MIB}, called the Holevo bound. Unlike the SLD and RLD bounds, it is explicitly a scalar bound - it does not bound $I\left[\theta \middle| P^{(\rho;\, M)}\right]$ directly. Fix $\theta \in \Theta$ and let

$$\mathbb{X}_\theta = \{(X_1, \ldots, X_k) \in \mathrm{Herm}(\mathcal{H})^k \mid \mathrm{Tr}[\partial_i \rho(\theta) X_j] = \delta_{ij}\}, \tag{2.39}$$

$$Z : \mathbb{X}_\theta \to \mathcal{M}_k(\mathbb{C}),\ Z[X]_{ij} = \mathrm{Tr}[\rho(\theta) X_i X_j]. \tag{2.40}$$

The Holevo bound is given by the optimisation

$$C^{\mathrm{H}}(\cdot|\rho) : \Theta \to [0, \infty],\ C^{\mathrm{H}}(\theta|\rho) = \inf_{X \in \mathbb{X}_\theta, U \in \mathcal{M}_k(\mathbb{R})} \{\mathrm{Tr}(U) \mid U \geq Z[X]\}. \tag{2.41}$$

The Holevo bound C^{H} directly bounds C^{MIB}, and is never worse than the SLD or RLD bounds:

Lemma 2.11: Let $\rho : \Theta \to \mathcal{D}(\mathcal{H})$ be a C^2 parameterised quantum state. Then, $\forall \theta \in \Theta$, we find that

$$C^{\mathrm{MIB}}(\theta|\rho) \geq C^{\mathrm{H}}(\theta|\rho), \tag{2.42}$$

$$\geq \max\{C^{\mathrm{SLD}}(\theta|\rho), C^{\mathrm{RLD}}(\theta|\rho)\}. \tag{2.43}$$

Additionally, $C^{\mathrm{H}}(\theta|\rho) \leq 2C^{\mathrm{SLD}}(\theta|\rho)$.

For more detail, see Ref. [32]. In general, C^{SLD} and C^{RLD} are incomparable (their relative size can change depending on the model). We see that C^{SLD} cannot be much worse than C^{H}, however C^{RLD} may be arbitrarily smaller than C^{SLD} and C^{H}. In general, C^{H} may be strictly smaller than C^{MIB}, see Ref. [36] for an example. However, when ρ is pure, or we work with single parameter states, many of the inequalities collapse:

Lemma 2.12: Let $\rho : \Theta \to \mathcal{D}(\mathcal{H})$ be a C^2 parameterised quantum state.

1. If $\Theta \subseteq \mathbb{R}$ (single parameter), then $\forall \theta \in \Theta$, we find that

$$C^{\mathrm{MIB}}(\theta|\rho) = C^{\mathrm{H}}(\theta|\rho) = C^{\mathrm{SLD}}(\theta|\rho). \tag{2.44}$$

2. If ρ is a pure parameterised state, then $\forall\theta \in \Theta$, we find that

$$C^{\mathrm{MIB}}(\theta|\rho) = C^{\mathrm{H}}(\theta|\rho). \tag{2.45}$$

See Refs. [32, 37] for a proof.

We note that C^{MIB} is not defined operationally, only in terms of the classical Fisher information. Operationally, one should consider the quantum generalisation O_Q of the local minimax bound O (defined in equation (2.16)):

$$O_Q : \Theta \to [0,\infty],\ O_Q(\theta_0) = \inf_{M\in\mathbb{M}(\mathcal{H})} \inf_{\hat{\theta}^{M_n}} \lim_{\delta\to 0} \limsup_{n\to\infty} \sup_{\|\theta-\theta_0\|\leq\delta} nR(\hat{\theta}_n,\theta). \tag{2.46}$$

There are three issues with attempting to apply Lemma 2.5 to the quantum case, in order to relate O_Q to C^{MIB}:

(i) Given $\rho_n = \sigma^{\otimes n}$, there is no requirement that one should measure the separate copies of σ with identical measurements (which would lead to iid samples). Indeed, one could collectively measure the whole ensemble ρ_n with a non-seperable POVM.

(ii) The set of POVMs that saturate C^{MIB} will generically depend on the unknown parameter θ.

(iii) For any given measurement $M \in \mathbb{M}(\mathcal{H})$, the model $P_\theta^{(\rho;\,M)}$ may not be identifiable.

In the pure state case, it has been shown that these issues do not apply: $O_Q = C^{\mathrm{H}}(= C^{\mathrm{MIB}})$, in agreement with the classical case (see Ref. [38], Section 5). A simpler scheme for attainability is given in Ref. [39], at the cost of assuming that for each θ one can pick a POVM $M_\theta \in \mathbb{M}(\mathcal{H})$ that saturates C^{MIB} and whose model $P_\theta^{(\rho;\,M_\theta)}$ is identifiable.

Somewhat surprisingly, recent work [37, 40–42] has found that if $\rho(\theta)$ is full rank for every θ, one can achieve risk scaling as $C^{\mathrm{H}}(\theta|\rho)/n$. As noted above, C^{H} may be strictly less than C^{MIB}, showing a deviation from the classical setting. Roughly speaking, their estimation protocol consists of two steps

1. Perform full state tomography on n^α copies of $\sigma(\theta)$, for some $\alpha \in (0,1)$. Tomography produces an estimate $\hat{\sigma}$ of σ, which can be used to get a rough estimate $\tilde{\theta}$ of θ.

2. Measure the remaining $n - n^\alpha$ copies of σ with a POVM, chosen using the intial estimate $\tilde{\theta}$. Use the outcome of this measurement to refine $\tilde{\theta}$ to a final estimate $\hat{\theta}$.

The first step aims to circumvent the issues (ii) and (iii) above. If the tomography is sufficiently accurate, one can essentially restrict θ to lie in a small volume, in which there is one POVM that approximately saturates C^{H}, and whose model $P_\theta^{(\rho;\,M)}$ is identifiable. However, it is not known what the asymptotic optimal prefactor is for states ρ of arbitrary rank.

Point (i) proves much harder to address. The Cramér-Rao bound (equation 2.22) lower bounds the risk of any locally-unbiased estimator in terms of C^{MIB}, but most estimators will not be locally unbiased, limiting practicability. J_{SLD} and J_{RLD} can be used in the full CR bound to bound the risk of any estimator, but the bias terms appearing in the bound limit its utility.

Alternatively, one can bound the asymptotic risk prefactor for various "nice" classes of estimation strategies. The Cramér-Rao bound can be used to bound the risk of asymptotically unbiased estimators [37, 40], i.e. those satisfying

$$\lim_{n\to\infty} B_{\hat{\theta}_n}(\theta),\ \lim_{n\to\infty} DB_{\hat{\theta}_n}(\theta) = 0. \tag{2.47}$$

However, it is not known whether asymptotically unbiased estimators exist for general parameterised quantum states [40], limiting the applicability of this bound.

A more complex class of boundable estimators was given in Ref. [40]. Fix $\theta, h \in \mathbb{R}^k$, and define a "local" sequence of states $\widetilde{\rho}_n(h) = \rho_n(\theta + h/\sqrt{n})$. Consider a measurement-estimator pair $(M_n, \hat{\theta}^{M_n})$. Define the sequence of random variables $X_n^{\theta,h} = \sqrt{n}(\hat{\theta}^{M_n}(Y_n) - \theta)$, where Y_n is a random variable denoting the outcome of measuring $\widetilde{\rho}_n(h)$ with M_n. If $\hat{\theta}^{M_n}(Y_n)$ estimates of $\theta + h/\sqrt{n}$, then $X_n^{\theta,h}$ estimates of the "local shift" h. We say that $(M_n, \hat{\theta}^{M_n})$ is locally asymptotically covariant (LAC) at θ if

1. For all $h \in \mathbb{R}^k$, $X_n^{\theta,h}$ converges in distribution to some random variable $X^{\theta,h}$.
2. For all $h \in \mathbb{R}^k$, $X^{\theta,h}$ has the same distribution as $h + X^{\theta,0}$.

If $(M_n, \hat{\theta}^{M_n})$ is LAC at θ, then $\liminf_{n\to\infty} nR(\hat{\theta}^{M_n}, \theta) \geq C^{\mathrm{H}}(\theta|\rho)$ (see Ref. [40], theorem 2). Moreover, the two step estimation process described above is LAC (see Ref. [40], theorem 10), and thus LAC estimators exist for a wide range of parameterised quantum states. We summarise these asymptotic statements in a single Lemma:

Lemma 2.13: Let $\rho : \Theta \to \mathcal{D}(\mathcal{H})$ be a "sufficiently regular" parameterised quantum state and suppose that Θ is compact. Then:

(i) If ρ is pure for every $\theta \in \Theta$, and θ_0 is an interior point of Θ, then $O_Q(\theta_0) = C^{\mathrm{H}}(\theta_0) = C^{\mathrm{MIB}}(\theta_0)$.

(ii) If ρ is full rank for every $\theta \in \Theta$, then there exists a measurement-estimator pair $(M_n, \hat{\theta}^{M_n})$ such that

$$\limsup_{n\to\infty} nR(\hat{\theta}^{M_n}, \theta) = C^{\mathrm{H}}(\theta). \tag{2.48}$$

(iii) If ρ is full rank for every $\theta \in \Theta$, and $\hat{\theta}_n$ is LAC, then for every $\theta \in \Theta$

$$\liminf_{n\to\infty} nR(\hat{\theta}_n, \theta) \geq C^{\mathrm{H}}(\theta). \tag{2.49}$$

We see that the asymptotic limit is only well-understood for pure and full-rank parameterised states. However, intermediate rank states are not understood. Moreover, a unifying framework for pure and full rank-states, as well as simple estimation strategies that achieve asymptotic limits are yet to be discovered.

2.1.7 Quantum Filtering

In this Section, we give an overview of quantum filtering, which is essential for all of our results in Section 2.3. Quantum filtering, a special case of a more general technique known as postselection, allows for enormous increases in parameter estimation accuracy, which have been practically demonstrated

[43]. The application of postselection to quantum metrology is a relatively new theory, and thus we give a short, self-contained overview that covers most existing results.

In some experimental systems (particularly in optics), one can produce states $\rho(\theta)$ (through the interaction of a probe and a system of interest) much faster than one can measure them [44, 45]. Thus, experimental resources are limited by the number of measurements one takes, not the number of copies of a parameterised state. Moreover, collective, entangled measurements on a large number of copies of the state are often practically infeasible, limiting measurements to single copies of $\rho(\theta)$.

Suppose, per unit time, one could produce N copies of $\rho(\theta)$, but can only measure $M \leq N$. Only preparing M states is equivalent to discarding a fraction $(1 - M/N)$ of the available information. Instead of discarding a random fraction, one could use a more complex "filter", distilling the information in many copies of $\rho(\theta)$ into a fewer states. Such a filtering technique was first described in Ref. [46], and further refined in Ref. [47].

A quantum filter is described by a two outcome measurement. Since the post-measurement state of the filter is important, we cannot describe the measurement by a POVM. Instead, we explicitly consider its two Kraus operators $K, \widetilde{K}$ satisfying $K^\dagger K + \widetilde{K}^\dagger \widetilde{K} = \mathbb{1}$. The corresponding POVM operators are given by $F = K^\dagger K$, $\widetilde{F} = \widetilde{K}^\dagger \widetilde{K} = \mathbb{1} - F$. Filtering corresponds to postselecting on the measurement outcome corresponding to K; states that have outcome $\widetilde{K}$ are blocked by the filter. Thus, the filter gives a new parameterised quantum state

$$\rho_K^{\mathrm{ps}} : \Theta \to \mathcal{D}(\mathcal{H}),\ \rho_K^{\mathrm{ps}}(\theta) = \frac{K\rho(\theta)K^\dagger}{\mathrm{Tr}(\rho(\theta)F)}. \tag{2.50}$$

In order to measure the information content in ρ^{ps}, it is common to consider $J_{\mathrm{SLD}}(\cdot|\rho^{\mathrm{ps}})$. For notational brevity, we shall shorten J_{SLD} to J when discussing quantum filtering. The QFI of the postselected state ρ_K^{ps} may be much larger than the QFI of the original state. However, the gain in information from filtering is always balanced by the small probability of successful postselection. Indeed, for a filter with Kraus operator K, we define the probability of successful postselection as

$$p_{\rho,\,K}^{\mathrm{ps}} : \Theta \to [0,1],\ p_{\rho,\,K}^{\mathrm{ps}}(\theta) = \mathrm{Tr}(\rho(\theta)F). \tag{2.51}$$

The rate of information arriving at the detector is thus given by

$$\mathcal{R}_{\rho,\,K} : \Theta \to \mathcal{M}_k(\mathbb{R}),\ \mathcal{R}_{\rho,\,K}(\theta) = p_{\rho,\,K}^{\mathrm{ps}}(\theta) J\left(\theta \mid \rho_K^{\mathrm{ps}}\right). \tag{2.52}$$

The factor of $p_{\rho,\,K}^{\mathrm{ps}}(\theta)$ in $\mathcal{R}_{\rho,\,K}$ accounts for states that are blocked by the filter; $\mathcal{R}_{\rho,\,K}$ measures the rate of information arriving at the detector. A version of the data-processing inequality [48] applied to the QFI says that it cannot be increased by CPTP maps, and thus that

$$\forall\theta \in \Theta,\ \mathcal{R}_{\rho,\,K}(\theta) \leq J(\theta|\rho). \tag{2.53}$$

We say that the filter K is lossless at the point $\theta \in \Theta$ if there is equality in equation (2.53). If $p_{\rho,\,K}^{\mathrm{ps}}(\theta) = p$ is fixed, then a lossless filter achieves the maximum possible information amplification, i.e. it maximises $J(\theta|\rho_K^{\mathrm{ps}})$ over all filters with $p_{\rho,\,K}^{\mathrm{ps}}(\theta) \geq p$.

Ref. [47] gave an example of a lossless filter for pure states: the Jenne Arvidsson-Shukur Lupu-Gladstein (JAL) filter. The JAL filter depends on a parameter $t \in (0, 1]$ and $\theta_0 \in \Theta$:

$$K^{\mathrm{JAL}}_{\theta_0} = (t-1)\,|\psi(\theta_0)\rangle\langle\psi(\theta_0)| + \mathbb{1}. \tag{2.54}$$

The performance of the JAL filter is summarised by the following Lemma.

Lemma 2.14: Let $\rho : \Theta \to \mathcal{D}(\mathcal{H})$ be a pure parameterised quantum state. The JAL filter leads to unbounded, lossless amplification of information at the point $\theta_0 \in \Theta$. In particular,

$$J(\theta_0|\rho^{\mathrm{ps}}_{K^{\mathrm{JAL}}_{\theta_0}}) = J\left(\theta_0 \mid \rho\right)/t^2, \quad p^{\mathrm{ps}}_{\rho,\,K^{\mathrm{JAL}}_{\theta_0}}(\theta_0) = t^2. \tag{2.55}$$

For a proof, see Ref. [47]. This remarkable unbounded amplification of J has been shown to be a genuinely quantum effect (see Section 2.3.3). There are two main drawbacks of the JAL filter:

(i) The amplification is only described in terms of the QFI J, not an operational quantity such as the risk of an estimator.

(ii) The JAL filter depends on the parameter to be estimated (similar in spirit to how a measurement that saturates C^{MIB} depends on the parameter).

Point two was partially addressed in Ref. [47], where they show that if θ_0 is sufficiently close to the true parameter $\theta \in \Theta$, the JAL filter is still approximately lossless:

Lemma 2.15: Let $\theta \in \Theta$, and let $\delta = \theta - \theta_0$, then

$$J(\theta_0|\rho^{\mathrm{ps}}_{K^{\mathrm{JAL}}_{\theta_0}}) = J\left(\theta_0 \mid \rho\right)/t^2 + O(\|\delta\|^2/t^2), \quad p^{\mathrm{ps}}_{\rho,\,K^{\mathrm{JAL}}_{\theta_0}}(\theta_0) = t^2 + O(\|\delta\|^2/t^2). \tag{2.56}$$

Lemma 2.15 shows that that $\|\delta\| \ll t$ is sufficient for good performance of the JAL filter. In general, it is also expected that this condition is necessary. The constraint $\|\delta\| \ll t$ may be interpreted in two ways. Firstly, one could consider a desired filter strength characterised by t, in which case one needs to estimate θ to accuracy $\sim t$ before one filter to the desired level. Alternatively, given a guarantee on $\|\delta\|$, the maximum possible filter strength is given by $\sim \|\delta\|$.

We conclude by noting that a proof-of-principle experiment using the JAL filter was recently carried out in Ref. [43]. They obtained an increase in information content by up to two orders of magnitude, demonstrating the practicality of the JAL filter. However, noise limitations meant that information amplification could not be made unbounded.

2.2 Non-Asymptotic Quantum Metrology

We have seen that the asymptotic limit allows one to define a meaningful of notion of optimality for measurements and estimators. However, its practicality is limited; in reality one receives a finite number of copies of a quantum state. Moreover, in many scenarios, the number of copies is severely limited - they are often very expensive to produce (or measure). Thus, the non-asymptotic regime, whilst mathematically inconvenient, is a closer match to experimental reality. In Section 2.2, we present a series of foundational results in non-asymptotic metrology (based off Ref. [1]), generalising many of the concepts from Section 2.1.1 to quantum metrology.

2.2.1 Comparing Measurements

To capture the non-asymptotic limit, it is enough to consider a single state $\rho : \Theta \to \mathcal{D}(\mathcal{H})$, and a single measurement estimator pair $(M, \hat{\theta}^M)$. As before, one can take $\rho = \sigma^{\otimes n}$ to include the case of multiple copies of a quantum state. In this setting, Lemma 2.9 tells us that there is no optimal choice of measurement *and* estimator. However, this does not preclude the possibility of an optimal measurement. As shown in Lemma 2.1, there is usually no best choice of estimator (and therefore trivially no best choice of measurement *and* estimator). In order to decide the existence of an optimal measurement, we must define a notion of optimality of measurements that is agnostic to a subsequent choice of estimator.

We begin by defining an operational relation $\preceq$ on measurements [1]. We say that a measurement $M \in \mathbb{M}(\mathcal{H})$ is as least as good as a measurement $F \in \mathbb{M}(\mathcal{H})$, written $M \preceq F$, if

$$\forall \hat{\theta}^F \; \exists \hat{\theta}^M \text{ s.t. } \; \hat{\theta}^M \leq \hat{\theta}^F. \tag{2.57}$$

If $M \preceq F$, then, from the perspective of risk, one should always change measurements from F to M. Whatever estimation strategy $\hat{\theta}^F$ was in use before, there is always a new one $\hat{\theta}^M$ that is at least as good as $\hat{\theta}^F$, regardless of the true underlying parameter. If $M \not\preceq F$, then there will be at least one estimation strategy $\hat{\theta}^F$ such that whatever new strategy $\hat{\theta}^M$ one picks, there are some values of the underlying parameter where one would expect to do worse. Thus, our definition is the natural and only way one can define M as being at least as good as F, without further assumptions on the estimation problem. We say that M dominates F, written $M \prec F$, if $M \preceq F$, but $F \not\preceq M$. In analogue with the classical case, we say that a measurement M is admissible if no other measurement dominates it. Clearly, as in the classical case, admissibility of a measurement is strongly desirable.

2.2.2 Characterising Optimal Measurements

In view of equation (2.57), there is a natural definition of an optimal measurement - a minimal element of $\preceq$. That is, we say that $M \in \mathbb{M}(\mathcal{H})$ is optimal if for all other possible measurements $F \in \mathbb{M}(\mathcal{H})$, $M \preceq F$. We emphasise that this is the strongest and most general definition of optimality one could make; any other definition must make some additional modelling assumptions (e.g., some *a priori* belief about the unknown parameter). Nonetheless, and somewhat surprisingly, optimal measurements do exist for a family of parameter estimation problems, as shown by the following Lemma.

Lemma 2.16: Let $\Theta \subseteq \mathbb{R}^k$ and let L be a choice of loss function that is convex in its first argument. Suppose that $\rho(\theta) = \sum_i p_\theta(i) |i\rangle\langle i|$, for some probability distribution $\{p_\theta(i)\}$ and a fixed, θ-independent, basis $\{|i\rangle\}$, then the measurement $M_i = |i\rangle\langle i|$ is optimal.

Proof: Suppose $F \in \mathbb{M}(\mathcal{H})$ is some other measurement with a choice of estimator $\hat{\theta}^F$. Fix $\theta \in \Theta$. Let $m_{k,i} = \mathrm{Tr}(F_k |i\rangle\langle i|)$. Note that $m_{k,i} \geq 0, \sum_k m_{k,i} = 1$ and $\sum_i p_\theta(i) m_{k,i} = p_\theta^{(\rho;\, F)}(k)$. Define an estimator using M by $\hat{\theta}^M(i) = \sum_k m_{k,i} \hat{\theta}^F(k)$. Then

$$R(\hat{\theta}^M, \theta) = \sum_i p_\theta(i) L\left(\sum_k m_{k,i} \hat{\theta}^F(k), \theta\right), \tag{2.58}$$

$$\leq \sum_{i,k} p_\theta(i) m_{k,i} L(\hat{\theta}^F(k), \theta) = R(\hat{\theta}^F, \theta). \tag{2.59}$$

Consequently, by swapping $\hat{\theta}^F$ to $\hat{\theta}^M$ we can only decrease risk (regardless of the value of the parameter) and thus $M \preceq F$. □

Parameterised states of the form $\rho(\theta) = \sum_i p_\theta(i) |i\rangle\langle i|$ are known as classical[5] [32]. The Gibbs state (example 2.3) is the canonical example of a classical parameterised state; Lemma 2.16 implies that measuring the energy of a Gibbs state is always optimal for estimating (inverse) temperature. The following condition shows that classical states may be alternatively characterised by commutation

Proposition 2.1: A paramaterised state $\rho : \Theta \to \mathcal{D}(\mathcal{H})$ is classical iff.

$$\forall \theta_1, \theta_2 \in \Theta,\ [\rho(\theta_1), \rho(\theta_2)] = 0, \tag{2.60}$$

Proof: Clearly, if ρ is classical, equation (2.60) holds. Suppose equation (2.60) holds, fix θ_1 in Θ and decompose $\mathcal{H}$ as a direct sum of $\rho(\theta_1)$'s (distinct, orthogonal) eigenspaces $\mathcal{H} = \oplus_j V_j$. If this is an eigenspace decomposition of $\rho(\theta)$ for every $\theta \in \Theta$, then by definition ρ is a classical state. Otherwise, take θ_2 such that some V_j is not an eigenspace of $\rho(\theta_2)$. But, $[\rho(\theta_1), \rho(\theta_2)] = 0$ and thus V_j must decompose a sum of eigenspaces of $\rho(\theta_2)$, $V_j = \oplus V_j'$, giving a refined decomposition of $\mathcal{H}$ into eigenspaces of $\rho(\theta_1)$ and $\rho(\theta_2)$. We may repeat this process, noting that it must terminate as each eigenspace must have dimension at least one, to see that $\rho(\theta)$ must be classical. □

Lemma 2.16 may lead one to hope that the issues of Lemma 2.9 all stem from the (classical) choice of estimator, and that there is always an optimal choice of (quantum) measurement. We extend example 2.4 to show that this is not the case

Example 2.5: No optimal measurement for the Phase-Encoded State

Recall the phase-encoded state is

$$|\psi(\theta)\rangle = \frac{1}{\sqrt{2}}(|0\rangle + e^{i\theta} |1\rangle), \tag{2.61}$$

[5]Classical is an unfortunately overloaded term in quantum information. Indeed, the states $|i\rangle$ appearing in the decomposition of ρ may be very "non-classical" according to other definitions, e.g. they could be very entangled, or have a high magic content. Parameterised states of the form $\rho(\theta) = \sum_i p_\theta(i) |i\rangle\langle i|$ could perhaps be better described as *classically encoded*, but this terminology is non-standard.

for $\theta \in [0, 2\pi)$. Consider a measurement M of the state in the $\{|+\rangle, |-\rangle\}$ basis. Define an estimator by $\hat{\theta}^M(+) = 0$, $\hat{\theta}^M(-) = \pi$. Note that $R(\hat{\theta}^M, 0) = R(\hat{\theta}^M, \pi) = 0$. Thus, any measurement that is as least as good as M must discriminate $|+\rangle$ and $|-\rangle$ (with certainty). Consider a different measurement F, with respect to the rotated basis

$$|e_1\rangle = \frac{1}{\sqrt{2}}(|0\rangle + e^{i\pi/4}|1\rangle),$$
$$|e_2\rangle = \frac{1}{\sqrt{2}}(e^{-i\pi/4}|0\rangle - |1\rangle), \tag{2.62}$$

and define an estimator $\hat{\theta}^F(1) = \pi/4$, $\hat{\theta}^F(2) = 3\pi/4$. As above, $R(\hat{\theta}^F, \pi/4) = R(\hat{\theta}^F, 3\pi/4) = 0$ and thus any measurement that is as least as good as F must discriminate $|e_1\rangle$ and $|e_2\rangle$. It is not difficult to show no measurement exists that simultaneously discriminates $|+\rangle, |-\rangle, |e_1\rangle$ and $|e_2\rangle$. Thus, no optimal measurement exists. We note that M and F saturate C^{MIB} (for all θ) [27].

In fact (for L the least square loss function), we show that the existence of an optimal measurement for ρ implies that it must be classical (under mild assumptions on the parameter space):

Lemma 2.17: Suppose that $\rho : \Theta \to \mathcal{D}(\mathcal{H})$ is some parameterised quantum state, where $\Theta \subseteq \mathbb{R}^k$ is convex. If ρ admits an optimal measurement (under least-squares loss), then it is classical.

We prove Lemma 2.17 in Section 2.2.5, it requires a long series of technical propositions. The Lemma shows that the difficulty of quantum metrology is (generically) twofold that of classical parameter estimation: there is no best choice of measurement and there is subsequently no best choice of estimator. It is important to note that not all measurements are good choices, only that there are many measurements that are incomparable.

We combine Lemmas 2.16 and 2.17 into a single theorem:

Theorem 2.1: Suppose that $\rho : \Theta \to \mathcal{D}(\mathcal{H})$ is some parameterised quantum state, where $\Theta \subseteq \mathbb{R}^k$ is convex. Then ρ admits an optimal measurement (under least-squares loss) iff. it is classical.

Theorem 2.1 provides a complete characterisation of when a parameterised quantum state admits an optimal measurement. It may be viewed as alternative generalisation of Lemma 2.1, and therefore it is a foundational result in non-asymptotic quantum metrology. It may be intuitively understood through pairwise distinguishablity of states. For $\sigma, \nu \in \mathcal{D}(\mathcal{H})$ quantum states, the optimal measurement to distinguish them [28] is with respect to the eigenbasis of $\sigma - \nu$. If ρ is a classical parametrised state, then for any $\theta_1, \theta_2 \in \Theta$ the eigenbasis of $\rho(\theta_1) - \rho(\theta_2)$ is the same, and thus the same measurement is optimal for distinguishing all possible pairs of states. Since it is pairwise optimal, it is intuitive that this measurement is then "globally" optimal for parameter estimation. Conversely, if such a globally optimal measurement exists, it must intuitively also be "pairwise" optimal.

2.2.3 Approximately Optimal Measurements

The previous Section shows that optimal measurements (under least-squares loss) only exist for the very restrictive class of classical parametrised states. It is natural to ask whether this result is robust: does an approximately optimal measurement imply that a state is approximately classical? Does an

approximately classical state imply the existence of an approximately optimal measurement? In this Section we show the latter of these to hold, i.e. we give a robust version of Lemma 2.16.

There are three clear candidates for when a measurement is approximately optimal:

(i) A measurement M is ϵ-additively optimal if, for any other measurement-estimator pair $(F, \hat{\theta}^F)$ there exists an estimator $\hat{\theta}^M$ such that for all $\theta \in \Theta$,

$$R(\hat{\theta}^M, \theta) \leq R(\hat{\theta}^F, \theta) + \epsilon. \tag{2.63}$$

(ii) A measurement M is η-multiplicatively optimal if, for any other measurement-estimator pair $(F, \hat{\theta}^F)$ there exists an estimator $\hat{\theta}^M$ such that for all $\theta \in \Theta$,

$$R(\hat{\theta}^M, \theta) \leq (1 + \eta) R(\hat{\theta}^F, \theta). \tag{2.64}$$

(iii) A measurement is δ-locally optimal if, for any other measurement-estimator pair $(F, \hat{\theta}^F)$ there exists an estimator $\hat{\theta}^M$ and subset $S \subseteq \Theta$ such that S has measure less than or equal to δ and for all $\theta \in S^c$,

$$R(\hat{\theta}^M, \theta) \leq R(\hat{\theta}^F, \theta). \tag{2.65}$$

We can combine the third definition with either of the first two. For example, a measurement is ϵ-additively and δ-locally optimal if, for any other measurement-estimator pair $(F, \hat{\theta}^F)$ there exists an estimator $\hat{\theta}^M$ and subset $S \subseteq \Theta$ such that S has measure less than or equal to δ and for all $\theta \in S^c$,

$$R(\hat{\theta}^M, \theta) \leq R(\hat{\theta}^F, \theta) + \epsilon. \tag{2.66}$$

For each of the three definitions of approximate optimality, there is a corresponding notion of "closeness" of quantum states such that if ρ is close to being classical, there is an approximately optimal measurement: approximately classical implies an approximately optimal measurement.

We start with additive optimality, considering closeness in trace norm: for any matrix A, define its trace norm as

$$\|A\|_1 = \mathrm{Tr}\left(\sqrt{A^\dagger A}\right). \tag{2.67}$$

We denote $\sqrt{A^\dagger A}$ as $|A|$. We show that closeness in trace norm implies an approximately additively optimal measurement:

Lemma 2.18: Suppose $\rho : \Theta \to \mathcal{D}(\mathcal{H})$ is some parametrised state and $\sigma : \Theta \to \mathcal{D}(\mathcal{H})$ is some classical state. Suppose further that Θ has finite diameter $d = \sup_{\theta,\phi} L(\theta, \phi)$. If, for all $\theta \in \Theta$, $\|\rho(\theta) - \sigma(\theta)\|_1 \leq \epsilon$, then there is a $(2d\epsilon)$-additively optimal measurement M.

Proof: Since σ is classical, we can fix some optimal measurement M. Fix another measurement

$F \in \mathbb{M}(\mathcal{H})$, estimator $\hat{\theta}^{\mathrm{F}}$ and $\theta \in \Theta$. Then note that

$$|R_\rho(\hat{\theta}^F, \theta) - R_\sigma(\hat{\theta}^F, \theta)| = \left| \sum_i \mathrm{Tr}[F_i(\rho[\theta] - \sigma[\theta])] L(\hat{\theta}^F(i), \theta) \right|, \tag{2.68}$$

$$\leq d \sum_i |\,\mathrm{Tr}[F_i(\rho[\theta] - \sigma[\theta])]|. \tag{2.69}$$

Fix some $F_i \geq 0$. Diagonalising $\rho(\theta) - \sigma(\theta) = \sum_j \lambda_j |j\rangle\langle j|$, note that

$$|\,\mathrm{Tr}[F_i(\rho(\theta) - \sigma(\theta))]| \leq \sum_j |\lambda_j| \langle j| F_i |j\rangle, \tag{2.70}$$

$$= \mathrm{Tr}(F_i |\rho(\theta) - \sigma(\theta)|). \tag{2.71}$$

Substituting inequality (2.71) into (2.69), we see that

$$|R_\rho(\hat{\theta}^F, \theta) - R_\sigma(\hat{\theta}^F, \theta)| \leq d\epsilon. \tag{2.72}$$

Since M is optimal with respect to σ there exists $\hat{\theta}^M$ such that for all θ, $R_\sigma(\hat{\theta}^M, \theta) \leq R_\sigma(\hat{\theta}^F, \theta)$. Then applying (2.72) twice, we see

$$R_\rho(\hat{\theta}^M, \theta) - R_\rho(\hat{\theta}^F, \theta) \leq R_\sigma(\hat{\theta}^M, \theta) - R_\sigma(\hat{\theta}^F, \theta) + 2d\epsilon, \tag{2.73}$$

$$\leq 2d\epsilon. \tag{2.74}$$

□

For multiplicative robustness, we need a different notion of closeness, namely the maximum relative entropy [49]. For quantum states $\rho, \sigma \in \mathcal{D}(\mathcal{H})$, we define

$$D_{\max}(\rho||\sigma) = \inf\{\gamma : \rho \leq e^\gamma \sigma\}. \tag{2.75}$$

The maximum relative entropy has the following useful property (relating it to the measured maximum relative entropy):

Proposition 2.2: For two states ρ, σ

$$D_{\max}(\rho||\sigma) = \log \sup_{0 \leq M \leq \mathbb{1}} \frac{\mathrm{Tr}(M\rho)}{\mathrm{Tr}(M\sigma)}. \tag{2.76}$$

We omit the proof, it is given in Ref. [49]. We use this proposition to show that closeness in maximum relative entropy implies an approximately multiplicatively optimal measurement:

Lemma 2.19: Suppose $\rho : \Theta \to \mathcal{D}(\mathcal{H})$ is some parametrised state and $\sigma : \Theta \to \mathcal{D}(\mathcal{H})$ is some classical state. If, for all θ, we have that $\exp\{D_{\max}(\rho[\theta]||\sigma[\theta]) + D_{\max}(\sigma[\theta]||\rho[\theta])\} \leq 1 + \eta$, then there is a η-multiplicatively optimal measurement M.

Proof: Fix some measurement $F \in \mathbb{M}(\mathcal{H})$, and estimator $\hat{\theta}^F$, note that

$$\frac{R_\rho(\hat{\theta}^F, \theta)}{R_\sigma(\hat{\theta}^F, \theta)} = \frac{\sum_i \mathrm{Tr}(\rho(\theta)F_i)L(\hat{\theta}^F(i), \theta)}{\sum_i \mathrm{Tr}(\sigma(\theta)F_i)L(\hat{\theta}^F(i), \theta)}, \tag{2.77}$$

$$\leq \max_i \frac{\mathrm{Tr}(\rho(\theta)F_i)}{\mathrm{Tr}(\sigma(\theta)F_i)}, \tag{2.78}$$

$$\leq e^{D_{\max}(\rho||\sigma)}, \tag{2.79}$$

where we used $\mathrm{Tr}(\rho(\theta)F_i) = \mathrm{Tr}(\sigma(\theta)F_i)(\mathrm{Tr}(\rho(\theta)F_i)/\mathrm{Tr}(\sigma(\theta)F_i))$ in the first inequality, and Lemma 2.2 in the second. Fixing an optimal measurement for σ and using the inequality (2.79) twice, the result follows. □

The final closeness result is for local optimality. For a subset of real vectors $Y \subseteq \mathbb{R}^k$, denote its Lebesgue measure by $|Y|$. We find the following locally-optimal result:

Lemma 2.20: Suppose $\rho : \Theta \to \mathcal{D}(\mathcal{H})$ is some parametrised state, which is classical on some subset $\Gamma \subseteq \Theta$. Then there is a $(|\Theta| - |\Gamma|)$-locally optimal measurement M.

Proof: Take the optimal measurement for $\rho|_\Gamma$. This clearly has the desired property. □

2.2.4 Admissibility of Measurements

In classical parameter estimation, there has been significant effort to characterise the set of admissible estimators for common choices of model P_θ [50–53], which allows one to ensure a choice of estimator is indeed admissible. In this setting we provide the beginnings of a such a theory for quantum metrology, proving two necessary conditions on a measurement for it to be admissible:

(i) The probabilities of the measurement outcomes depend on the parameter.

(ii) None of the post-measurement states of the measurement depend on the parameter.

These desirability of these conditions is evident: if (ii) does not hold the system can be measured further, allowing for a better estimate of the parameter, if (i) does not hold then the measurement is clearly unfit for purpose. However, without further specifying the parameterised quantum state ρ, it is difficult to give general criteria for when a measurement is inadmissible.

We prove (i) and (ii) for any loss function satisfying the following pair of properties:

(P1) L is strictly convex in its first argument.

(P2) For any prior π on Θ, the Bayes estimator $\hat{\theta}^{\mathrm{B}}$ is unique and is given by the posterior mean $\hat{\theta}^{\mathrm{B}}(x) = \mathbb{E}_\pi[\theta|\text{outcome } x]$.

Condition (P2) may seem strong, both (P1) and (P2) are satisfied by a ubiquitous family of Loss functions, known as Bregman divergences (see Ref. [54]). Least-squares loss and the KL divergence are both examples of Bregman divergences. We require three additional technical results on admissibility of estimators.

Proposition 2.3: Let L be a loss function that satisfies (P2). Suppose the model P_θ admits a parameterised probability mass function $\{p_\theta(i)\}$, and that there exist $\theta_1, \theta_2 \in \Theta$ such that $p_{\theta_1}(1)p_{\theta_2}(2) \neq p_{\theta_1}(2)p_{\theta_2}(1)$. Then, there exists an admissible estimator $\hat{\theta}$ where $\hat{\theta}(1) \neq \hat{\theta}(2)$.

Proof: Define a Prior on Θ of uniformly random choice of θ_1 or θ_2, that is to say,

$$\mathbb{P}(X = \theta) = \begin{cases} 1/2, & \theta \in \{\theta_1, \theta_2\}, \\ 0, & \text{o.w.} \end{cases} \tag{2.80}$$

In this case, by property (P2), we know that

$$\hat{\theta}^{\mathrm{B}}(i) = \mathbb{E}_\pi[\theta|i], \tag{2.81}$$

$$= \Pi(\theta_1|i)\theta_1 + \Pi(\theta_2|i)\theta_2. \tag{2.82}$$

Since this is a convex combination of θ_1 and θ_2 which must be distinct to satisfy the conditions of the proposition, we deduce that $\hat{\theta}^{\mathrm{B}}(1) = \hat{\theta}^{\mathrm{B}}(2)$ iff. $\Pi(\theta_1|1) = \Pi(\theta_1|2)$. But

$$\Pi(\theta_1|i) = \frac{p_{\theta_1}(i)}{p_{\theta_1}(i) + p_{\theta_2}(i)}. \tag{2.83}$$

So

$$\Pi(\theta_1|1) = \Pi(\theta_1|2) \Leftrightarrow \frac{p_{\theta_1}(1)}{p_{\theta_1}(1) + p_{\theta_2}(1)} = \frac{p_{\theta_1}(2)}{p_{\theta_1}(2) + p_{\theta_2}(2)}, \tag{2.84}$$

$$\Leftrightarrow p_{\theta_1}(1)p_{\theta_2}(2) = p_{\theta_1}(2)p_{\theta_2}(1). \tag{2.85}$$

Thus, by the assumption, $\hat{\theta}^{\mathrm{B}}(1) \neq \hat{\theta}^{\mathrm{B}}(2)$. Moreover, by Lemma 2.2, $\hat{\theta}^{\mathrm{B}}$ is admissible. □

Proposition 2.4: Let L be a loss function that satisfies (P1). Suppose that $\hat{\theta}$ is an admissible estimator, and that $\tilde{\theta} \leq \hat{\theta}$. Then $\hat{\theta} = \tilde{\theta}$ with probability 1 (for all values of the parameter).

Proof: Let $\hat{\theta}'(x) = [\hat{\theta}(x) + \tilde{\theta}(x)]/2$. By strict convexity of L in its first argument, $R(\hat{\theta}', \cdot) \leq R(\hat{\theta}, \cdot)$. Since $\hat{\theta}$ is admissible, there must be equality for all θ and the result follows from strict convexity. □

Proposition 2.5: Let L be a loss function that satisfies (P1). Suppose the model P_θ does not depend on θ. Then an estimator $\hat{\theta}$ is admissible iff. it is constant with probability 1.

Proof: Let $\hat{\theta}$ be an estimator. Consider the constant estimator $\hat{\theta}^{\theta_0} = \mathbb{E}_{X \sim P_\theta}[\hat{\theta}(X)]$, which is well defined by the assumption that P_θ is constant in θ. Then,

$$R(\hat{\theta}, \theta) = \mathbb{E}_{X \sim P_\theta}[L(\hat{\theta}(X), \theta)], \tag{2.86}$$

$$\text{Jensen's inequality and (P1)}, \quad \geq L(\theta_0, \theta) = R(\hat{\theta}^{\theta_0}, \theta). \tag{2.87}$$

By proposition 2.4, $\hat{\theta}$ is admissible only if it equals θ_0 with probability 1. □

We can now prove the two necessary conditions for admissibility. Firstly, we show condition (i)

Lemma 2.21: Let L be a loss function on Θ that satisfies (P1) and (P2), and let $\rho : \Theta \to \mathcal{D}(\mathcal{H})$ be some non-constant parametrised state. Suppose that $F \in \mathbb{M}(\mathcal{H})$ is a measurement whose outcome probabilities are independent of θ, then F is inadmissible.

Proof: Suppose there exists a measurement $M \in \mathbb{M}(\mathcal{H})$ and a non-constant admissible estimator $\hat{\theta}^{\mathrm{M}}$. By proposition 2.5, every admissible estimator of F is constant (with probability one) and thus $M \preceq F$. Suppose there exists a non-constant admissible estimator $\hat{\theta}^{\mathrm{M}}$. If there exists $\hat{\theta}^{\mathrm{F}} \leq \hat{\theta}^{\mathrm{M}}$, then (by proposition 2.3), WLOG $\hat{\theta}^{\mathrm{F}} = \hat{\theta}^{\theta_0}$ is constant. Then, by proposition 2.4, $\hat{\theta}^{\mathrm{M}}$ is constant, contradicting our assumption. Thus if such an $\hat{\theta}^{\mathrm{M}}$ exists we see that $F \not\preceq M$, completing the proof.

By the Helstrom bound [28], there is some measurement $M \in \mathbb{M}_2(\mathcal{H})$ where $p_{\theta_1}^{(\rho;\, M)}(1) \neq p_{\theta_2}^{(\rho;\, M)}(1)$. Then, note that

$$p_{\theta_1}^{(\rho;\, M)}(1)p_{\theta_2}^{(\rho;\, M)}(2) = p_{\theta_1}^{(\rho;\, M)}(2)p_{\theta_2}^{(\rho;\, M)}(1), \tag{2.88}$$

$$\Leftrightarrow p_{\theta_1}^{(\rho;\, M)}(1) = p_{\theta_2}^{(\rho;\, M)}(1). \tag{2.89}$$

An application of proposition 2.3 gives a non-constant admissible estimator $\hat{\theta}^{\mathrm{M}}$. □

In order to consider condition (ii), we must consider the post-measurement state of the system. Fix a measurement $N = (K, M) \in \mathbb{M}(\mathcal{H})$ and suppose it admits Kraus operators $\{F_i\}_{i=1}^{K}$, i.e. $M_i = F_i^\dagger F_i$. If one observes outcome i, the post measurement state is

$$\rho_i = \frac{F_i \rho F_i^\dagger}{\mathrm{Tr}\Big(F_i \rho F_i^\dagger\Big)} : \Theta \to \mathcal{D}(\mathcal{H}). \tag{2.90}$$

We say that the measurement M is refineable if it admits a choice of Kraus operators, such that at least one of the post-measurement states (WLOG the first) is non-constant. That is, there are $\theta_1, \theta_2 \in \Theta$ such that

1. $\rho_1(\theta_1) \neq \rho_1(\theta_2)$.
2. $p_{\theta_1}^{(\rho;\, M)}(1),\, p_{\theta_2}^{(\rho;\, M)}(1) \neq 0$.

The second of these conditions is to ensure that the outcome 1 is possible at θ_1 and θ_2. Otherwise, the post measurement state is not well-defined. We show that refineable measurements are inadmissible:

Lemma 2.22: Let L be a loss function on Θ that satisfies (P1) and (P2), and let $\rho : \Theta \to \mathcal{D}(\mathcal{H})$ be some non-constant parametrised state. Suppose that $F \in \mathbb{M}_K(\mathcal{H})$ is a refineable measurement, then F is inadmissible.

Proof: Fix some choice of Kraus operators $F_i = G_i^\dagger G_i$, and $\theta_1, \theta_2 \in \Theta$ such that conditions 1 and 2 above are satisfied. Since $\rho_1(\theta_1) \neq \rho_1(\theta_2)$, by the Helstrom Bound [28], there is some measurement $M \in \mathbb{M}_2(\mathcal{H})$ where $p_{\theta_1}^{(\rho_1;\, M)}(1) \neq p_{\theta_2}^{(\rho_1;\, M)}(1)$. Consider the concatenated measurement $MF \in \mathbb{M}_{2K}(\mathcal{H})$ corresponding to measuring first with F and then M. That is,

$$MF_{(i,j)} = G_i^\dagger M_j G_i, \tag{2.91}$$

for $i \in [K]$, $j \in [2]$. Note that for any estimator $\hat{\theta}^F$ we can construct an estimator $\hat{\theta}_F^{MF}(i,j) =$

$\hat{\theta}^F(i)$ satisfying $R(\hat{\theta}^{MF}_F, \cdot) = R(\hat{\theta}^F, \cdot)$. Thus, $MF \preceq F$. It remains to show $F \not\preceq MF$.

Suppose we find an MF-admissible estimator $\hat{\theta}^{MF}$ satisfying $\hat{\theta}^{MF}(1,1) \neq \hat{\theta}^{MF}(1,2)$. If there is an estimator $\hat{\theta}^F$ such that $\hat{\theta}^F \leq \hat{\theta}^{MF}$, then using the construction above, we find $\hat{\theta}^{MF}_F \leq \hat{\theta}^{MF}$. But, by construction, $\hat{\theta}^{MF}_F \neq \hat{\theta}^{MF}$ (with non-zero probability), contradicting proposition 2.4, and hence $F \not\preceq MF$. Thus, it suffices to find such an estimator $\hat{\theta}^{MF}$.

By proposition 2.3 it is sufficient to show $p^{(\rho;\,MF)}_{\theta_1}(1,1)p^{(\rho;\,MF)}_{\theta_2}(1,2) \neq p^{(\rho;\,MF)}_{\theta_1}(1,2)p^{(\rho;\,MF)}_{\theta_2}(1,1)$. But by expanding probabilities and dividing through by $p^{(\rho;\,F)}_{\theta_1}(1)p^{(\rho;\,F)}_{\theta_2}(1)$, we find that

$$p^{(\rho;\,MF)}_{\theta_1}(1,1)p^{(\rho;\,MF)}_{\theta_2}(1,2) = p^{(\rho;\,MF)}_{\theta_1}(1,2)p^{(\rho;\,MF)}_{\theta_2}(1,1), \tag{2.92}$$

$$\Leftrightarrow p^{(\rho_1;\,M)}_{\theta_1}(1)p^{(\rho_1;\,M)}_{\theta_2}(2) = p^{(\rho_1;\,M)}_{\theta_1}(2)p^{(\rho_1;\,M)}_{\theta_2}(1), \tag{2.93}$$

$$\Leftrightarrow p^{(\rho_1;\,M)}_{\theta_1}(1) = p^{(\rho_1;\,M)}_{\theta_2}(1). \tag{2.94}$$

However, by the construction of M, the final equality (2.94) cannot hold. □

Finally, we provide a method to show a measurement M is admissible, generalising Lemma 2.2 to the quantum setting.

Lemma 2.23: Suppose that $\rho : \Theta \to \mathcal{D}(\mathcal{H})$ is some parameterised quantum state, and that π is some prior distribution on Θ. If $M \in \mathbb{M}(\mathcal{H})$ is a unique Bayes measurement (with respect to π), then M is admissible

Proof: Suppose that F dominates M. Let $\hat{\theta}^M$ be the Bayes estimator of M, as M is a Bayes measurement, $\hat{\theta}^M$ minimises the Bayes risk across all choices of measurement and estimator. However, as $F \preceq M$, there exists $\hat{\theta}^F$ such that $\hat{\theta}^F \leq \hat{\theta}^M$. Then $R_\pi(\hat{\theta}^F) \leq R_\pi(\hat{\theta}^M)$ and F must also be Bayes. This contradicts the uniqueness of M. □

2.2.5 Proof of Lemma 2.17

We now present the proof of Lemma 2.17 as a series of propositions. We first consider the single parameter ($\Theta \subseteq \mathbb{R}$) case. We fix $\theta_1, \theta_2 \in \Theta$ and consider the restricted parameter space $\{\theta_1, \theta_2\} \subseteq \Theta$ and show that if an optimal measurement M exists, then $\rho(\theta_1)$ and $\rho(\theta_2)$ commute. This is achieved by considering the Bayesian measurements associated with all possible priors on $\{\theta_1, \theta_2\}$. This shows that if an optimal measurement exists on Θ, all $\{\rho(\theta) : \theta \in \Theta\}$ must commute, and thus ρ is classical. In higher dimensions, we restrict the parameter estimation problem to lines, which allows us to use the single parameter result. Throughout this Section we return to the convention that our loss function L is always least-squares. We now give the series of propositions needed for Lemma 2.17.

Proposition 2.6: Suppose $\rho : \Theta \to \mathcal{D}(\mathcal{H})$ is some parametrised state with an optimal measurement M. Then M is a Bayes measurement for any prior π on Θ.

Proof: Suppose we have some Bayes estimator $\hat{\theta}^F$. Then since M is optimal, $M \preceq F$, and thus there is an estimator $\hat{\theta}^M$ satisfying $\hat{\theta}^M \leq \hat{\theta}^F$. But then, by definition, $\hat{\theta}^M$ must also be Bayes.

Thus for any prior, we can always measure with M and pick a Bayes estimator based on the resulting probability distribution. □

We introduce some notation from [29], useful in single parameter quantum Bayesian estimation. For any measurement-estimator pair $(F, \hat{\theta}^F)$, we define

$$\Lambda = \sum_i F_i \hat{\theta}^F(i), \quad \Lambda_2 = \sum_i F_i \hat{\theta}^F(i)^2, \tag{2.95}$$

Furthermore, for any prior π on Θ, we define two operators

$$\bar{\rho} = \int d\theta \pi(\theta)\rho(\theta), \quad \bar{\rho}' = \int d\theta \pi(\theta)\rho(\theta)\theta. \tag{2.96}$$

Proposition 2.7: Suppose $\rho : \Theta \to \mathcal{D}(\mathcal{H})$ is some single parameter state ($\Theta \subseteq \mathbb{R}$). Suppose further we have some prior π on Θ and that $\bar{\rho}$ has full rank. Then a measurement $F \in \mathbb{M}(\mathcal{H})$ is Bayesian iff. the following hold

(i) F is the projective measurement onto Λ's eigenspaces, or a fine-graining of it.

(ii) $\{\Lambda, \bar{\rho}\} = 2\bar{\rho}'$.

Proof: Note that this proposition, and most of this proof, is presented in [29] as a sufficient condition. We prove that the conditions (i) and (ii) are also necessary. Suppose some measurement-estimator pair $(F, \hat{\theta}^F)$ is Bayesian. Note that we may write the Bayesian risk in terms of Λ and Λ_2:

$$R_\pi(\hat{\theta}^F) = \int d\theta \sum_i \pi(\theta) \operatorname{Tr}(F_i \rho(\theta))(\hat{\theta}^F(i) - \theta)^2, \tag{2.97}$$

$$= \operatorname{Tr}(\Lambda_2 \bar{\rho}) - 2\operatorname{Tr}\big(\Lambda \bar{\rho}'\big) + \int d\theta \pi(\theta)\theta^2. \tag{2.98}$$

We note that

$$0 \leq \sum_i (\hat{\theta}^F(i)\mathbb{1} - \Lambda)^\dagger F_i (\hat{\theta}^F(i)\mathbb{1} - \Lambda), \tag{2.99}$$

$$= \Lambda_2 - \Lambda^2, \tag{2.100}$$

and thus $\operatorname{Tr}(\Lambda_2\bar{\rho}) \geq \operatorname{Tr}\big(\Lambda^2\bar{\rho}\big)$. If F does not saturate this inequality, let M be the projective measurement onto the eigenspaces of Λ, and $\hat{\theta}^{\mathrm{M}}$ output the corresponding eigenvalue. Then $R_\pi(\hat{\theta}^{\mathrm{M}}) < R_\pi(\hat{\theta}^{\mathrm{F}})$, contradicting that $\hat{\theta}^{\mathrm{F}}$ is Bayesian, and we deduce that $\operatorname{Tr}(\Lambda_2\bar{\rho}) = \operatorname{Tr}\big(\Lambda^2\bar{\rho}\big)$. and thus equation (2.99) must be an equality. Note that for any positive semi-definite operator A,

$$\operatorname{Tr}(\bar{\rho}A) = 0 \Leftrightarrow \operatorname{Tr}\Big((\sqrt{\bar{\rho}}\sqrt{A})^\dagger \sqrt{\bar{\rho}}\sqrt{A}\Big) = 0, \tag{2.101}$$

$$\Leftrightarrow \sqrt{\bar{\rho}}\sqrt{A} = 0, \tag{2.102}$$

$$\Leftrightarrow A = 0, \tag{2.103}$$

where we used the assumption that $\bar{\rho}$ is full rank, and hence that $\sqrt{\bar{\rho}}$ is invertible. Thus, $\sum_i (\hat{\theta}^F(i)\mathbb{1} - \Lambda)^\dagger F_i (\hat{\theta}^F(i)\mathbb{1} - \Lambda) = 0$. But since this is a sum of positive semi-definite operators, this implies that, for each i, $(\hat{\theta}^F(i)\mathbb{1} - \Lambda)^\dagger F_i (\hat{\theta}^F(i)\mathbb{1} - \Lambda) = 0$. If $\hat{\theta}^F(i)$ is not an eigenvalue of Λ, then $F_i = 0$. Otherwise (by prop. 1.1 (i)), F_i must only have support on the $\hat{\theta}^F(i)$ eigenspace of Λ. Since $\Lambda = \sum F_i \hat{\theta}^F(i)$, condition (i) follows.

We see that Λ must minimise the function

$$\mathcal{C} : \mathrm{Herm}(\mathcal{H}) \to \mathbb{R},\ C(X) = \mathrm{Tr}\big(X^2 \bar{\rho}\big) - 2\,\mathrm{Tr}\big(X \bar{\rho}'\big), \tag{2.104}$$

over all Hermitian operators X. By differentiating with respect to X and setting the derivative to zero, we see that there is a unique minimum Λ and that it satisfies

$$\frac{1}{2}\{\Lambda, \bar{\rho}\} = \bar{\rho}'. \tag{2.105}$$

□

Proposition 2.8: Suppose $\rho : \Theta \to \mathcal{D}(\mathcal{H})$ is some single parameter state with an optimal measurement M under least-squares loss. Fix $\theta_1, \theta_2 \in \Theta$ distinct. If $\rho(\theta_1)$, $\rho(\theta_2)$ both have full rank, then there exist simultaneously diagonalisable Hermitian maps Γ, Γ' satisfying

$$\{\Gamma, \rho(\theta_2)\} = \rho(\theta_1), \quad \{\Gamma', \rho(\theta_1)\} = \rho(\theta_2). \tag{2.106}$$

Proof: For notational convenience, let $\rho_i = \rho(\theta_i)$ for $i = 1, 2$. For $p \in (0, 1)$ fix a prior on Θ where we are given ρ_1 with probability p and ρ_2 with probability $1 - p$. Note that for any $p \in (0, 1)$, $\bar{\rho}$ has full rank. By proposition 2.6 we see that M must be Bayesian for any value of p - let $\hat{\theta}_p^M$ be a Bayes estimator for a fixed value of p. Then $\Lambda = \sum_i M_i \hat{\theta}_p^M(i)$ and M must satisfy condition (i) of proposition 2.7. We say $i \sim j$ if, for all $p \in (0, 1)$, $\hat{\theta}_p^M(i) = \hat{\theta}_p^M(j)$. If $i \nsim j$, take p such that $\hat{\theta}_p^M(i) \neq \hat{\theta}_p^M(j)$. Then M_i only has support on the $\hat{\theta}_p^M(i)$ eigenspace of Λ, whereas M_j only has support on the $\hat{\theta}_p^M(j)$ eigenspace, and thus M_i and M_j have orthogonal supports. Hence, the subspaces $S_i = \oplus_{j \in [i]} \mathrm{supp}(M_j)$, are orthogonal. Since $\sum_i M_i = \mathbb{1}$, it follows that $\sum_{j \in [i]} M_j = \Pi_{S_i}$. Hence, we may expand $\Lambda = \sum_k \mu_k(p) \, |k\rangle\langle k|$ - in an eigenbasis that does not depend on p (but whose eigenvalues will generically depend on p).

By condition (ii) of Lemma 2.7, for any value of $p \in (0, 1)$ we know that Λ satisfies

$$\frac{1}{2}\{\Lambda, p\rho_1 + (1-p)\rho_2\} = p\theta_1\rho_1 + (1-p)\theta_2\rho_2. \tag{2.107}$$

As well as being invertible, the smallest eigenvalue of $\bar{\rho}$, $\lambda_{\min}(p)$ is bounded below by the minimum of the eigenvalues of ρ_1 and ρ_2. Consider equation (2.107) in the limit of $p \to 0$. Let Γ be defined implicitly as the solution to the equation

$$\{\Gamma, \rho_2\} = \rho_1. \tag{2.108}$$

By expanding equation (2.108) in the eigenbasis of ρ_2, and using that ρ_2 is invertible, we see that

Γ is well defined and is furthermore Hermitian. For some "remainder" matrix R, we expand Λ as $\Lambda = \theta_2 \mathbb{1} + 2p(\theta_1 - \theta_2)\Gamma + R$. Substituting this ansatz into equation (2.107), we see that R must satisfy the equation

$$\frac{1}{2}\{R, \bar{\rho}\} = p^2(\theta_2 - \theta_1)\{\Gamma, \rho_1 - \rho_2\}. \tag{2.109}$$

Expanding in an eigenbasis of $\bar{\rho}$ and recalling that all the eigenvalues of $\bar{\rho}$ are bounded below by $\lambda_{\min}$ (which is independent of p), we see that $R = O(p^2)$. Consider an eigenstate $|\psi\rangle$ of Λ with eigenvalue $\mu(p)$, then note that

$$\left\| \Gamma |\psi\rangle - \frac{\mu(p) - \theta_2}{2p(\theta_1 - \theta_2)} |\psi\rangle \right\| = O(p). \tag{2.110}$$

Taking the limit as $p \to 0$, we see that $|\psi\rangle$ must be an eigenstate of Γ. But, by symmetry, in the limit $p \to 1$, we may define Γ' by the equation

$$\{\Gamma', \rho_1\} = \rho_2, \tag{2.111}$$

and $|\psi\rangle$ must also be an eigenstate of Γ'. Thus Γ and Γ' must be simultaneously diagonalisable in Λ's eigenbasis. □

Proposition 2.9: Suppose A, B are positive definite and Γ, Γ' are simultaneously diagonalisable Hermitian maps satisfying

$$\{\Gamma, A\} = B, \quad \{\Gamma', B\} = A. \tag{2.112}$$

Then A and B commute.

Proof: Let Γ and Γ' be simultaneously diagonalisable in the orthonormal basis $\{|i\rangle\}_{i=1}^n$. Let $\Gamma |i\rangle = \lambda_i |i\rangle$, $\Gamma' |i\rangle = \mu_i |i\rangle$, $\langle i| A |j\rangle = A_{ij}$ and $\langle i| B |j\rangle = B_{ij}$, for $i, j \in [n]$. By positive definiteness, note that $A_{ii}, B_{ii} > 0$ for any i. Rewriting equations (2.112) in components of this basis, we reach the set of equations

$$(\lambda_i + \lambda_j)A_{ij} = B_{ij}, \quad (\mu_i + \mu_j)B_{ij} = A_{ij}. \tag{2.113}$$

Setting $i = j$ we deduce that

$$\lambda_i = B_{ii}/2A_{ii}, \quad \mu_i = A_{ii}/2B_{ii}. \tag{2.114}$$

Equation (2.113) shows that $A_{ij} = 0 \Leftrightarrow B_{ij} = 0$. Suppose $A_{ij} \neq 0$, then multiplying the two individual equations in (2.113) and dividing by $A_{ij}B_{ij}$ we see that

$$\begin{aligned} &(A_{ii}/2B_{ii} + A_{jj}/2B_{jj})\,(B_{ii}/2A_{ii} + B_{jj}/2A_{jj}) = 1, \\ &\Leftrightarrow \frac{A_{ii}B_{jj}}{B_{ii}A_{jj}} + \frac{B_{ii}A_{jj}}{A_{ii}B_{jj}} = 2. \end{aligned} \tag{2.115}$$

But $x + 1/x = 2$ iff. $x = 1$ and we deduce that

$$\frac{A_{ii}}{B_{ii}} = \frac{A_{jj}}{B_{jj}}. \tag{2.116}$$

Consider a graph G on n nodes, with an edge between i and j iff. $A_{ij} \neq 0$. If G is connected, then for any nodes i, j in G there is path between them. Assuming connectedness, we apply the result of equation (2.116) along the path, we deduce that $A_{ii}B_{jj} = B_{ii}A_{jj}$ for every $i, j \in [n]$. But then summing over j, we deduce that $A_{ii} = \frac{\mathrm{Tr}(A)}{\mathrm{Tr}(B)}B_{ii}$. Substituting this into equations (2.113) and (2.114), we see that A and B are proportional and thus commute. If G is not connected, then we can apply the above procedure to each connected component of G. The result is that A and B are block diagonal, with the diagonal matrix entries proportional and thus A and B commute. □

Proposition 2.10: Suppose $\rho : \Theta \to \mathcal{D}(\mathcal{H})$ is some single parameter state ($\Theta \subseteq \mathbb{R}$) with an optimal measurement $M \in \mathbb{M}_L(\mathcal{H})$. Fix $\theta_1, \theta_2 \in \Theta$ distinct, then $\rho(\theta_1)$, $\rho(\theta_2)$ commute (regardless of their rank).

Proof: Again, for notational convenience, let $\rho_i = \rho(\theta_i)$ for $i = 1, 2$. Redefine our parameter space $\Theta = \{\theta_1, \theta_2\}$, but we still allow estimators to take any real value. Note that, by definition, M must also be optimal for this parameter estimation problem.

Suppose that $\mathcal{U} = \ker(\rho_2) \cap \ker(\rho_1) \neq \{0\}$. Let $\mathcal{U}^\perp$ be the orthogonal complement of $\mathcal{U}$ and let $\Pi_{\mathcal{U}^\perp}$ be the orthogonal projection matrix onto $\mathcal{U}^\perp$. Then note that replacing M with the measurement $(\Pi_{\mathcal{U}^\perp} M_i \Pi_{\mathcal{U}^\perp}, \mathbb{1} - \Pi_{\mathcal{U}^\perp}) \in \mathbb{M}_{L+1}(\mathcal{H})$ does not change any of the measurement probabilities when measuring ρ_1 or ρ_2. Thus, we may WLOG restrict our Hilbert space to $\mathcal{U}^\perp$, i.e. we take $\mathcal{H} = \mathcal{U}^\perp$, which still has an optimal measurement for this restricted parameter estimation problem (given above). Furthermore $\Pi_{\mathcal{U}}\rho_1\Pi_{\mathcal{U}}$ and $\Pi_{\mathcal{U}}\rho_1\Pi_{\mathcal{U}}$ commute iff. ρ_1 and ρ_2 do.

Suppose that ρ_2 is not full rank, i.e. it has some non-trivial kernel $K \leq \mathcal{H}$. Take a measurement $E = (\Pi_K, \Pi_{K^\perp}) \in \mathbb{M}_2(\mathcal{H})$, along with an estimator $\hat{\theta}^E(K) = \theta_1$, $\hat{\theta}^E(K^\perp) = \theta_2$ so that

$$R(\hat{\theta}^E, \theta_2) = 0, \tag{2.117}$$

$$R(\hat{\theta}^E, \theta_1) = (\theta_1 - \theta_2)^2[1 - \mathrm{Tr}(\Pi_K \rho_1)]. \tag{2.118}$$

Since M is optimal, we have that $M \preceq E$. Then, in particular, there must be an estimator $\hat{\theta}^M$ satisfying $\hat{\theta}^M \leq \hat{\theta}^E$. In particular, $\hat{\theta}^M$ must have zero risk at θ_2 and thus if $\mathrm{Tr}(\rho_2 M_i) \neq 0$, then we must have $\hat{\theta}^M(i) = \theta_2$. Let $I = \{i \mid \mathrm{Tr}(\rho_2 M_i) = 0\}$, then, by the above, $\hat{\theta}^M$ must satisfy

$$R(\hat{\theta}^M, \theta_1) \geq (\theta_1 - \theta_2)^2 \left(1 - \sum_{i \in I} \mathrm{Tr}(\rho_1 M_i)\right), \tag{2.119}$$

with equality iff. for every $i \in I$, $\hat{\theta}^M(i) = \theta_1$. WLOG, assume that this holds, so that we saturate equation (2.119). If $\mathrm{Tr}(\rho_2 M_i) = 0$, then, as $M_i \geq 0$, every eigenvector of M_i with non-zero eigenvalue must lie in K and thus $\sum_{i \in I} M_i \leq \Pi_K$. Thus, $R(\hat{\theta}^M, \theta_1) \geq R(\hat{\theta}^E, \theta_2)$ with equality iff. $\rho_1 \left(\Pi_K - \sum_{i \in I} M_i\right) = 0$. Since ρ_1 does not kill any vector in K (as we restricted to $\mathcal{U}^\perp$) we get equality iff. $\sum_{i \in I} M_i = \Pi_K$. Thus, as $\sum_{i=1}^L M_i = \mathbb{1}$, for $i \notin I$ and $|k\rangle \in K$, $M_i |k\rangle = 0$. Using proposition 1.1 (ii), we deduce that, if $i \in I$, then $M_i = \Pi_K M_i \Pi_K$ and $M_i \Pi_{K^\perp} = 0$, and if $i \notin I$, then $M_i = \Pi_{K^\perp} M_i \Pi_{K^\perp}$ and $M_i \Pi_K = 0$

Now, for $p \in (0,1)$ fix a prior on Θ where we are given ρ_1 with probability p and ρ_2 with probability $1-p$. Note that (as we restricted to $\mathcal{U}^\perp$), $\bar{\rho}$ has full rank. By proposition 2.6, there exists a Bayesian estimator $\hat{\theta}_p^M$. Taking $\Lambda = \sum_i M_i \hat{\theta}_p^M(i)$ we find that Λ decomposes as $\Lambda = \Lambda_K + \Lambda_{K^\perp}$, where $\Lambda_K = \Pi_K \Lambda_K \Pi_K$ and $\Lambda_{K^\perp} = \Pi_{K^\perp} \Lambda_{K^\perp} \Pi_{K^\perp}$. Furthermore, in order for our estimator to be Bayes, we must always guess θ_1 in the case of a K outcome i.e. $\Lambda_K = \theta_1 \Pi_K$.

Taking the inner product of condition (ii) of Lemma 2.7 with $|k\rangle \in K$ and $|\ell\rangle \in K^\perp$, we find that

$$\langle \ell | \Lambda_{K^\perp} \rho_1 | k \rangle = \theta_1 \langle \ell | \rho_1 | k \rangle . \tag{2.120}$$

Note that as $p \to 0$, we must have $\Lambda_{K^\perp} \to \Pi_{K^\perp} \theta_2$, as our estimates must approach θ_2 for our estimator to be Bayes. Thus the only way for equation (2.120) to be satisfied for all $p \in (0,1)$ is for $\langle \ell | \rho_1 | k \rangle = 0$. Thus ρ_1 fixes K and we may decompose it as $\rho_1 = \Pi_K \rho_1 \Pi_K + \Pi_{K^\perp} \rho_1 \Pi_{K^\perp}$. Thus equation (2.107) becomes

$$\begin{aligned} &\frac{1}{2}\{\Lambda_{K^\perp}, p\Pi_{K^\perp}\rho_1\Pi_{K^\perp} + (1-p)\Pi_{K^\perp}\rho_2\Pi_{K^\perp}\}, \\ &= p\theta_1 \Pi_{K^\perp}\rho_1\Pi_{K^\perp} + (1-p)\theta_2\Pi_{K^\perp}\rho_2\Pi_{K^\perp}. \end{aligned} \tag{2.121}$$

We observe that ρ_1 and ρ_2 commute iff. $\Pi_{K^\perp}\rho_1\Pi_{K^\perp}$ and $\Pi_{K^\perp}\rho_2\Pi_{K^\perp}$ do.

Consider the restricted estimation problem $\tilde{\rho} : \{\theta_1, \theta_2\} \to \mathcal{D}(K^\perp)$, $\tilde{\rho}(\theta) = \Pi_{K^\perp}\rho(\theta)\Pi_{K^\perp}$. Suppose that $F \in \mathbb{M}_J(K^\perp)$, and we have some estimator $\hat{\theta}^F$, we can extend it to a measurement $\bar{F} \in \mathbb{M}_{J+1}(\mathcal{H})$ by

$$\bar{F}_i = \begin{cases} F_i, & i \neq J+1, \\ \Pi_K, & i = J+1. \end{cases} \tag{2.122}$$

Similarly, we extend $\hat{\theta}^F$ to an estimator $\hat{\theta}^{\bar{F}}$ by $\hat{\theta}^{\bar{F}}|_{[J]} = \hat{\theta}^F$ and $\hat{\theta}^{\bar{F}}(J+1) = \theta_1$. By optimality of M, there is some estimator $\hat{\theta}^M \leq \hat{\theta}^{\bar{F}}$, where WLOG (as this cannot increase the risk), for $i \in I$, we take $\hat{\theta}^M(i) = \theta_1$. Consider the restricted measurement $\tilde{M} \in \mathbb{M}(\mathcal{H})$ given by $\tilde{M} = (\Pi_{K^\perp} M_i \Pi_{K^\perp})_{i \notin I}$, with corresponding estimator $\hat{\theta}^{\tilde{M}} = \hat{\theta}^M|_{[L]\setminus I}$. A brief calculation shows that $\hat{\theta}^{\tilde{M}} \leq \hat{\theta}^F$. Thus, $\tilde{\rho}$ admits an optimal measurement $\tilde{M}$. Hence, we may restrict our Hilbert space to $K^\perp$, on which ρ_2 has full rank.

We may then repeat the above procedure, reducing our Hilbert space to one on which both ρ_1 and ρ_2 have full rank. The key insight is that if $\rho_2 = \Pi_K \rho_2 \Pi_K + \Pi_{K^\perp}\rho_2\Pi_{K^\perp}$ has full rank, then so too does $\Pi_{K^\perp}\rho_2\Pi_{K^\perp}$, and thus ρ_2 remains full rank after repeating the procedure. We then apply propositions 2.8 and 2.9, and the result follows. □

Note that the proof of proposition 2.10 may be seen as a generalisation of example 2.5. In light of propositions 2.1 and 2.10, we see that Lemma 2.17 holds in the single parameter case, for any parameter space Θ, without the assumption of convexity. We can use the single parameter case to prove the full version of the Lemma:

Lemma 2.2: Suppose that $\rho : \Theta \to \mathcal{D}(\mathcal{H})$ is some parameterised quantum state, where $\Theta \subseteq \mathbb{R}^k$ is convex. If ρ admits an optimal measurement (under least-squares loss), then it is classical.

Proof: Fix $\theta_1, \theta_2 \in \Theta$ and let $T = [0, ||\theta_2 - \theta_1||]$. By convexity, for $t \in T$, $\gamma(t) = \theta_1 + t(\theta_2 - \theta_1)/||\theta_2 - \theta_1||$ is in Θ. Thus $\tilde{\rho} : T \to \mathcal{D}(\mathcal{H})$, $\tilde{\rho}(t) = \rho(\gamma(t))$ gives a single parameter parameterised state. We have constructed this state such that the least-squares loss functions agree - that is for $s, t \in T$,

$$||\gamma(t) - \gamma(s)||^2 = |t - s|^2. \tag{2.123}$$

But then (by projecting any estimators $\hat{\theta}^M$ onto the line segment between θ_1 and θ_2) M must also be optimal for $\tilde{\rho}$, and thus, by proposition 2.10, $\rho(\theta_1)$ and $\rho(\theta_2)$ commute. □

2.2.6 Outlook

Above, we have introduced, and provided several central results, within the theory of admissibility of measurements in quantum metrology. The theory of admissibility of classical estimators is rich, and the corresponding theory for measurements should prove more complex. This complexity is already evident in the disparity between the existence of optimal measurements (see Lemma 2.16) and lack of existence of optimal estimators (see Lemma 2.1).

With regards to optimal measurements, it would be desirable to generalise theorem 2.1 to other loss functions, such as other p-norms, or the loss functions used in thresholding. It would also be desirable to relax the requirements on Θ. For example, the parameter space S^1 of the phase-encoded state (example 2.4) is naturally embeddable into $\mathbb{R}^2$, but it is not convex. It can also be thought of as $S^1 \simeq [0, 2\pi)$, but the least squares loss function on $[0, 2\pi)$ does not respect the periodicity of the problem. Indeed, a more sensible choice of loss function, which locally reproduces least-squares, but respects global periodicity, would be $L(\theta_1, \theta_2) = 4\sin^2([\theta_1 - \theta_2]/2)$ [6], to which our result does not apply.

With regards to approximately optimal measurements, we anticipate it to be harder to construct a robust converse result (i.e. a robust version of Lemma 2.17), given the more involved nature of the original proof. Nonetheless, it would be desirable to characterise the existence of approximately optimal measurements, and moreover develop fast classical algorithms to find approximately optimal measurements when they exist.

In the asymptotic limit, we expect (approximately) optimal measurements to exist (as discussed in Section 2.1.6. A large number of copies of a pure state $|\Psi(\theta)\rangle = |\psi(\theta)\rangle^{\otimes n}, n \gg 1$ is never classical, yet we expect approximately optimal measurements to exist that saturate the QFI. Note that $[\,|\Psi(\theta_1)\rangle\langle\Psi(\theta_1)|\,, |\Psi(\theta_2)\rangle\langle\Psi(\theta_2)|\,] \sim |\,\langle\psi(\theta_1)|\psi(\theta_2)\rangle\,|^n$, and thus we expect exponential decay in the size of commutators. Thus, the state $|\Psi(\theta)\rangle$ looks more classical in the asymptotic limit. We hope that such intuition could be made more rigorous to show the existence of asymptotically optimal measurements in the sense of $\preceq$, without reference to a particular estimator.

Finally, we hope that one could make a complete characterisation of the admissible measurements for a given parameterised state ρ. To the best of our knowledge, there is no complete classical characterisation of admissible estimators in full generality, and thus a full quantum solution may prove challenging.

[6] This loss function also arises as the restriction of the $\mathbb{R}^2$ least squares loss function to S^1.

2.3 Performance of Quantum Filters

Section 2.1.7 provided a theoretical background, including motivation, for quantum filtering. In Section 2.3, we give a series of novel results in quantum filtering. Firstly, we completely characterise lossless quantum compression for parameterised pure states. Secondly, we show that the JAL filter (and indeed other optimal filters) are not optimal in the presence of noise (or more abstractly for mixed parameterised quantum states). Thirdly, we give an operational justification for why lossless compression by filtering is a genuine quantum effect. Finally, we give a concrete iterative scheme for quantum filtering in the ubiquitous model of Gaussian shift estimation (defined below). The results are mostly drawn from Refs. [2, 3]

2.3.1 Conditions for an Optimal Filter

In order to design quantum filters, it is essential to understand under what conditions a filter is lossless. In this Section we completely characterise lossless filters for parameterised pure states. We will use the following useful expression for the QFI, derived in Ref. [47].

Proposition 2.11: Let $|\psi(\cdot)\rangle : \Theta \to S(\mathcal{H})$ be a parameterised (pure) quantum state and let K be a Kraus operator of a filter. Then, for $\theta \in \Theta$, the QFI of the postselected state has the form

$$J(\theta|\rho_K^{\mathrm{ps}})_{i,j} = 4\,\mathrm{Re}\Bigg(\frac{1}{p^{\mathrm{ps}}_{\rho,\,K}(\theta)}\,\langle\partial_i\psi(\theta)|F|\partial_j\psi(\theta)\rangle - \frac{1}{p^{\mathrm{ps}}_{\rho,\,K}(\theta)^2}\,\langle\partial_i\psi(\theta)|F|\psi(\theta)\rangle\,\langle\psi(\theta)|F|\partial_j\psi(\theta)\rangle\Bigg). \tag{2.124}$$

With proposition 2.11, we can give a characterisation of when a filter is lossless:

Theorem 2.2: Let $|\psi(\cdot)\rangle : \Theta \to S(\mathcal{H})$ be a parameterised (pure) quantum state, where $\Theta \subseteq \mathbb{R}^k$. Fix $\theta \in \Theta$, and assume that $J(\theta|\psi)_{i,i} \neq 0$, for every $i \in [k]$. Let $\mathcal{U} = \mathrm{span}\{|\psi(\theta)\rangle\,, |\partial_i\psi(\theta)\rangle\,|i \in [k]\}$ and $\Pi_{\mathcal{U}}$ be the orthogonal projection onto $\mathcal{U}$. Let K be a Kraus operator of a filter,

(i) The postselected Fisher information $J(\cdot|\rho_K^{\mathrm{ps}}) : \Theta \to \mathcal{M}_k(\mathbb{R})$ only depends on $F_u := \Pi_{\mathcal{U}} K^\dagger K \Pi_{\mathcal{U}}$, not K.

(ii) K is lossless at θ iff. $F_u = (p-1)\,|\psi(\theta)\rangle\langle\psi(\theta)| + \Pi_{\mathcal{U}}$, for some $p \in (0, 1]$.

Proof: (i) follows immediately from proposition 2.11.
We note that for all $\theta \in \Theta$, $\langle\psi(\theta)|\psi(\theta)\rangle = 1$. Taking the derivative of this expression, we see that for all $\theta \in \Theta$ and $j \in [k]$,

$$\mathrm{Re}(\langle\partial_j\psi(\theta)|\partial\psi(\theta)\rangle) = 0. \tag{2.125}$$

Fix $\theta \in \Theta$, in view of equation (2.125), we may expand

$$|\partial_j\psi(\theta)\rangle = ix_j\,|\psi(\theta)\rangle + \alpha_j\left|\psi_j^{\perp}(\theta)\right\rangle, \tag{2.126}$$

where $x_j \in \mathbb{R}$, $\alpha_j \in \mathbb{C}$ and $\left|\psi_j^\perp(\theta)\right\rangle \in S(\mathcal{H})$ satisfies $\left\langle\psi(\theta)\middle|\psi_j^\perp(\theta)\right\rangle = 0$. We define

$$\langle\psi(\theta)|F_u|\psi(\theta)\rangle = p^{\text{ps}}_{\rho,\,K}(\theta) \in [0,1], \tag{2.127}$$

$$\left\langle\psi_j^\perp(\theta)\middle|F_u\middle|\psi_j^\perp(\theta)\right\rangle = B_j \in [0,1], \tag{2.128}$$

$$\left\langle\psi(\theta)\middle|F_u\middle|\psi_j^\perp\right\rangle = C_j \in \mathbb{C}. \tag{2.129}$$

Using proposition 2.11, we find that for $j \in [k]$,

$$J(\theta|\rho^{\text{ps}}_K)_{j,j} = \frac{4}{p^{\text{ps}}_{\rho,\,K}(\theta)}\left(x_j^2 p^{\text{ps}}_{\rho,\,K}(\theta) + |\alpha_j|^2 + ix_j[\alpha_j^* C_j^* - \alpha_j C_j]\right) - \frac{4}{p^{\text{ps}}_{\rho,\,K}(\theta)^2}|ix_j p^{\text{ps}}_{\rho,\,K}(\theta) + \alpha_j C_j|^2, \tag{2.130}$$

$$= 4\left[\frac{|\alpha_j|^2 B_j}{p^{\text{ps}}_{\rho,\,K}(\theta)} - \frac{|\alpha_j C_j|^2}{p^{\text{ps}}_{\rho,\,K}(\theta)^2}\right]. \tag{2.131}$$

By considering the case of $K = F = \mathbb{1}$, we see that

$$J(\theta|\rho) = 4|\alpha_j|^2. \tag{2.132}$$

We deduce $p^{\text{ps}}_{\rho,\,K}(\theta)J(\theta|\rho^{\text{ps}}_K)_{j,j} = J(\theta|\rho)_{j,j}$ iff.

$$|\alpha_j|^2(B_j - 1) - \frac{|\alpha_j C_j|^2}{p^{\text{ps}}_{\rho,\,K}(\theta)} = 0. \tag{2.133}$$

Since $\alpha_j \neq 0$ by assumption, we deduce that this condition holds iff. $B_j = 1$ and $C_j = 0$. Note that $\mathbb{1} - F \geq 0$, so using proposition 1.1 (i), we see that the diagonal entries of $p^{\text{ps}}_{\rho,\,K}(\theta)J(\theta|\rho^{\text{ps}}_K)$ and $J(\theta|\rho)$ are equal iff.

$$F_u = (p^{\text{ps}}_{\rho,\,K}(\theta) - 1)\,|\psi(\theta)\rangle\langle\psi(\theta)| + \Pi_{\mathcal{U}}. \tag{2.134}$$

To prove point (ii) it remains to show that for such F_u, the off diagonal entries of $J(\theta|\rho^{\text{ps}}_K)$ and $J(\theta|\rho)$ are equal. This is a straightforward calculation. □

The condition that $J(\theta|\rho)_{j,j}$ be non-zero for all $j \in [k]$ is a natural requirement - it is equivalent to saying that there is information on θ contained in ρ to begin with, so that it can be compressed. Indeed, if $J(\theta|\rho)_{j,j} = 0$ for a single $j \in [k]$, then there exists a vector $v \in \mathbb{R}^k$ such that $v^T J(\theta|\rho)v = 0$, and thus by proposition 1.1 (i), $J(\theta|\rho)$ is non-invertible.

It may be that $\mathcal{U}$ has dimension u less than $d = \dim\mathcal{H}$. For example, this will always happen if $k + 1 < d$, where k is the number of parameters: $\Theta \subseteq \mathbb{R}^k$. In general, if $u < d$, then there is not a unique lossless filter. For a fixed $\theta \in \Theta$, the subspace $\mathcal{U}$ encodes the local changes in the state $|\psi\rangle$ that can arise from infinitesimal variations in the parameter, see figure 2.4 for a intuitive discussion. We note that the JAL filter is always lossless, regardless of $\mathcal{U}$. In this sense, it is thus the canonical choice of lossless filter.

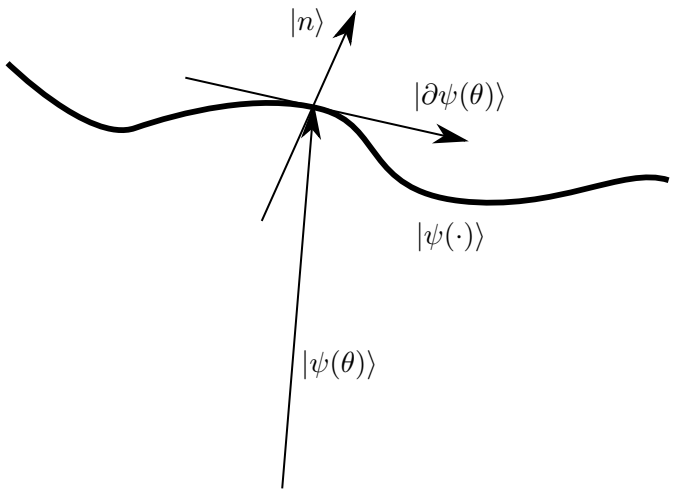


Figure 2.4: Schematic diagram of a Hilbert space with a single-parameter parameterised quantum state $|\psi(\cdot)\rangle$. At the point $\theta \in \Theta$, we see that there is one direction in $\mathcal{H}$ arising from infinitesimal variations in the parameter: the tangent $|\partial\psi(\theta)\rangle$. The third direction $|n\rangle$ is (locally) unimportant.

2.3.2 Non-optimality of the JAL Filter with Noise

Filtering of mixed states is more complicated than pure states, see Ref. [2] for a discussion of filtering pure states that have been acted upon by depolarising noise. One finds that unbounded amplification of information is no longer possible, and that lossless compression may not exist. In this Section, we give an explicit example where the JAL filter is not optimal for a parameterised mixed state. Our parameterised mixed state is a depolarised parameterised pure state in a 3-dimensional Hilbert space: $\dim \mathcal{H} = 3$.

First, we prove a useful result for calculating the QFI of postselected states in the single parameter case

Proposition 2.12: Suppose that $\rho : \Theta \to \mathcal{D}(\mathcal{H})$ is a single parameter ($\Theta \subseteq \mathbb{R}$) parameterised quantum state, and that $p : \Theta \to (0, 1]$ is a parameterised probability. Let $\rho' : \Theta \to \mathcal{B}(\mathcal{H})$ be defined by

$$\rho = \frac{\rho'}{p}. \tag{2.135}$$

Then, for all $\theta \in \Theta$,

$$J(\theta|\rho) = \frac{1}{p(\theta)}\left[J(\theta|\rho') - \frac{(\mathrm{Tr}[\partial\rho'(\theta)])^2}{p(\theta)}\right]. \tag{2.136}$$

Here, we extend the definition of the QFI to include non-normalised positive semi-definite operators in the natural way.

Proof: We calculate that

$$\partial\rho = \frac{\partial\rho'}{p} - \frac{\partial p}{p}\rho. \tag{2.137}$$

Fix $\theta \in \Theta$, let $\Lambda = L^{\mathrm{SLD}}(\theta)$. We define a new operator Λ' by the equation

$$\Lambda = \Lambda' - \frac{\partial p(\theta)}{p(\theta)}\mathbb{1}. \tag{2.138}$$

Using the defining property of Λ, we see that Λ' satisfies the reduced equation

$$\frac{1}{2}\{\Lambda', \rho'(\theta)\} = \partial\rho'(\theta), \tag{2.139}$$

and thus is a choice of SLD for ρ'. We expand the QFI of ρ as

$$\begin{aligned} J(\theta|\rho) &= \mathrm{Tr}\big(\rho(\theta)\Lambda^2\big), && (2.140)\\ &= \mathrm{Tr}\left(\frac{1}{2}\{\rho(\theta), \Lambda\}\Lambda\right), && (2.141)\\ &= \mathrm{Tr}(\partial\rho(\theta)\Lambda), && (2.142)\\ &= \mathrm{Tr}\left[\left(\frac{\partial\rho'(\theta)}{p(\theta)} - \frac{\partial p(\theta)}{p(\theta)}\rho(\theta)\right)\left(\Lambda' - \frac{\partial p(\theta)}{p(\theta)}\mathbb{1}\right)\right], && (2.143)\\ &= \frac{1}{p(\theta)}\left[J(\theta|\rho') - \frac{\partial p(\theta)}{p(\theta)}\mathrm{Tr}\big(\rho'(\theta)\Lambda'\big) - \frac{\partial p(\theta)}{p(\theta)}\mathrm{Tr}\big[\partial\rho'(\theta)\big] + \frac{[\partial p(\theta)]^2}{p(\theta)}\right]. && (2.144) \end{aligned}$$

Taking the trace of equation (2.137) (noting that $\mathrm{Tr}(\rho)$ is constant, and therefore $\mathrm{Tr}(\partial\rho)$) vanishes), we find that

$$\mathrm{Tr}\big(\partial\rho'(\theta)\big) = \partial p(\theta). \tag{2.145}$$

Substituting this into equation (2.144) and using that $\mathrm{Tr}(\rho'(\theta)\Lambda') = \mathrm{Tr}(\{\rho'(\theta), \Lambda'\})/2 = \mathrm{Tr}[\partial\rho'(\theta)]$, the result follows. □

We now consider the effect of depolarising noise on the JAL filter. We consider some parameterised pure state $|\psi(\cdot)\rangle : \Theta \to S(\mathcal{H})$, which is affected by depolarising noise. For notational brevity, we consider a noise rate of $1/2$, so that the preselected state has the form

$$\rho : \Theta \to \mathcal{D}(\mathcal{H}),\ \rho(\theta) = \frac{1}{2}(|\psi(\theta)\rangle\langle\psi(\theta)| + \mathbb{1}/3), \tag{2.146}$$

Fix a filter with Kraus operator K, which gives rise to the non-normalised postselected state

$$\rho' : \Theta \to \mathcal{B}(\mathcal{H}),\ \rho'(\theta) = K\,|\psi(\theta)\rangle\langle\psi(\theta)|\,K^\dagger + KK^\dagger/3. \tag{2.147}$$

Fix some $\theta \in \Theta$. As in the proof of Lemma 2.2, we decompose

$$|\partial\psi(\theta)\rangle = ix\,|\psi(\theta)\rangle + \alpha\left|\psi^\perp\right\rangle, \tag{2.148}$$

where $x \in \mathbb{R}, \alpha \in \mathbb{C}$, and $\left|\psi^\perp\right\rangle$ is normalised and orthogonal to $|\psi(\theta)\rangle$. We further assume that $\alpha \neq 0$, i.e. that $J(\theta|\psi) \neq 0$, so that there is information to compress. Since $\mathcal{H}$ is 3-dimensional, we can find a state $|n\rangle \in S(\mathcal{H})$ such that $\{|\psi(\theta)\rangle, \left|\psi^\perp\right\rangle, |n\rangle\}$ is an orthonormal basis of $\mathcal{H}$. In this basis, consider a filter with a single off-diagonal term:

$$K = \begin{pmatrix} t & 0 & b \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{2.149}$$

where $t, b \in [0,1]$. K has POVM operator

$$F = K^\dagger K = \begin{pmatrix} t^2 & 0 & tb \\ 0 & 1 & 0 \\ tb & 0 & b^2 \end{pmatrix}. \tag{2.150}$$

The eigenvalues of F are $0, 1, t^2 + b^2$. Thus, for $b \in [0, \sqrt{1-t^2}]$, we find that $0 \leq F \leq \mathbb{1}$ and therefore K is a valid filter. One can explicitly calculate

$$\rho'(\theta) = \begin{pmatrix} (4/3)t^2 & 0 & tb/3 \\ 0 & 1/3 & 0 \\ tb/3 & 0 & b^2/3 \end{pmatrix}, \quad \text{and} \quad \partial\rho'(\theta) = \begin{pmatrix} 0 & \alpha^* t & 0 \\ \alpha t & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{2.151}$$

We also explicitly find Λ':

$$\Lambda' = \begin{pmatrix} 0 & \Lambda'^*_{12} & 0 \\ \Lambda'_{21} & 0 & \Lambda'^*_{23} \\ 0 & \Lambda'_{32} & 0 \end{pmatrix}, \tag{2.152}$$

We note that $\partial\rho'(\theta)$ is traceless, and thus by proposition 2.12

$$\mathcal{R}_{\rho,\,K}(\theta) = J(\theta|\rho'). \tag{2.153}$$

Finally, we calculate

$$\Lambda'^*_{12} = \Lambda'_{21} = \frac{6(1+b^2)t\alpha}{1+b^2+(4+3b^2)t^2}, \tag{2.154}$$

$$\Lambda'^*_{23} = \Lambda'_{32} = -\frac{6bt^2\alpha}{1+b^2+(4+3b^2)t^2}. \tag{2.155}$$

We can further calculate

$$\mathcal{R}_{\rho,\,K}(\theta) = J(\theta|\rho') = \frac{12|\alpha|^2(1+b^2)t^2}{1+b^2+(4+3b^2)t^2}, \tag{2.156}$$

$$= \frac{12|\alpha|^2 t^2}{1+\left(3+\dfrac{1}{1+b^2}\right)t^2}, \tag{2.157}$$

which is manifestly increasing in b. Surprisingly, we see that the off-diagonal term b is beneficial to filter performance. Therefore, it is, counter-intuitively, beneficial to mix the noisy subspace spanned by $|n\rangle$ with $\mathcal{U}$.

We compare K to a diagonal Kraus operator, a generalisation of the JAL filter

$$\tilde{K} = \begin{pmatrix} t & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & r \end{pmatrix}, \tag{2.158}$$

where $r \in [0,1]$ controls the probability of transmitting an $|n\rangle$ state (which can only come from noise).

When $r = 1$, $\tilde{K}$ is the JAL filter. Similarly, one can calculate

$$\mathcal{R}_{\rho,\tilde{K}}(\theta) = \frac{12|\alpha|^2 t^2}{1+4t^2}, \tag{2.159}$$

As r is increased, the postselection probability increases, but the information in a transmitted states decreases. These two effects cancel, so that $\mathcal{R}_{\rho,\tilde{K}}(\theta)$ does not depend on r. Comparing equations (2.157) and (2.159), we see that $\mathcal{R}_{\rho,K}(\theta) \geq \mathcal{R}_{\rho,\tilde{K}}(\theta)$, with equality iff. $b = 0$. Thus, the JAL filter is only optimal in the degenerate case $t = 1$ (no filtering), for all $t < 1$ it is outperformed by K. We conclude that the JAL filter may not be optimal for mixed states. Finding such optimal filters remains an open question.

2.3.3 A New Perspective on Classical Filtering

Ref. [46] considers a quasi-probabilitic description of quantum filtering, using the (extended) Kirkwood-Dirac (KD) distribution [55]. They showed that if the postselected QFI $J(\theta|\rho_K^{\text{ps}})$ passes a certain threshold, then the KD-distribution must contain some negativity. They label such negativity as a hallmark of quantum effects, given the well-established nature of the KD distribution in quantum foundations. Note, if the required postselection probability is sufficiently small, then by this definition of quantumness, any lossless filter implies quantum effects.

In this Section, we consider a more operational approach to proving that lossless compression is a quantum effect, and consider classical filtering of arbitrary random variables. We show that classical filtering can lead to an increase in postselected FI, but that this filtering is almost never lossless.

Let $X : \Omega \to \chi$ be some random variable, with parameter-dependent law P_θ. We assume that P_θ admits a pdf. $f(x|\theta)$. A classical filter is defined by a (measurable) function $t^2 : \chi \to [0,1]$, where $t^2(x)$ is the probability that x passes the filter. We call the function t^2 a (classical) filter. A sample of the "postselected" variable Y can be obtained via the following procedure:

1. Sample X, giving outcome $x \in X$.
2. Generate a number u uniformly in $[0,1]$, if $t^2(x) \leq u$, then output x. Otherwise return to step 1.

We calculate the probability that a single sample of X passes the filter, i.e. the probability of transmission:

$$p(\theta) = \int_\chi f(x|\theta) t^2(x) dx. \tag{2.160}$$

We require that $p(\theta) > 0$ for every $\theta \in \Theta$, so that (almost surely) the procedure to generate a sample of Y terminates in a finite number of steps.

In analogy with the quantum case, we say that a classical filter is lossless at $\theta \in \Theta$ if $p(\theta)I(Y|\theta) = I(X|\theta)$. We will find that, in contrast to the quantum setting, lossless classical compression of information via filtering is almost always impossible.

Proposition 2.13: Let $X : \Omega \to \chi$ be a random variable which admits a parameter dependent pdf. $f(x|\theta)$, and let $t^2 : \chi \to [0,1]$ be a filter. The postselected variable Y admits a pdf

$$g(y|\theta) = \frac{f(x|\theta)}{p(\theta)} t^2(y), \tag{2.161}$$

where $p(\theta)$ is the probability of transmission, as defined in equation (2.160).

Proof: Let $B \subseteq \chi$ be a measurable subset. Then,

$$\mathbb{P}_\theta(Y \in B) = \int_B f(x|\theta) t^2(x) dx + (1 - p(\theta)) \mathbb{P}_\theta(Y \in B). \tag{2.162}$$

Rearranging, we find that

$$\mathbb{P}_\theta(Y \in B) = \int_B \frac{f(x|\theta)}{p(\theta)} t^2(x) dx, \tag{2.163}$$

and the result follows. □

Proposition 2.14: Let $X : \Omega \to \chi$ be a random variable which admits a parameter dependent pdf. $f(x|\theta)$, and let $t^2 : \chi \to [0,1]$ be a filter. Denote the postselected variable by Y. Fix $\theta \in \Theta$, then

$$p(\theta) I(\theta|Y) = \int_\chi \frac{[\partial_\theta f(y|\theta)]^2}{f(y|\theta)} t^2(y) dy - \frac{[\partial_\theta p(\theta)]^2}{p(\theta)}, \tag{2.164}$$

where $p(\theta)$ is the probability of transmission.

Proof: First, we calculate the derivative of the pdf. of Y (see proposition 2.13):

$$\partial_\theta g(y|\theta) = \left[\partial_\theta f(y|\theta) - \frac{f(y|\theta)}{p(\theta)} \partial_\theta p(\theta) \right] \frac{t^2(y)}{p(\theta)}. \tag{2.165}$$

We can then directly calculate

$$\begin{aligned}
p(\theta) I(\theta|Y) &= \int_\chi p(\theta) \frac{[\partial_\theta g(y|\theta)]^2}{g(y|\theta)} dy, && (2.166)\\
&= \int_\chi \left[\partial_\theta f(y|\theta) - \frac{f(y|\theta)}{p(\theta)} \partial_\theta p(\theta) \right]^2 \frac{t^2(y)}{f(y|\theta)} dy, && (2.167)\\
&= \int_\chi \frac{[\partial_\theta f(y|\theta)]^2}{f(y|\theta)} t^2(y) dy && (2.168)\\
&\quad - 2 \frac{\partial_\theta p(\theta)}{p(\theta)} \int_\chi \partial_\theta f(y|\theta) t^2(y) dy + \frac{[\partial_\theta p(\theta)]^2}{p(\theta)^2} \int_\chi f(y|\theta) t^2(y) dy. && (2.169)
\end{aligned}$$

Finally, we note that $p(\theta) = \int_\chi f(y|\theta) t^2(y) dy$ and $\partial_\theta p(\theta) = \int_\chi \partial_\theta f(y|\theta) t^2(y) dy$, and the result follows. □

For a (measureable) subset $A \subseteq \chi$, we define the information contained in A (with respect to X) $I_A(X|\theta)$ as its contribution to the total Fisher information:
We can now characterize the set of lossless classical filters.

Lemma 2.24: Let $t^2 : \chi \to [0,1]$ be a filter on a random variable $X : \Omega \to \chi$. Then t^2 is lossless at $\theta \in \Theta$ iff. $\{x \in X : \partial_\theta f(x|\theta) \neq 0,\ t^2(x) < 1\}$ has measure 0.

Proof: Fix $\theta \in \Theta$, using proposition 2.14, we know that

$$p(\theta)I(\theta|Y) = \int_\chi \frac{[\partial_\theta f(y|\theta)]^2}{f(y|\theta)} t^2(y)dy - \frac{[\partial_\theta p(\theta)]^2}{p(\theta)}. \tag{2.170}$$

Let $Z = \{x \in X : \partial_\theta f(x|\theta) \neq 0,\ t^2(x) < 1\}$. We find that

$$\begin{aligned} p(\theta)I(\theta|Y) &\leq \int_\chi \frac{[\partial_\theta f(y|\theta)]^2}{f(y|\theta)} t^2(y)dy, && (2.171)\\ &= \int_Z \frac{[\partial_\theta f(y|\theta)]^2}{f(y|\theta)} t^2(y)dy + \int_{Z^c} \frac{[\partial_\theta f(y|\theta)]^2}{f(y|\theta)} t^2(y)dy, && (2.172)\\ &\leq \int_Z \frac{[\partial_\theta f(y|\theta)]^2}{f(y|\theta)} dy + \int_{Z^c} \frac{[\partial_\theta f(y|\theta)]^2}{f(y|\theta)} dy, && (2.173)\\ &= I(\theta|X). && (2.174) \end{aligned}$$

We have equality in (2.171) iff. $\partial_\theta p(\theta) = 0$. The inequality (2.173) is saturated iff. Z has measure 0. Note that if Z has meausre 0, then $\partial_\theta p(\theta) = \int \partial_\theta f(x|\theta) t^2(x) dx = \int \partial_\theta f(x|\theta) = 0$. We conclude that Z having zero measure is necessary and sufficient for the filter t^2 to be lossless. □

Lemma 2.24 shows that classical filtering is lossless if and only if one filters on a subset that has no local parameter dependence. If, as is often the case, $\partial_\theta f(x|\theta)$ is non-vanishing for all $x \in \chi$, then only the trivial filter $t^2 = 1$ (almost everywhere) is lossless. This is in stark contrast to the quantum case, where one can find a lossless filter for any choice of (non-zero) transmission probability $p(\theta)$. This demonstrates the truly quantum nature of filtering in an operational fashion, without reference to any other preconceived notions of classicality.

2.3.4 Iterative Filtering

As noted in Section 2.1.7, existing work on postselected metrology has focused on theoretical calculations using the quantum Fisher information, rather than concrete protocols. In this Section, we introduce the first explicit postselected quantum parameter estimation scheme. We give an algorithm for a specific parameterised quantum state, where one starts with no prior knowledge on the parameter to be estimated, that repeatedly updates the quantum filter. As the filter is updated, the risk of the corresponding estimator (the output if the algorihtm was terminated) decreases arbitrarily below its non-postselection version, demonstrating the advantage that postselection can give. Additionally, we consider the application of the James-Stein estimator (see example 2.2) to postselection. We see that the (genuinely quantum) effect of postselection can boost the James-Stein estimator's advantage over the MLE.

Unlike other Sections, we take $\mathcal{H}$ to be infinite dimensional, and consider an explicit parameterised pure state. Specifically, for some $k \in \mathbb{N}$, we take

$$\mathcal{H} = L^2(\mathbb{R}^k, \mathbb{C}) := \left\{ f : \mathbb{R}^k \to \mathbb{C} \;\middle|\; \int d^k x\, |f(x)|^2 < \infty \right\} / \sim, \tag{2.175}$$

where $f \sim g$ iff. $f = g$ almost everywhere. We consider a parameterised state with parameter space $\Theta = \mathbb{R}^k$:

$$|\psi_\cdot\rangle : \mathbb{R}^k \to S(\mathcal{H}),\ \theta \mapsto |\psi_\theta\rangle\ ,\ |\psi_\theta\rangle\,(x) = C^{k/2} e^{-\|x-\theta\|^2/B}, \tag{2.176}$$

for some known $B > 0$ and $C = \sqrt{2/\pi B}$. In the position basis, $|\psi_\theta\rangle$ is a Gaussian, with mean θ and variance $\sqrt{B/2}$.

Suppose that we use filter with Kraus operator K on $|\psi\rangle$. We denote the postselected state by

$$\left|\psi^K\right\rangle : \Theta \to S(\mathcal{H}),\ \theta \mapsto \left|\psi_\theta^K\right\rangle = \frac{K\,|\psi_\theta\rangle}{\|K\,|\psi_\theta\rangle\|}. \tag{2.177}$$

Given its theoretical properties, we will use the JAL filter, with parameters $\theta_0 \in \Theta$ and $t \in [0, 1]$. We will show that the postselected state can be well approximated as a Gaussian state, for θ_0 close to θ. We require a brief technical proposition to do so.

Proposition 2.15: Let $a \in \mathbb{R}^k$, then

$$\sup_{x \in \mathbb{R}^k} e^{-\|x\|^2} |e^{a\cdot x} - (1 + a \cdot x)| = O(\|a\|^2), \tag{2.178}$$

as $a \to 0$.

Proof: From Lagrange's form of the remainder of a Taylor series [56], we know that

$$|e^{a\cdot x} - (1 + a \cdot x)| = \frac{1}{2}(a \cdot x)^2 e^\xi, \tag{2.179}$$

where $\xi \in \mathbb{R}$ is bounded: $|\xi| \leq |a \cdot x|$. By Cauchy-Schwartz, we deduce that

$$|e^{a\cdot x} - (1 + a \cdot x)| \leq \frac{1}{2}\|a\|^2 \|x\|^2 e^{\|a\|\|x\|}. \tag{2.180}$$

Using that $\|x\|^2 \leq e^{\|x\|}$, we find that

$$\begin{aligned} e^{-\|x\|^2} |e^{a\cdot x} - (1 + a \cdot x)| &\leq \frac{1}{2}\|a\|^2 \|x\|^2 e^{\|a\|\|x\| - \|x\|^2}, && (2.181)\\ &\leq \frac{1}{2}\|a\|^2 e^{(1+\|a\|)\|x\| - \|x\|^2}, && (2.182)\\ &= \frac{1}{2}\|a\|^2 e^{-(\|x\| - (1+\|a\|)/2)^2 + (1+\|a\|)^2/4}, && (2.183)\\ &\leq \frac{1}{2}\|a\|^2 e^{(1+\|a\|)^2/4}. && (2.184) \end{aligned}$$

Taking the supremum of both sides, the result follows. □

We now show an approximation for the postselected state. Suppose that the true value of the parameter is $\theta \in \Theta$, and let $\delta = \theta - \theta_0$. We first show following approximation

Proposition 2.16: Suppose $K = K_\theta^{\mathrm{JAL}}$, and $|\psi_\theta\rangle$ is defined as in equation (2.176). Then, for $x \in \mathbb{R}^k$

$$\left|\psi_\theta^K\right\rangle(x) = C^{k/2} e^{-\|x - \theta_0 - \delta/t\|^2/B} [1 + O(\|\delta\|^2/t^2)], \tag{2.185}$$

(almost everywhere). Furthermore, the probability of postselection satisfies

$$p^{\mathrm{ps}}_{\rho,K}(\theta) = t^2 + O(\|\delta\|^2). \tag{2.186}$$

Proof: First, we calculate the overlap

$$\langle\psi_{\theta_0}|\psi_\theta\rangle = C^k \int_{-\infty}^{\infty} d^k x \;\; e^{-(\|x-\theta\|^2+\|x-\theta_0\|^2)/B}, \tag{2.187}$$

$$= C^k \int_{-\infty}^{\infty} d^k x \;\; e^{-(2\|x-\theta/2-\theta_0/2\|^2+\|\theta-\theta_0\|^2/2)/B}, \tag{2.188}$$

$$= e^{-\|\delta\|^2/2B}. \tag{2.189}$$

We deduce that $K|\psi_\theta\rangle = |\psi_\theta\rangle - (1-t)e^{-\|\delta\|^2/2B}|\psi_{\theta_0}\rangle$. Define $f \in \mathcal{H}$ by $f(x) = K|\psi(\theta)\rangle(x+\theta_0)$. For $x \in \chi$, we find that

$$f(x) = C^{k/2}(e^{-\|x-\delta\|^2/B} - (1-t)e^{-\|\delta\|^2/2B}e^{-\|x\|^2/B}), \tag{2.190}$$

$$= C^{k/2}e^{-\|\delta\|^2/B}e^{-\|x\|^2/B}[e^{2x\cdot\delta/B} - (1-t)e^{\|\delta\|^2/2B}]. \tag{2.191}$$

Using proposition 2.15 twice, we find that

$$f(x) = C^{k/2}e^{-\|\delta\|^2/B}e^{-\|x\|^2/B}[t + 2x\cdot\delta/B + O(\|\delta\|^2)], \tag{2.192}$$

$$= tC^{k/2}e^{-\|\delta\|^2/B}e^{-\|x\|^2/B}[1 + 2x\cdot\delta/tB + O(\|\delta\|^2/t^2)], \tag{2.193}$$

$$= tC^{k/2}e^{-\|\delta\|^2/B}e^{-\|x\|^2/B}e^{2x\cdot\delta/tB}[1 + O(\|\delta\|^2/t^2)], \tag{2.194}$$

$$= tC^{k/2}e^{-\|x-\delta/t\|^2/B}e^{\|\delta\|^2(1/t^2-1)/B}[1 + O(\|\delta\|^2/t^2)], \tag{2.195}$$

$$= tC^{k/2}e^{-\|x-\delta/t\|^2/B}[1 + O(\|\delta\|^2/t^2)]. \tag{2.196}$$

Finally, we note that $p^{\mathrm{ps}}_{\rho,K}(\theta) = \|K|\psi_\theta\rangle\|^2 = \int dx|f(x)|^2$. □

Proposition 2.16 may be seen as a non-asymptotic version of Lemma 2.15 for the particular example of a Gaussian state. The amplification of J by a factor of $1/t^2$ is demonstrated by the increased sensitivity of the postselected state to changes in θ (by a factor of $1/t$). We will show how to use this approximation to construct an explicit protocol for estimating θ.

Because we require that $t \gg \|\delta\|$, one cannot decrease t (increasing the strength of the filter) until one has a good estimate θ_0 of θ. We start with no knowledge of θ, and hence no suitable guess θ_0. We run an iterative algorithm (described below), where one estimates δ (using sample variance) and then decreases t when one is confident that δ is sufficiently small. We remark that this algorithm is heuristic, we do not explicitly prove its convergence properties. We do, however, emperically demonstrate its performance. . In this way, the Gaussian approximation of proposition 2.16 is expected to hold at all steps in the algorithm.

Algorithm 2.1:

Input: An unknown parameter value $\theta \in \mathbb{R}^k$ and $N \in \mathbb{N}$, the number of measurements to be taken

Output: An estimate of the parameter θ

1. Set $\theta_0 = 0$, $t = 1$, $i = 1$
2. Measure the position of the postselected state $\left|\psi_\theta^{K_{\theta_0}^{\mathrm{JAL}}}\right\rangle$, denote the outcome X_i. We let $Y_i = \theta_0 + t(X_i - \theta_0)$. By proposition 2.16, if $\|\delta\| \ll t$, we expect that $Y_i \sim \mathcal{N}(\theta, Bt^2/4\mathbb{1})$,
3. Let $\hat{\theta}_i = \hat{\theta}(Y_i)$ be an estimate of θ made using Y_i. Update a collated estimate given by $\bar{\theta}_i$, the sample mean of $\hat{\theta}_1, \ldots, \hat{\theta}_i$.
4. Let $\bar{\theta}_i$ be the sample mean of $\hat{\theta}_1, \ldots, \hat{\theta}_i$, a collated estimate of θ. Let $\hat{\sigma}_i$ be the sample standard deviation of $\hat{\theta}_1, \ldots, \hat{\theta}_i$. We let $\hat{\delta}_i = \hat{\sigma}_i/\sqrt{i}$, an estimate of δ
5. If δ_i is less than $0.3t$, we set t to $3\hat{\delta}_i$ and θ_0 to $\bar{\theta}_i$: our current best estimate of θ.
6. If $i = N$, then output $\bar{\theta}_i$ as our estimate of θ. Otherwise, increase i by 1 and return to step 2.

We reiterate that we are working with the approximation that the limiting factor in estimating θ is measuring the states $\left|\psi_\theta^{K_{\theta_0}^{\mathrm{JAL}}}\right\rangle$. As one decreases t, decreasing the chance of successful postselection, one can increase the intensity of probes input to the system to produce states $\left|\psi_\theta^{K_{\theta_0}^{\mathrm{JAL}}}\right\rangle$.

Algorithm 2.1 relies on an estimator $\hat{\theta} : \mathbb{R}^k \to \mathbb{R}^k$ for estimating the mean of an unknown Gaussian distribution $\mathcal{N}(\theta, \Sigma)$, with known covariance matrix Σ. We first consider the algorithm with the MLE: $\hat{\theta}^{\mathrm{MLE}}(x) = x$, and denote the corresponding collated estimates $\bar{\theta}_n$ by $\hat{\theta}_n^{\mathrm{pMLE}}$ (shorthand for postselected MLE).

Suppose that when X_i is sampled, t has value t_i. Under the Gaussian approximation of proposition 2.16, we thus expect

$$\hat{\theta}_n^{\mathrm{pMLE}} \sim \mathcal{N}\left(\theta, \frac{B}{4n}\,\bar{t}_n^2\,\mathbb{1}\right), \quad R(\hat{\theta}^{\mathrm{pMLE}}, \theta) = \frac{kB}{4n}\,\bar{t}_n^2, \tag{2.197}$$

where $\bar{t}_n^2 = (1/n)\sum_{i=1}^n t_i^2$ is the average value of t^2. For comparison $\hat{\theta}_n^{\mathrm{MLE}} \sim \mathcal{N}(\theta, B/4n\,\mathbb{1})$ and $R(\hat{\theta}_n^{\mathrm{MLE}}, \theta) = kB/4n$. We see that postselection improves the risk by a factor of $\bar{t}_n^2$.

We expect $t_n \to 0$ as $n \to \infty$ and thus, by Cesàro's Lemma, we expect $\bar{t}_n^2 \to 0$, and that postselection gives an unbounded advantage in estimating θ. Of course, in practice, one cannot truly decrease $t \to 0$, as this would make the postelection probability arbitrarily small. There will be some lower bound $t \geq t_{\min}$ corresponding to the maximum intensity at which one can input probes to a system of interest. In this case, $\bar{t}_n^2 \to t_{\min}^2$, and postselection leads to a constant (but potentially very large) improvement.

An explicit demonstration of the advantage from postselection is seen in figure 2.5. After an initial period (in which an initial estimate θ_0 is collected), we see that the advantage of postselection appears to increase linearly (and unboundedly) with the number of samples.

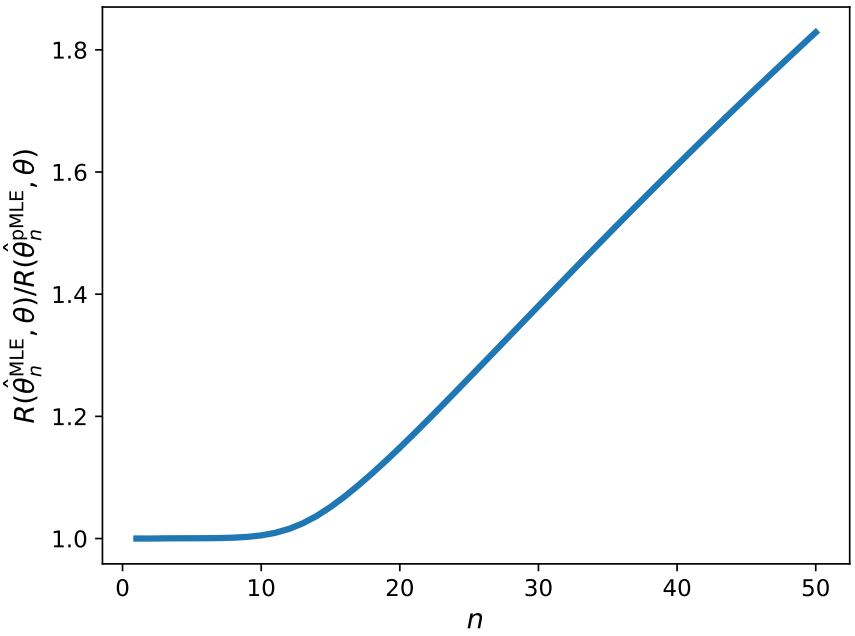


Figure 2.5: Plot showing the advantage of postselection when using the MLE. Here, $k = 4$, $B = 1$ and $\theta = (1, -2, 3, 1.5)/30$.

As shown by example 2.2, the MLE is inadmissible for estimating the mean of a normal distribution when $k \geq 3$. Thus, we consider our protocol using the James-Stein estimator (see example 2.2). However, we note that $\hat{\theta}^{\mathrm{JS}}$ only has a non-negligible advantage over $\hat{\theta}^{\mathrm{MLE}}$ when θ is close to some known value[7] $\theta_0 \in \mathbb{R}^k$. But in this setting, we can immediately decrease t, which will shift θ to $\theta_0 + (\theta_0 - \theta)/t$, which is far from θ_0 and thus destroys the JS advantage. Thus, we consider a modified version of the James-Stein estimator $\hat{\theta}^{\mathrm{mJS}}$ that can be applied to our scenario. Instead of shrinking the vector x towards the origin like $\hat{\theta}^{\mathrm{JS}}$, $\hat{\theta}^{\mathrm{mJS}}$ shrinks x in the direction of the isotropic vector $e = (1, \ldots, 1) \in \mathbb{R}^k$. To be precise, for $x \in \mathbb{R}^k$, we define $x_m = (\sum_{i=1}^k x_i/k)e \in \mathbb{R}^k$, and for $k \geq 4$ we define $\hat{\theta}^{\mathrm{mJS}}$ as

$$\hat{\theta}^{\mathrm{mJS}}(x) = x - \frac{(k-3)\Sigma^{-1}}{\|\Sigma^{-1}(x - x_m)\|^2}(x - x_m). \tag{2.198}$$

The risk of $\hat{\theta}^{\mathrm{mJS}}$ is given by

$$R(\hat{\theta}^{\mathrm{mJS}}, \theta) = \mathrm{Tr}(\Sigma) - (k-3)^2 \mathbb{E}\left[\frac{1}{\|\Sigma^{-1}(X - X_m)\|^2}\right], \tag{2.199}$$

so that $\hat{\theta}^{\mathrm{mJS}}$ also dominates $\hat{\theta}^{\mathrm{MLE}}$. In analogy to $\hat{\theta}^{\mathrm{pMLE}}$, we denote the collated estimates $\bar{\theta}_n$ when using $\hat{\theta} = \hat{\theta}^{\mathrm{mJS}}$ by $\hat{\theta}_n^{\mathrm{pmJS}}$.

In order to consider the effect of postselection on the JS advantage, we consider the "postselected

[7] By shifting the origin, we can take this value to be non-zero.

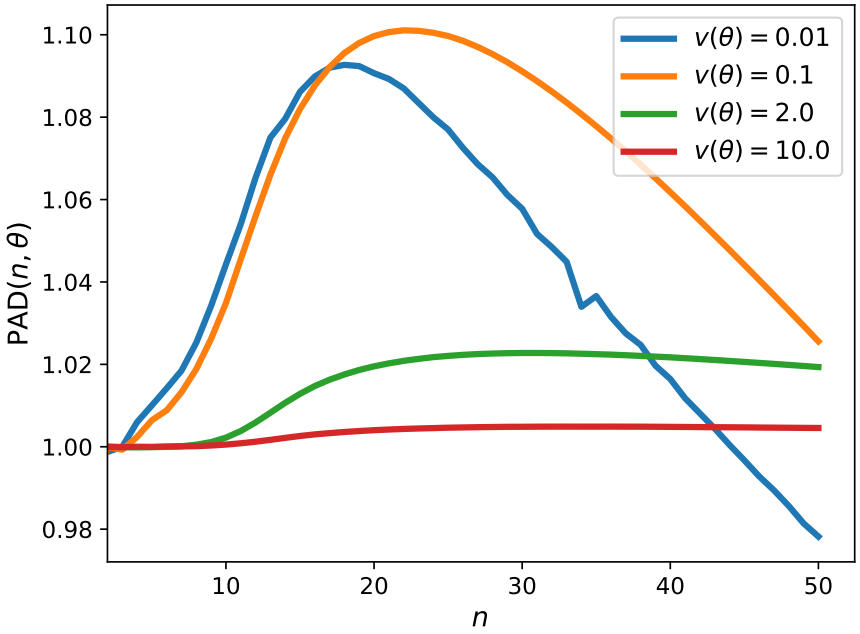


Figure 2.6: Comparison of the effect of postselection on the James-Stein advantage for various θ, $B = 1$. The small kink in the blue curve is due to computational shot noise.

advantage" (PAD) given by the ratio

$$\text{PAD} : \mathbb{N} \times \mathbb{R}^k \to [0, \infty), \ \text{PAD}(n, \theta) = \frac{R(\hat{\theta}_n^{\text{pMLE}}, \theta)}{R(\hat{\theta}_n^{\text{pmJS}}, \theta)} \cdot \frac{R(\hat{\theta}_n^{\text{mJS}}, \theta)}{R(\hat{\theta}_n^{\text{MLE}}, \theta)}. \tag{2.200}$$

If PAD > 1, then postselection has increased the James-Stein advantage (compared to the non-postselected case). In light of equation (2.199), we expect that the behaviour of PAD(n, θ) will depend on how "isotropic" θ is. We quantify this by

$$v(\theta) = \sum (\theta_i - \bar{\theta})^2, \tag{2.201}$$

related to the variance of θ. In figure 2.6, we plot PAD(n, θ) against n for 4 different values of $\theta \in \mathbb{R}^4$, each with a different $v(\theta)$. the James-Stein estimator allows stronger postselection for smaller n, so that PAD > 1. However, as n increases, the $1/t$ factor in sensitivity increases $v(\theta)$ by a factor of $1/t^2$, which degrades the James-Stein advantage. Thus, PAD eventually decreases. These effects are magnified when $v(\theta)$ is smaller.

Our results demonstrate the subtle interplay between postselection and parameter estimation. Postselection gives a potentially unbounded, uniform amplification in estimation accuracy, as demonstrated in figure 2.5. However, it can have a much more complex effect on the advantage of one estimator over another, as demonstrated in figure 2.6. Moreover, this effect depends on the value of the underlying parameter.

2.3.5 Outlook

In this Section we have further explored the application of postselection to quantum metrology. We have provided a range of results, each of which is a valuable contribution within their own right. Despite recent theoretical advancements [2, 3, 47] in quantum filtering, and even a theoretical demonstration [43], there are still many open problems.

Firstly, we need a better understanding of concrete algorithms for filtering. The scheme in Section 2.3.4 could be further optimised, with a systematic study of when it is best to reduce the value of t, and by how much it should be reduced. Additionally, it is important to consider explicit algorithms for other common parameterised states, such as the phase-encoded state (example 2.4). More generally, it would be desirable to have an iterative procedure that works for any parameterised state ρ, with guarantees on performance of the estimation strategy.

Secondly, filtering of mixed states is very poorly understood. It remains an open problem to maximise $\mathcal{R}_{\rho,K}$ over choices of filter K when ρ is mixed. The example in Section 2.3.2 shows that existing results for pure states do not generalise easily to the mixed case. Some sufficient conditions for lossless compression have been found in Ref. [57], but they are not known to be necessary. Mixed states are particularly important for accounting for noise, which is always present in physical systems. Filtering with depolarising noise and/or photon loss has been explored in Ref. [2], but the case of depolarising noise is not fully solved. Section 2.3.2 shows that this problem is complex, even in dimension three. Moreover, other noise models (such as individual qubit noise) have not yet been studied.

Chapter 3

Probably Approximately Correct Learning

3.1 Background

3.1.1 Classical PAC Learning

In recent years, machine learning has shown remarkable success. Applications have been found in text generation [58], image processing [59], cybersecurity [60], healthcare [61] and biochemistry [62]; to say this list is non-exhaustive is undoubtedly an understatement. Most modern machine learning has its mathematical foundations in *probably approximately correct* (PAC) learning, as introduced by Valiant in 1984 [63]. In this Section, we introduce the classical framework of PAC learning, in a self-contained manner. This is an essential prerequisite to understanding our results, presented in Section 3.2. As in the previous chapter, we give more exposition than is strictly required for our work, in order to give sufficiently motivating context.

In this thesis, we will focus on PAC learning with the simplest type of object, called a classifier, which partitions a set into YES and NO instances. For example, a facial recognition algorithm decides whether or not an image contains a specific face. In analogy to decision problems in complexity theory, these simple YES/NO objects are among the most studied object in theoretical machine learning; much of the theory developed for classifiers can be applied to more general cases [64].

Mathematically, a classifier is a function $f : \mathcal{X} \to \{0, 1\}$, where $\mathcal{X}$ is the set of valid inputs to f, called the input space. We will focus on the case where $\mathcal{X}$ is finite; often the input to f is a length n bistring, i.e. $\mathcal{X} = \{0, 1\}^n$. It is certainly possible to work with infinite $\mathcal{X}$, but one has to deal with technical aspects of measurability [64]. We denote the set of all possible classifiers on $\mathcal{X}$ by $\{0, 1\}^{\mathcal{X}}$. Sometimes, a classifier f is thought of as the subset $f^{-1}(\{1\})$, or its "truth-table" - a list of $f(x)$ for every $x \in \mathcal{X}$.

If there was no known structure in learning problems, machine learning would be extremely difficult - there are $2^{|\mathcal{X}|}$ possible classifiers, so an exponential amount of information is needed to even describe a generic classifier. Yet, there is anthropological evidence that it *is* possible to efficiently learn faces - young children can recognise faces. Our brains filter out most of the irrelevant information in the input image, extracting important details to recognise a face. More abstractly, the rich structure of the problem allows it to be solved efficiently. In general, the structure of a machine learning algorithm

is given by specifying a subset $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$, called a concept class. For example, $\mathcal{C}$ could be the set of functions that do not change if their input is reflected.

A key insight in PAC learning is that it is enough to learn a function f approximately, rather than exactly - if you are correct 99.999% of the time, that is usually good enough in practice. This is modelled by some probability distribution $\mathcal{D}$ on $\mathcal{X}$, representing the distribution of inputs to the classifier. For example, a facial recognition algorithm is much more likely to receive an image corresponding to a photograph than pixelated noise. The distribution induces a distance on classifiers[1]

$$d : \{0,1\}^{\mathcal{X}} \times \{0,1\}^{\mathcal{X}} \to [0,1],\ d(f,g) = \mathbb{P}_{X\sim\mathcal{D}}[f(X) \neq g(X)]. \tag{3.1}$$

We say that f is an ϵ-approximation to g if $d(f,g) \leq \epsilon$. In the simplest type of learning, known as realisable learning, there is some unknown classifier $c \in \mathcal{C}$. An algorithm succeeds if its output g satisfies $d(c,g) \leq \epsilon$, i.e. the output is a good approximation to the true classifier.

A machine learning algorithm receives data (often called the training phase), and then outputs a classifier g. If we insist $g \in \mathcal{C}$, the learning algorithm is called proper, otherwise (if g may be in $\{0,1\}^{\mathcal{X}} \setminus \mathcal{C}$) the algorithm is called improper. In this Section, we only consider one type of access to data, known as random labelled examples. See Section 3.2.1 for a discussion of various different access models. In realisable learning, a random labelled example is a tuple $(X, c(X)) \in \mathcal{X} \times \{0,1\}$

Suppose a learning algorithm $\mathcal{A}$ receives a number of random labelled examples in its training phase. Then, with respect to the randomness of the training data (and any other randomness in the learning algorithm), the output g of $\mathcal{A}$ is itself random. In analogy to allowing a small error in the output g, we also allow for the learning algorithm to fail with some small probability δ. To be precise, $\mathcal{A}$ is said to be an (ϵ, δ) learner if, with probability at least $1-\delta$, it outputs g satisfying $d(g,c) \leq \epsilon$, for all possible inputs (i.e. unknown classifiers c and distributions $\mathcal{D}$).

In practice, it is often difficult to pin down a simple concept class $\mathcal{C}$ that the true classifier belongs to, i.e. the structure of the problem is hard to describe. Instead, we specify some $\mathcal{C}$ (e.g. the set of possible configurations of a neural network), without the promise that an unknown classifier f is in $\mathcal{C}$. We define the optimal error of $\mathcal{C}$ with respect to an unknown classifier f by

$$\mathrm{OPT}_{\mathcal{C}}(f) = \min_{g\in\mathcal{C}} d(f,g). \tag{3.2}$$

In this type of learning, known as functional agnostic learning, instead of outputting an ϵ-approximation to f, we wish to output a classifier g whose error is not much worse than the best case using $\mathcal{C}$. That is, an algorithm succeeds if $d(f,g) \leq \mathrm{OPT}_{\mathcal{C}}(f) + \epsilon$. One hopes that by choosing a suitably "expressive" concept class $\mathcal{C}$, $\mathrm{OPT}_{\mathcal{C}}(f)$, and therefore the error of the output of the learning algorithm, will be small.

In some cases, an ideal classifier f may not even exist. For example, it can be ambiguous whether or

[1] In general, when learning functions of the form $f : \mathcal{X} \to \mathcal{Y}$, we introduce a loss function (as in metrology) $L : \mathcal{Y} \times \mathcal{Y} \to [0,\infty]$ and take $d(f,g) = \mathbb{E}_{X\sim\mathcal{D}}[L(f(X), g(X))]$. If $\mathcal{Y}$ is finite, L is usually taken to be discrete loss (see Section 2.1.1).

Setting	Distribution	Unknown learning object	Succeeds if output g has $d(g,\cdot) \leq$	Random labelled example (aka. sample)
Realisable	$\mathcal{D}$ on $\mathcal{X}$	$c \in \mathcal{C}$	ϵ	$(X, c(X))$, $X \sim \mathcal{D}$
Functional Agnostic	$\mathcal{D}$ on $\mathcal{X}$	$f \in \{0,1\}^{\mathcal{X}}$	$\mathrm{OPT}_{\mathcal{C}}(f) + \epsilon$	$(X, f(X))$, $X \sim \mathcal{D}$
Distributional Agnostic	$\mathcal{D}'$ on $\mathcal{X} \times \{0,1\}$	$\mathcal{D}'$	$\mathrm{OPT}_{\mathcal{C}}(\mathcal{D}') + \epsilon$	$(X, Y) \sim \mathcal{D}'$

Table 3.1: Summary of different settings for PAC learning.

not an image contains a face, or data can be noisy. This is modelled by a joint distribution $\mathcal{D}'$ over $\mathcal{X} \times \{0,1\}$. The conditional probability $\mathcal{D}'(x,y)/\mathcal{D}'(x)$ represents the likelihood that the input x is classified as y. We extend our notion of distance to include distributions:

$$d : \{0,1\}^{\mathcal{X}} \times \Delta(\mathcal{X} \times \{0,1\}),\ d(f, \mathcal{D}') = \mathbb{P}_{(X,Y)\sim\mathcal{D}'}[f(X) \neq Y]. \tag{3.3}$$

Again, we define the optimal error of $\mathcal{C}$ with respect to a distribution $\mathcal{D}'$ as

$$\mathrm{OPT}_{\mathcal{C}}(\mathcal{D}') = \min_{g \in \mathcal{C}} d(g, \mathcal{D}'). \tag{3.4}$$

This type of learning is known as distributional agnostic learning. In this paradigm, a random labelled example is a tuple (X, B), drawn from the distribution $\mathcal{D}'$. A learning algorithm succeeds if its output g satisfies $d(g, \mathcal{D}') \leq \mathrm{OPT}_{\mathcal{C}}(\mathcal{D}') + \epsilon$. Note that functional agnostic learning is a special case of distributional agnostic learning, and that realisable learning is a special case of functional agnostic learning.

The different learning scenarios discussed above are summarised in table 3.1.

The sample [resp. time] complexity of a learning algorithm $\mathcal{A}$ is the number of samples T used by [resp. runtime of] $\mathcal{A}$. The sample [resp. time] complexity of a learning task is the minimum sample [resp. time] complexity of any (ϵ, δ)-learner. Time complexity is more practically relevant than sample complexity, but harder to bound; one can give upper bounds for specific concept classes $\mathcal{C}$, but it is difficult to give generic upper bounds. Moreover, it is hard to give explicit, non-trivial, lower bounds. Sample complexity certainly lower bounds time complexity, but it can be much smaller:

Lemma 3.1: There exist a family of concept classes $\mathcal{C}_n$ for which the sample complexity of realisable learning grows at most polynomially (in $1/\epsilon$, $1/\delta$ and n), but if the time complexity of realisable learning were to also grow at most polynomially, then $\mathsf{RP} = \mathsf{NP}$.

Proof: For a proof of the $\mathsf{RP} = \mathsf{NP}$ claim see Ref. [65], chapter 1 theorem 1.3. The sample complexity growing polynomially is a consequence of Lemma 3.2 and theorem 3.1 below. □

Here, RP is the set of languages L for which there exists a polynomial time randomised algorithm A such that (i) if $x \notin L$, then $A(x) = 0$ with certainty, (ii) if $x \in L$ then $A(x) = 1$ with probability at least $1/2$. Note that RP is contained in $\mathsf{BPP} \cap \mathsf{NP}$, where the random bits consumed by A act as a

witness.

In this thesis, we will focus on sample complexity, which we shall call complexity. The complexity of PAC learning a concept class $\mathcal{C}$ depends on three things: ϵ, δ and some measure of structure within the class $\mathcal{C}$. For sample complexity, the correct measure of complexity of $\mathcal{C}$ is known to be its Valiant-Chapernikis (VC) dimension [66]. For a subset $Y \subseteq X$, and a concept $c \in \mathcal{C}$, let $c|_Y : Y \to \{0,1\}$ $y \mapsto c(y)$ denote the restriction of c to the subset Y. Furthermore, we define $\mathcal{C}|_Y := \{c|_Y : c \in \mathcal{C}\}$ as the restriction of the concept class to Y. We say that $\mathcal{C}$ shatters Y if $\mathcal{C}|_Y = \{0,1\}^Y$, i.e. if all possible labellings of Y appear in concepts in $\mathcal{C}$. The VC dimension of $\mathcal{C}$, $\mathrm{VC}(\mathcal{C})$, is the maximum size of a shattered subset

$$\mathrm{VC}(\mathcal{C}) = \max\{|Y| : Y \subseteq \mathcal{X} \text{ is shattered by } \mathcal{C}\}. \tag{3.5}$$

On shattered sets, $\mathcal{C}$ has no exploitable structure; even if one learns the value of c on every element of Y bar one, there is no way to infer the value of c on the remaining element of Y. It is often difficult to exactly calculate $\mathrm{VC}(\mathcal{C})$, but it can be conveniently bounded in terms of $|\mathcal{C}|$:

Lemma 3.2:

$$\mathrm{VC}(\mathcal{C}) \leq \log|\mathcal{C}| \leq \log\left(\sum_{i=0}^{\mathrm{VC}(\mathcal{C})} \binom{|\mathcal{X}|}{i}\right) \leq \mathrm{VC}(\mathcal{C}) \log|\chi|. \tag{3.6}$$

Proof: For $Y \subseteq \mathcal{X}$, $|\{0,1\}^Y| = 2^{|Y|}$ and thus if $\mathcal{C}$ shatters Y, $2^{|Y|} \leq |\mathcal{C}|$. The lower bound follows by taking Y a shattered set of maxmimal size. The upper bound on $\log|\mathcal{C}|$ is known as the Sauer-Shelah Lemma [67]. The last inequality follows from a simple counting argument. □

The importance of the VC dimension is shown by the following theorem, sometimes called the "Fundamental Theorem of PAC learning" [68].

Theorem 3.1: Suppose $\mathcal{C}$ is a concept class with $\mathrm{VC}(\mathcal{C}) = d$. Let $\epsilon, \delta > 0$. Then the sample complexity for realisable PAC learning of $\mathcal{C}$, denoted $T_C(\epsilon, \delta, \mathcal{C})$, scales as

$$T_C = \Theta\left[\frac{1}{\epsilon}\left(d + \log\left[\frac{1}{\delta}\right]\right)\right]. \tag{3.7}$$

Moreover, the sample complexity for distributional agnostic PAC learning of $\mathcal{C}$, denoted T_C^{dag}, scales as

$$T_C^{\mathrm{dag}} = \Theta\left[\frac{1}{\epsilon^2}\left(d + \log\left[\frac{1}{\delta}\right]\right)\right]. \tag{3.8}$$

The proof of the lower bound on T_C is from Ref. [69], which also gave an upper bound that was only worse by a factor of $\log\epsilon$. The matching upper bound on T_C was given in Ref. [70]. The lower bound on T_C^{dag} was given in Ref. [71], the upper bound was given in Ref. [72]. Functional agnostic learning is more complex: some classes require $\Omega(d/\epsilon^2)$ samples, whilst others require only $O[d/\epsilon(\log[d/\epsilon] + \log[1/\delta])]$, see Ref. [73] for a discussion. For the remainder of this thesis, we will only focus on realisable learning. It is perhaps the least realistic access model, but provides a simple setting from which results can be generalised. Thus, we say "PAC learn" to mean "realisably PAC learn".

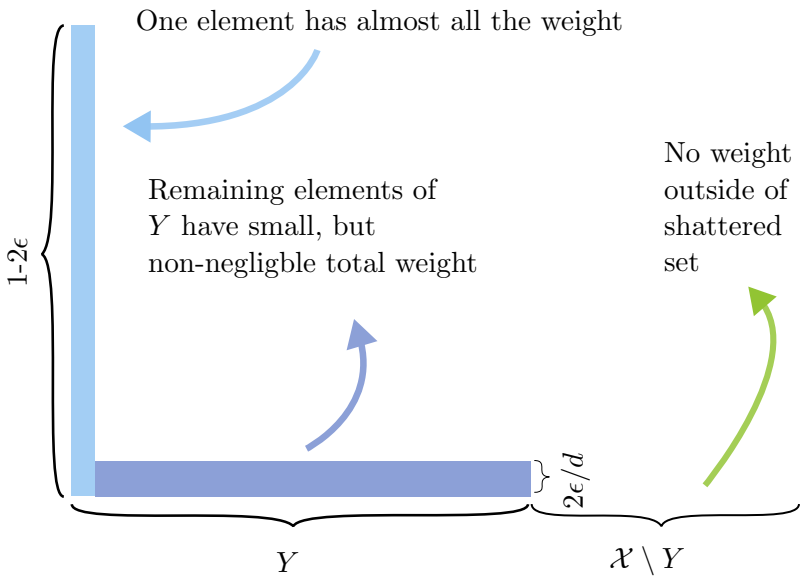


Figure 3.1: The worst case distribution for PAC learning algorithms

We give a brief sketch on the lower bound of T_C, which will be useful in motivating our quantum results. Suppose that $\mathrm{VC}(\mathcal{C}) = d + 1 \geq 2$. Let $Y \subseteq \mathcal{X}$ be shattered by $\mathcal{C}$, with $|Y| = \mathrm{VC}(\mathcal{C})$. Pick some element $y \in Y$, let $Z = Y \setminus \{y\}$, and define a "perturbed" delta function distribution on $\mathcal{X}$:

$$\mathcal{D}_{\mathrm{wc}}(z) = \begin{cases} 1 - 2\epsilon, & z = y, \\ 2\epsilon/d, & z \in Z, \\ 0, & \text{o.w.} \end{cases} \tag{3.9}$$

See figure 3.1 for a pictorial distribution of $\mathcal{D}_{\mathrm{wc}}$. Since $\mathcal{D}_{\mathrm{wc}}$ only has support on Y, a shattered set, there is no structure to exploit. It takes, on average, $\Omega(1/\epsilon)$ samples to learn the value of c on an element of Z, and we must learn at least half of these values to learn an ϵ-approximation to c. Thus, intuitively, $\Omega(d/\epsilon)$ samples should be required. This can be made rigorous: see [69]. Even if one knows the distribution is $\mathcal{D}_{\mathrm{wc}}$ beforehand, one still requires $\Omega(d/\epsilon)$ samples. Since this lower bound matches the known upper bound, $\mathcal{D}_{\mathrm{wc}}$ is a worst case distribution for a learning algorithm.

The upper bound on T_C from from Ref. [69] follows from a very simple algorithm, known as "Occams razor"[2]. The learning algorithm $\mathcal{A}$ takes T samples $S = (x_1, y_1), \ldots (x_T, y_T)$, and outputs any concept in $\mathcal{C}$ that is consistent with the samples. If T is chosen large enough (only a log factor worse than T_C), then the algorithm is a proper (ϵ, δ)-learner. In practice (as shown by Lemma 3.1) it may not be possible to find such a classifier, so one attempts to find a classifier f in $\mathcal{C}$ that agrees with a large fraction of the samples, or equivalently minimises the empirical risk:

$$\hat{R}_S(f) = \frac{1}{T} \sum_i \mathbb{1}_{\{f(x_i) \neq y_i\}}(i). \tag{3.10}$$

The distance $d(f, c)$ of f from the true concept c is known as its generalisation error. Empirical risk minimisation, and bounding generalisation error, forms much of the basis of modern machine learning algorithms.

[2]This name is inspired by the original, philosophical definition.

3.1.2 Quantum PAC Learning

In this Section we describe quantum PAC learning, the basis for our machine learning results. Quantum machine learning has received much attention as a potential application of quantum computers, see Ref. [74] for a review. We will briefly review the fundamental theory of quantum PAC learning, and provide a brief discussion of known quantum advantages. Then, we describe a recent result that calls into question generic quantum machine learning advantages. Our main results, presented in the subsequent Section, show how to circumvent this no-go result, and we give the first generic, rigorous theory of quantum advantage in PAC learning.

The quantum generalisation of PAC learning was first formalised by Bshouty and Jackson [75] in 1996. Instead of a sample $(X, c(X))$, one receives a quantum state

$$|\psi_c\rangle = \sum_{x\in\mathcal{X}} \sqrt{\mathcal{D}(x)}\,|x\ c(x)\rangle\,, \tag{3.11}$$

chosen so that measuring $|\psi_c\rangle$ in the computational basis gives a random labelled example. The state $|\psi_c\rangle$ is called a quantum sample; the sample complexity of a quantum learning algorithm is the number of quantum samples it uses. Since one can always measure the quantum state, and then use a classical algorithm, the quantum learning complexity of $\mathcal{C}$ cannot be worse than the classical one.

In the special case of quantum PAC learning under the uniform distribution, it has been shown that one can obtain quantum sample complexity advantages in specific learning tasks [75–77]. These advantages rely on Fourier sampling, in which one applies the Hadamard transform on every qubit followed by a measurement of the resulting state in the computational basis. One observes a bit string s with probability given by its squared Fourier coefficient[3] $|\hat{c}_s|^2$ and can thus directly infer properties of the Fourier spectrum of the unknown function. However, such advantages rely on the distributions $\mathcal{D}$ being (approximately) uniform. For other specific concept classes, or closely related machine learning tasks, there are known time complexity advantages to quantum learning [79–81], which can be exponential. However, again, these advantages only apply to very specific concept classes, often pathologically constructed (e.g. based on factorisation). In addition, there are often caveats to when quantum exponential machine learning speedups apply, see Ref. [82] for a discussion.

In fact, for an arbitrary distribution $\mathcal{D}$, the quantum sample complexity has exactly the same asymptotic scaling as the classical learning complexity, ruling out everything but constant factor advantages.

Theorem 3.2: Suppose $\mathcal{C}$ is a concept class with $\mathrm{VC}(\mathcal{C}) = d$. Let $\epsilon, \delta > 0$. Then the sample complexity for quantum realisable PAC learning of $\mathcal{C}$, denoted $T_S(\epsilon, \delta, d)$, scales as

$$T_C = \Theta\left[\frac{1}{\epsilon}\left(d + \log\left(\frac{1}{\delta}\right)\right)\right]. \tag{3.12}$$

Note that the upper bounds on T_C follow from the classical case, the lower bounds are proved in Refs. [83, 84]. The worst case distributions for the quantum learners are exactly the same as those for the classical learners. The main idea used in the proof of the lower bound is to use coding theory to

[3] For a function $f : \{0,1\}^n \to \{0,1\}$ and bitstring $s \in \{0,1\}^n$, its Fourier coefficient $\hat{c}_s$ is given by $1/(2^n)\sum_{x\in\mathbb{Z}_2^n}(-1)^{f(x)+x\cdot s}$. For further detail, see Ref. [78]

reduce quantum PAC learning to quantum state discrimination, and utilise existing bounds on state discrimination. Our main result is a new approach to quantum machine learning that circumvents this lower bound.

3.2 Quantum PAC learning with the source code

3.2.1 Access Model

We consider a generalisation of Bshouty and Jackson's quantum samples, in which ones has access to a unitary Q_c that prepares $|\psi_c\rangle$, as well as the inverse unitary $Q_c^\dagger$. Precisely, we assume that there is a concept-independent known state $|\mathrm{IN}\rangle$ such that

$$Q_c\,|\mathrm{IN}\rangle = |\psi_c\rangle\,. \tag{3.13}$$

We are not given any promise on how the oracle Q_c might act on other states in our Hilbert space. We define the learning complexity of any algorithm as the total number of queries to Q_c or $Q_c^\dagger$. The minimum learning complexity of any (ϵ, δ)-learner is denoted $T_O(\epsilon, \delta, \mathcal{C})$.

This access model (or strongly related access models) has recently received attention in many different areas of quantum information. It has been studied in Ref. [85] in the context of quantum state tomography, in Ref. [86] in the context of quantum channel tomography and in Ref. [87], in the context of estimating the mean of a random variable.

Since Valiant introduced PAC learning, a plethora of access models have been proposed, inspired by different real-world scenarios. We provide a brief discussion of the access model considered in this Section (access to Q_c and $Q_c^\dagger$), and its relationship to three other popular models. We compare to random labelled examples $(X, c(X))$, quantum samples $|\psi_c\rangle$, and membership queries - where a learner can submit an $x \in \mathcal{X}$ and receives $c(x)$ in return.

Random labelled examples are the most commonly considered access model in the machine learning literature. Even if learning complexity is polynomial, there may be no (time) efficient learner [88]. Some classes have time-efficient learners with random labelled examples and membership queries, but not with only random labelled examples [89]. Curiously, we are not aware of a work that gives a learning complexity lower bound on random labelled examples and membership queries. For certain concept classes (such as concepts which label exactly d elements of $\mathcal{X}$ as 1), we believe that the standard lower bound distribution for random labelled examples should give the same lower bound for random labelled examples and membership queries. However, for example, when considering the concept class of all possible functions, membership queries give a learning upper bound of $|\mathcal{X}|$, which may give a significant improvement over the $|\mathcal{X}|/\epsilon$ samples required by random labelled examples alone. Thus in general, membership queries can be much stronger than either random labelled examples or our model; given random labelled examples or quantum samples, it takes on average $1/\mathcal{D}(x)$ queries to find $c(x)$. Given our access model, one can perform amplitude amplification such that only $1/\sqrt{\mathcal{D}(x)}$ samples are required, which may also be large. Indeed, if $\mathcal{D}(x) = 0$, then a membership query of x is impossible using random labelled examples, quantum samples or our model. Thus, membership queries are incomparable with our model.

Quantum samples were introduced as the natural quantum generalisation of random labelled examples. However, in practice it is unclear how to prepare $|\psi_c\rangle$, given that it depends on c and $\mathcal{D}$. Whether or not this is reasonable depends on how the random labelled examples are classically generated - if

they are completely "black-box", then it seems unlikely that quantum samples will be an appropriate resource without very strong quantum data loading subroutines, such as QRAM. If c and $\mathcal{D}$ are "white-box", i.e. there is some circuit producing them, then an appropriate quantisation procedure will lead to Q_c. Given a circuit description of Q_c, one can theoretically perform $Q_c^\dagger$. Thus, in most scenarios where quantum samples are sensible for learning classical data, our model is also reasonable.

A more promising case for quantum samples is when data is inherently quantum. Suppose a quantum process produces $\left|\widetilde{\psi_c}\right\rangle = \sum_x \sqrt{\mathcal{D}(x)} e^{ih(x)} |x, c(x), g(x)\rangle$, where $h(x), g(x)$ are some extraneous functions. In this case, learning c allows us to make physical predictions of the $c(x)$ register given the x register, without knowledge of $g(x)$. This is useful, e.g., in learning far-range behaviour/correlations. By the Stinespring dilation theorem, the quantum process has some unitary representation Q_c, which can be taken as our oracle. Note that our algorithm is insensitive to the addition of the $g(x)$ register and $h(x)$ phases.

3.2.2 Grover Subroutine

An essential subroutine for our quantum advantage is to use calls to Q_c and $Q_c^\dagger$ to run a Grover search [90, 91]. This leads to a quadratic improvement in learning complexity (up to polylogarithmic factors) over classical PAC learning. In this Section, we describe our Grover subroutine.

Our Grover subroutine takes as an input a "good" subset $G \subseteq \{(x, b) : x \in \mathcal{X}, b \in \{0, 1\}\}$, where we wish to find an x such that $(x, c(x)) \in G$. We define a corresponding "good" subspace by

$$\mathcal{G} = \text{span}\{|x\ b\rangle : (x, b) \in G\}. \tag{3.14}$$

In order to implement Grover's search, we need to implement the Grover operator, defined by

$$D = (\mathbb{1} - 2\,|\psi_c\rangle\langle\psi_c|)(\mathbb{1} - 2\Pi_{\mathcal{G}}), \tag{3.15}$$

where $\Pi_{\mathcal{G}}$ is the orthogonal projection map onto $\mathcal{G}$. We show that implementing D requires a constant number of queries.

Proposition 3.1: For any choice of $\mathcal{G}$, one can implement the Grover operator D with one call to Q_c and one to $Q_c^\dagger$.

Proof: Note that $(\mathbb{1} - 2\Pi_{\mathcal{G}})$ is independent of c and, therefore, may be implemented by a (possibly exponentially sized circuit) without any queries. To implement $(\mathbb{1} - 2\,|\psi_c\rangle\langle\psi_c|)$, note that

$$\mathbb{1} - 2\,|\psi_c\rangle\langle\psi_c| = Q_c(\mathbb{1} - 2\,|\text{IN}\rangle\langle\text{IN}|)Q_c^\dagger, \tag{3.16}$$

Note that $(\mathbb{1} - 2\,|\text{IN}\rangle\langle\text{IN}|)$ is independent of c and, therefore, may be implemented by a (possibly exponentially sized circuit) without any queries. □

One can uniquely decompose

$$|\psi_c\rangle = \sin(\theta)\,|g\rangle + \cos(\theta)\,|b\rangle\,, \tag{3.17}$$

where $|g\rangle, |b\rangle$ are orthonormal, $\theta \in [0, \pi/2]$, $|g\rangle \in \mathcal{G}$ and $|b\rangle \in \mathcal{G}^{\perp}$. It is well known that [90]

$$D^n |\psi\rangle = \sin((2n+1)\theta) |g\rangle + \cos((2n+1)\theta) |b\rangle . \tag{3.18}$$

Thus, if we knew θ exactly, we could apply D^n such that $\sin((2n+1)\theta) \approx 1$. However, since θ depends on $\mathcal{D}$, which is unknown, this is impossible. Instead, we use a well-established [92] version of Grover's search for an unknown number of items. Our exact subroutine is given below; algorithm 3.1.

Algorithm 3.1:

Input: $G \subseteq \{(x, b) : x \in \mathcal{X}, b \in \{0, 1\}\}$ a good subspace, $\epsilon > 0$ a tolerance

Output: labelled example $(x, c(x))$. Succeeds if $(x, c(x)) \in G$

1. Produce $|\psi_c\rangle = Q_c |\text{IN}\rangle$
2. Pick N from $0, 1 \ldots, \lceil 2/\sqrt{\epsilon} \rceil - 1$ uniformly at random
3. Apply D, the Grover operator, N times to $|\psi_c\rangle$
4. Measure the resulting state in the computational basis

The properties of our algorithm are summarised in the following Lemma

Lemma 3.3: Let $G \subseteq \{(x, b) : x \in \mathcal{X}, b \in \{0, 1\}\}$ be a good subset, $\epsilon > 0$ be a fixed tolerance. Suppose that we run Algorithm 3.1 with these inputs, then

(i) In the worst case, the algorithm makes $O(1/\sqrt{\epsilon})$ oracle (or inverse oracle) calls.

(ii) If $\mathbb{P}_{X \sim \mathcal{D}} [(X, c(X)) \in G] \geq \epsilon$ then the algorithm succeeds, i.e., returns $(x, c(x)) \in G$, with probability at least $p = 0.09$.

(iii) Conditional on succeeding, the output of the algorithm $(X, c(X))$ is distributed according to

$$\mathbb{P} [(X, c(X)) | \text{algorithm succeeds}] = \frac{\mathbb{P}_{X \sim \mathcal{D}} [X]}{\mathbb{P}_{X \sim \mathcal{D}} [(X, c(X)) \in G]}. \tag{3.19}$$

Proof: Part (i): From the definition of the algorithm and Lemma 3.1, the worst case number of oracle calls is $1 + 2(\lceil 2/\sqrt{\epsilon} \rceil - 1) = O(1/\sqrt{\epsilon})$.

Part (ii): Let $M = \lceil 2/\sqrt{\epsilon} \rceil$, let θ be as in equation (3.17) and let ρ^{ps} be the probability that the algorithm succeeds. Note that $\mathbb{P}_{X \sim \mathcal{D}} [(X, c(X)) \in G] \geq \epsilon \Leftrightarrow \sin(\theta) \geq \sqrt{\epsilon}$. We use Lemma 2 (Section 6) from [92], which claims

$$\rho^{\text{ps}} = \frac{1}{2} - \frac{1}{4M} \frac{\sin(4M\theta)}{\sin(2\theta)}. \tag{3.20}$$

For $\sin(\theta) \in [\sqrt{\epsilon}, 1/\sqrt{2}]$:

$$M \geq \frac{2}{\sin(\theta)}, \tag{3.21}$$

$$\geq \frac{1}{\sin(2\theta)}, \tag{3.22}$$

and thus

$$\rho^{\text{ps}} \geq \frac{1}{2} - \frac{1}{4} = \frac{1}{4} > 0.09. \tag{3.23}$$

Note that for $\theta \in [\pi/4, \pi/2]$,

$$\sin(2\theta) \geq \frac{\pi/2 - \theta}{\pi/4}, \tag{3.24}$$

Thus, for $\theta \in [\pi/4, (1/2 - 1/4M)\pi]$, we have that

$$p_s(\theta) \geq \frac{1}{2} - \frac{1}{4M} \cdot \frac{4/\pi}{\pi/2 - (1/2 - 1/4M)\pi}, \tag{3.25}$$

$$= \frac{1}{2} - \frac{4}{\pi^2} > 0.09. \tag{3.26}$$

Finally, for $\theta \in [(1/2 - 1/4M)\pi, \pi/2]$, note that $\sin(2\theta) \geq 0$ and $\sin(4M\theta) \leq 0$ so that $p_s(\theta) \geq 1/2 > 0.09$.

Part (*iii*). This follows from the form of $D^n |\psi_c\rangle$; the relative magnitude of the amplitudes in $|g\rangle$ is unchanged by the Grover operator D. □

We discuss how to combine the Grover subroutine with the algorithm of Section 3.2.3 to achieve a quantum learning complexity advantage in Section 3.2.4.

3.2.3 Learning with Imperfect Equivalence Queries

Equivalence queries are an alternative learning model for PAC learning. It was recently shown [93] that PAC learning with equivalence queries gives an exponential advantage over learning with labelled examples. In this Section, we show how to use imperfect equivalence queries to PAC learn a concept class, which is a core component of our quantum algorihtm.

An (ideal) equivalence query consists of submitting a candidate hypothesis h for an underlying true concept c. If $h = c$ then we are told YES. Otherwise, we receive a labelled example $(x, c(x))$ where $c(x) \neq h(x)$ at random, according to the distribution $\mathbb{P}(y) = \mathcal{D}(y)/\mathcal{D}(\{x : c(x) \neq h(x)\})$. Such a labelled example where $h(x) \neq c(x)$ is called a counterexample. Equivalence queries are a very strong learning model, and perhaps unrealistic. Thus, we assume we can only implement them probabilistically.

An imperfect equivalence query consists of submitting a candidate hypothesis h for the underlying concept c. In return we receive some labelled example $(x, c(x))$ with the following promises

- There is some known probability p, independent of the concept c, and the values of ϵ and δ. Moreover, If $d(h, c) \geq \epsilon$, then the probability of receiving a counterexample (i.e. that $c(x) \neq h(x)$) is at least p

- The distribution of $(X, c(X))$ *conditional on being a counterexample* is the same as an ideal equivalence query. More precisely, $\mathbb{P}[\text{receive } (y, c(y)) \mid c(y) \neq h(y)] = \mathcal{D}(y)/\mathcal{D}(\{x : c(x) \neq h(x)\})$.

Note that we can tell whether our imperfect equivalence query failed or not - we can look at the result $(x, c(x))$ and check whether $h(x) = c(x)$. If they are equal, the equivalence query failed. Otherwise, it succeeded. Classically, we can implement an imperfect equivalence query using $1/\epsilon$ random labelled examples - we just sample $1/\epsilon$ times and see whether $c(x) \neq h(x)$ for any of our samples. On a quantum computer we can do this in $1/\sqrt{\epsilon}$ time using Grover's algorithm, as described in Section 3.2.2 in Lemma 3.3.

We need one additional tool from classical learning theory to run our algorithm, the weighted majority vote. Suppose we have a set of classifiers $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ and a distribution ρ on $\mathcal{H}$. Then the weighted majority vote [94], $\text{WMV}_{\mathcal{H},\rho} \in \{0,1\}^{\mathcal{X}}$ is defined such that it maximises

$$\mathbb{P}_{h\sim\rho}\left[\text{WMV}_{\mathcal{H},\rho}(x) = h(x)\right], \tag{3.27}$$

for every x (ties can be broken arbitrarily).

Suppose we have a classical algorithm $\mathcal{A}$ that uses $T_E(\epsilon, \delta, d)$ (ideal) equivalence queries to PAC learn a concept class $\mathcal{C}$. We show how to use $O(T_E + \log(1/\delta))$ imperfect equivalence queries to PAC learn the same concept class. The full detail of the algorithm is given below in algorithm 3.2. It works by running $\mathcal{A}$, replacing every equivalence query with repeated imperfect equivalence queries until one succeeds. We terminate if the learning algorithm $\mathcal{A}$ terminates or if we make a total of $R(T_E, \delta)$ imperfect equivalence queries.

Algorithm 3.2:

Input: $\delta > 0, \epsilon > 0$ (the usual PAC parameters) and $\mathcal{A}$ a classical equivalence query learning algorithm with worst case query complexity $T_E > 0$

Output: Hypothesis $h \in \{0,1\}^{\mathcal{X}}$

1. Set the maximum imperfect equivalence query budget as $R = 6T_E/p + (3/2p^2)\log(1/\delta)$. If R total imperfect equivalence queries have ever been made, go to step 3
2. Run $\mathcal{A}$, whenever it requires an equivalence query to a hypothesis h, repeatedly make imperfect equivalence queries until one succeeds (note, we can check whether the imperfect equivalence query succeeded, by comparing $c(x)$ and $h(x)$). If $\mathcal{A}$ terminates, output the output of $\mathcal{A}$
3. Let $\mathcal{H} = \{h_1, \ldots, h_k\}$ be the set of hypothesis we ran imperfect equivalence queries on (so that $k \leq T_E$). Suppose we spent n_i imperfect equivalence queries on h_i (so that $\sum n_i = R$). Let $\rho(h_i) = n_i/R$ and output $h = \text{WMV}_{\mathcal{H},\rho}$

We give some rough intuition for why the algorithm works before proving so. If $\mathcal{A}$ terminates, then with high probability, it outputs an approximately correct hypothesis. If we pick R large enough,

then with high probability T_E ideal queries to hypotheses h_i with $d(h_i, c) \geq \epsilon$ would all succeed in $< R/3$ imperfect equivalence queries. Thus, if the algorithm $\mathcal{A}$ does not terminate and we make R total imperfect equivalence queries, with high probability, we spent $> 2/3$ of our imperfect equivalence queries on hypotheses h_i with $d(h_i, c) < \epsilon$. Hence, if we take the weighted majority vote of all of the hypotheses we queried, weighted by the number of imperfect equivalence queries spent on each hypothesis, most of the vote will be decided by hypotheses that are close to the concept c. Thus, the weighted majority vote will also be close to c.

The full proof of why algorithm 3.2 works is given as two propositions. They require some new terminology. A transcript of a run of algorithm 3.2 is given by the list of hypotheses $\mathcal{H} = \{h_i\}$ that the algorithm queried along with a corresponding collection of natural numbers $n_i > 0$, where n_i is the number of imperfect equivalence queries spent on h_i. The corresponding time-spent distribution ρ is the probability distribution on $\mathcal{H}$ given by $\rho(h_i) = n_i / \sum_i n_i$. Finally, $F = \{i : d(h_i, c) \geq \epsilon\}$ is called the "feasible" set, where our imperfect equivalence query succeeds with probability at least p. Correspondingly $I = \{i : d(h_i, c) < \epsilon\}$ is the "infeasible" set, where there is no promise on the probability of success.

Firstly, we show that with high probability that a bounded number of queries is spent on the feasible set

Proposition 3.2: With probability $\geq 1 - \delta$ the total number of imperfect equivalence queries to feasible hypotheses is at most

$$2T_E/p + (1/2p^2)\log(1/\delta). \tag{3.28}$$

Proof: A imperfect equivalence query of a feasible hypothesis has (by definition) a chance $\geq p$ of succeeding, and the individual imperfect equivalence queries are independent. Additionally, there are at most T_E feasible hypotheses to query (since the classical algorithm makes at most T_E total equivalence queries). Thus, the probability that we succeed on all the feasible hypotheses using at most m imperfect queries feasible hypotheses is lower bounded by the probability of getting at least T_E successes from a binomial distribution $B(m, p)$. Thus, the chance of failure is lower bounded by the chance of fewer than T_E successes from $B(m, p)$.

Let $X \sim B(m, p)$. Applying Hoeffding's inequality [95], for $m \geq T_E/p$ we see that

$$\mathbb{P}\left[X < t\right] \leq e^{-2m(p - T_E/m)^2}. \tag{3.29}$$

Thus it is sufficient for

$$2m\left(p - \frac{T_E}{m}\right)^2 \geq \log(1/\delta). \tag{3.30}$$

In turn, it is sufficient that

$$2mp^2 - 4pT_E \geq \log(1/\delta), \tag{3.31}$$

whence we deduce our bound.

□

Next we prove that if we make enough imperfect equivalence queries on infeasible hypotheses, the weighted majority vote of the transcript must be close to the underlying concept c

Proposition 3.3: Suppose we make R total imperfect equivalence queries, and spend at least $2R/3$ imperfect equivalence queries on infeasible hypotheses. Then the weighted majority vote M of the transcript with the time-spent distribution has $d(M, c) < 4\epsilon$.

Proof: Fix the transcript $h_1, \dots h_k$. Let ρ be the time-spent distribution and let ρ' be the time-spent distribution conditioned on the infeasible set. That is, for $i \in I$, $\rho'(h_i) = \rho(h_i)/\rho(I)$. Similarly let $\tilde{\rho}$ the the time-spent distribution conditioned on the feasible set. We first show that if the infeasible set overwhelmingly votes for a bit y, then the whole transcript must also vote for that y. To be precise, if $\mathbb{P}_{h\sim\rho'}[h(x) = y] > 3/4$, then

$$\mathbb{P}_{h\sim\rho}[h(x) = y] = \mathbb{P}_{h\sim\rho'}[h(x) = y]\,\mathbb{P}_{h\sim\rho}[h \in I] + \mathbb{P}_{h\sim\tilde{\rho}}[h(x) = y]\,\mathbb{P}_{h\sim\rho}[h \in F], \tag{3.32}$$

$$> \frac{3}{4} \cdot \frac{2}{3}, \tag{3.33}$$

$$= \frac{1}{2}. \tag{3.34}$$

Let $M = \mathrm{WMV}_{\mathcal{H},\rho}$. By the above, if $\mathbb{P}_{h\sim\rho'}[h(x) = c(x)] > \frac{3}{4}$, then $M(x) = c(x)$. We deduce (inspired by [94]) that

$$\mathbb{P}_{X\sim\mathcal{D}}[M(X) \neq c(X)] \leq \mathbb{P}_{X\sim\mathcal{D}}\left[\mathbb{P}_{h\sim\rho'}[h(X) \neq c(X)] \geq \frac{1}{4}\right], \tag{3.35}$$

$$\text{Markov's inequality, } \leq 4\mathbb{E}_{X\sim\mathcal{D}}\mathbb{E}_{h\sim\rho'}[\mathbb{1}_{\{h(X)\neq c(X)\}}], \tag{3.36}$$

$$= 4\mathbb{E}_{h\sim\rho'}[d(h, c)], \tag{3.37}$$

$$\text{definition of infeasible set, } < 4\epsilon. \tag{3.38}$$

□

We can now prove the performance of our algorithm

Lemma 3.4: Let the maximum number of imperfect equivalence queries of algorithm 3.2 be

$$R(T_E(\epsilon, \delta, d), \delta) = 6T_E(\epsilon, \delta, d)/p + (3/2p^2)\log(1/\delta), \tag{3.39}$$

then algorithm 3.2 produces a hypothesis h with $d(h, c) \leq 4\epsilon$ with probability at least $1 - 2\delta$.

Proof: By proposition 3.2, with probability $\geq 1 - \delta$ we spend at most $R/3$ imperfect equivalence queries on feasible hypotheses - suppose this happens. If we succeed in an equivalence query for every hypothesis required by $\mathcal{A}$ then with probability at least $1 - \delta$, $\mathcal{A}$ outputs a hypothesis h with $d(h, c) \leq \epsilon$. Otherwise, we spend at least $2R/3$ imperfect equivalence queries on infeasible hypotheses (as we assumed the feasible ones took at most $R/3$ imperfect equivalence queries) and then by Lemma 3.3 the weighted majority vote $\mathrm{WMV}_{\mathcal{H},\rho}$ has $d(\mathrm{WMV}_{\mathcal{H},\rho}, c) < 4\epsilon$. Thus algorithm 3.2 outputs a 4ϵ-approximately correct hypothesis with probability at least $(1-\delta)^2 \geq 1 - 2\delta$. □

3.2.4 Upper bound on Quantum PAC Learning

Here, we combine the results of Sections 3.2.2 and 3.2.3 to give an upper bound on T_O, the learning complexity of PAC learning with a state preparation oracle Q_c (and its inverse).

Theorem 3.3: Let $\mathcal{C}$ be a concept class with VC dimension d. Then, for every $\epsilon, \delta > 0$, there exists a (ϵ, δ)-quantum PAC learner for $\mathcal{C}$ that makes at most

$$O\left(\frac{1}{\sqrt{\epsilon}}\left[d + \log\left(\frac{1}{\delta}\right)\right]\log^9(1/\epsilon)\right), \tag{3.40}$$

calls to an oracle that generates a quantum sample (Q_c) or its inverse ($Q_c^\dagger$).

Proof: Suppose that it takes $E(\epsilon)$ queries to perform an imperfect equivalence query for a hypothesis h. If we have a classical equivalence learning algorithm $\mathcal{A}$ with an equivalence query complexity of $T_E(\epsilon, \delta, d)$, then we can use algorithm 3.2 of Section 3.2.3 to get a quantum PAC learning algorithm with learning complexity

$$E(\epsilon/4)R(T_E(\epsilon/4, \delta/2, d), \delta/2). \tag{3.41}$$

The current best known T_E [93] has a worst-case query complexity of

$$T_E = O\left(\left[d + \log\left(\frac{1}{\delta}\right)\right]\log^9\left(\frac{1}{\epsilon}\right)\right). \tag{3.42}$$

If we use the Grover subroutine (Section 3.2.2 algorithm 3.1) with $G = \{(x, 1 - h(x)) : x \in \mathcal{X}\}$ to implement the imperfect equivalence queries, we find $E(\epsilon) = O(1/\sqrt{\epsilon})$. On substituting these values of T_E and E into the bound from equation (3.41), the result follows □

We note that theorem 3.3 is a square-root improvement (up to polylogarithmic factors) over the classical PAC learning sample complexity of theorem 3.1.

3.2.5 Lower bound on Quantum PAC Learning

In this Section, we prove a lower bound on quantum PAC learning with a state preparation oracle (and its inverse). We show that $\Omega(d/\sqrt{\epsilon})$ oracle calls are necessary. Up to polylogarithmic factors, this shows that our quadratic improvement is optimal.

Suppose we have a concept class $\mathcal{C}$ with VC dimension $d + 1$. Then there is a set Z of size $d + 1$ in $\mathcal{X}$ which is shattered by $\mathcal{C}$. We pick a marked element $x_0 \in Z$ and let $Y = Z \setminus \{x_0\}$. We define our distribution $\mathcal{D}$ as a perturbed delta-function, the standard distribution used to prove lower bounds in learning:

$$\mathcal{D}(x) = \begin{cases} 0, & \text{if } x \notin Z, \\ 1 - 4\epsilon, & \text{if } x = x_0, \\ 4\epsilon/d, & \text{if } x \in Y. \end{cases} \tag{3.43}$$

We also restrict our concept class to $\widetilde{\mathcal{C}} = \{c \in \mathcal{C} : c(x_0) = 0\}$. If our PAC algorithm works on $\mathcal{C}$, it will certainly work on $\widetilde{\mathcal{C}}$. Since our distribution is restricted to Z we need only identify the behaviour

of our concept on Z. Thus, we can index our concepts by bit-strings $u \in \{0,1\}^d$ and index them with elements of Y. To be precise, we identify a concept $c \in \widetilde{\mathcal{C}}$ with a bit-string $u \in \{0,1\}^d$, where $u_y = c(y)$.

For a given bit-string $u \in \{0,1\}^d$, the state preparation oracle acts as

$$Q_u \left|\mathrm{IN}\right\rangle = \sqrt{1-4\epsilon}\left|x_0\ 0\right\rangle + \sqrt{\frac{4\epsilon}{d}} \sum_{x \in Y} \left|x\ u_x\right\rangle . \tag{3.44}$$

Our main approach is to reduce to the following fact from Lemma 51 in [85].

Lemma 3.5: Let $u \in \{0,1\}^d$ be a bit string, and let O_u be a weak phase-kickback oracle, that is

$$O_u \left|x\right\rangle = e^{2i\eta u_x} \left|x\right\rangle . \tag{3.45}$$

Then recovering more than $3/4$ of the bits of u with high probability requires at least $\Omega(d/\eta)$ calls to O_u, its inverse or controlled versions of these.

Proof: See [85] □

We will use calls to controlled versions of O_u (denoted c-O_u) to implement the PAC state generation oracle Q_u. We fix $\eta \in [0, \pi/2]$ such that $\sin(\eta) = \sqrt{4\epsilon}$.

Proposition 3.4: One can implement Q_u using one call to c-O_u, one to c-$O_u^\dagger$ and two qubit-ancillae.

Proof: First, it is convenient to shift the phase to have a $\pm$ symmetry. Define a constant phase gate as

$$P_\alpha \left|x\right\rangle = e^{i\alpha} \left|x\right\rangle . \tag{3.46}$$

Then let

$$\widetilde{O}_u = P_\eta O_u^\dagger, \tag{3.47}$$

so that

$$\widetilde{O}_u \left|x\right\rangle = e^{i\eta \hat{u}_x} \left|x\right\rangle , \tag{3.48}$$

where

$$\hat{u}_x = (-1)^{u_x}. \tag{3.49}$$

We start by generating a uniform superposition of indices with the two-qubit ancillae in the $|+\rangle$ state:

$$\frac{1}{2\sqrt{d}} \sum_{x \in Y} \left|x\right\rangle [\left|00\right\rangle + \left|01\right\rangle + \left|10\right\rangle + \left|11\right\rangle]. \tag{3.50}$$

We next apply 4 controlled gates - either c-P_η, c-$P_{-\eta}$ c-$\widetilde{O}_u$ and c-$\widetilde{O}_u^\dagger$, such that each term in the

superposition in equation (3.50) picks up a different phase:

$$\mapsto \frac{1}{2\sqrt{d}} \sum_{x \in Y} |x\rangle \left[e^{i\eta} |00\rangle + e^{-i\eta} |01\rangle + e^{i\eta \hat{u}_x} |10\rangle + e^{-i\eta \hat{u}_x} |11\rangle \right]. \tag{3.51}$$

Note that this requires two calls to singly controlled versions of the oracle - we can implement a double-controlled version by using a CCNOT (Toffoli) gate followed by a controlled oracle. Next, we apply a Hadamard gate to the second qubit register

$$\mapsto \frac{1}{\sqrt{2d}} \sum_{x \in Y} |x\rangle \left[|0\rangle \left(\cos(\eta) |0\rangle + i \sin(\eta) |1\rangle \right) + |1\rangle \left(\cos(\eta \hat{u}_x) |0\rangle + i \sin(\eta \hat{u}_x) |1\rangle \right) \right]. \tag{3.52}$$

We then apply $S^\dagger$ to the second qubit register (to remove the factors of i). We also use the even/odd ness of cos/sin to regroup the terms:

$$\mapsto \frac{1}{\sqrt{2d}} \sum_{x \in Y} |x\rangle \left[\cos(\eta)(|0\rangle + |1\rangle) |0\rangle + \sin(\eta)(|0\rangle + \hat{u}_x |1\rangle) |1\rangle \right]. \tag{3.53}$$

We then apply a Hadamard gate to the first qubit register:

$$\mapsto \cos(\eta) \left(\frac{1}{\sqrt{d}} \sum_{x \in Y} |x\rangle \right) |00\rangle + \sin(\eta) \left(\frac{1}{\sqrt{d}} \sum_{x \in Y} |x \; u_x\rangle \right) |1\rangle. \tag{3.54}$$

Conditional on the final qubit being in the state $|0\rangle$, we apply a unitary to the first register that maps the uniform superposition over Y into the state $|x_0\rangle$:

$$\mapsto \cos(\eta) |x_0 \; 0 \; 0\rangle + \sin(\eta) \left(\frac{1}{\sqrt{d}} \sum_{x \in Y} |x \; u_x\rangle \right) |1\rangle. \tag{3.55}$$

Finally, conditional on the first register not being in the state $|x_0\rangle$, we apply an X gate to the second qubit register, followed by an H gate on the second qubit register:

$$\mapsto \left[\cos(\eta) |x_0 \; 0\rangle + \sin(\eta) \left(\frac{1}{\sqrt{d}} \sum_{x \in Y} |x \; u_x\rangle \right) \right] |+\rangle. \tag{3.56}$$

But by the definition of η, we see that this is exactly equal to the action of the PAC oracle:

$$(Q_u |\mathrm{IN}\rangle) |+\rangle. \tag{3.57}$$

□

We deduce our bound

Theorem 3.4: $T_O = \Omega\left(\frac{d}{\sqrt{\epsilon}}\right)$.

Proof: We can replace every call to Q_u (or its inverse) in our PAC algorithm with the unitary process described in proposition 3.4, which requires a constant number of calls to (a controlled) O_u (or its inverse). If the PAC algorithm outputs a correct hypothesis, then by construction of our distribution, it must agree on at least 3/4 of the bits of u. Thus, the algorithm replaced with

calls to O_u (and its inverse) satisfies the conditions of Lemma 3.5, and thus it must use at least $\Omega(d/\eta)$ calls to O_u. Hence, we reach a lower bound of

$$T_O = \Omega\left(\frac{d}{\arcsin\sqrt{4\epsilon}}\right) = \Omega\left(\frac{d}{\sqrt{\epsilon}}\right). \tag{3.58}$$

□

Note that our lower bound matches our upper bound (theorem 3.3), up to polylogarithmic factors.

3.2.6 Application to Learning k-juntas

A k-junta is a function $f : \{0,1\}^n \to 0,1$ that only depends on a subset of k bits. Letting $\mathcal{X} = \{0,1\}^n$, we can consider the concept class $\mathcal{C} = \{f \in \{0,1\}^{\mathcal{X}} : f \text{ is a } k \text{ junta}\}$. k-juntas are a particularly important concept class in PAC learning - they represent the fact that many systems depend on lots of parameters, but only a few are important. We bound

$$|\mathcal{C}| \leq \binom{n}{k} 2^{(2^k)}, \tag{3.59}$$

since there are $\binom{n}{k}$ ways to choose the k bits determining the junta, and then $2^{(2^k)}$ choices for the underlying function. Using Lemma 3.2 we deduce that

$$d \leq \log\left[\binom{n}{k}\right] + 2^k \leq k\log(en/k) + 2^k. \tag{3.60}$$

Thus, our learning algorithm can PAC learn a k-junta with

$$O\left(\frac{1}{\sqrt{\epsilon}}\left[k\log\left(\frac{n}{k}\right) + 2^k + \log\left(\frac{1}{\delta}\right)\right]\log^9(1/\epsilon)\right), \tag{3.61}$$

oracle calls. This has a worse scaling in n than algorithms presented in [77], but has a better scaling in ϵ and works for *any* underlying distribution, whereas previous work has focused on the uniform distribution.

3.3 Outlook

Our work leaves several interesting avenues for further research. Firstly, one could attempt to tighten the upper bound (theorem 3.3) to remove polylogarithmic factors and prove a tight matching lower bound. The removal of a $\log(1/\epsilon)$ factor in the query complexity for classical PAC learning took 27 years [8, 69]; we hope that the quantum case will be simpler. Moreover, in order to achieve $1/\sqrt{\epsilon}$ scaling with our method, one would require the optimal classical equivalence query learning complexity to have no ϵ dependence. Thus, a different approach is likely to be required.

Secondly, it is interesting to consider the power of quantum learning algorithms with access to the oracle Q_c, but not its inverse $Q_c^\dagger$. The inverse oracle seems necessary for Grover's search, and therefore it is unclear if a quantum advantage is possible. The lack of such an advantage would have interesting implications for understanding what makes quantum computing more powerful than classical computation. It is additionally interesting to consider if there is any suitable classical analogue for the

inverse oracle $Q_c^\dagger$.

Thirdly, one could attempt to find analogous sample complexity improvements in agnostic learning, in both the functional and distributional cases. There are many existing algorithms for estimating the mean of a random variable that admit quadratic quantum improvements [87, 96]. These algorithms could be applied to the distributional case.

Finally, we briefly consider potential applications to boosting. Boosting is a technique in classical machine learning that combines several "weak" hypotheses into a single "strong" hypothesis (see Ref. [97] for a review). The weak hypotheses are produced by a weak-learner, which is promised to always output a hypothesis h that is slightly better than random guessing: $d(h, c) \leq 1/2 - \gamma$ for some (small) parameter $\gamma \in (0, 1/2]$.

Boosting algorithms run the weak learner many times. Importantly, they change the distribution on $\mathcal{X}$ that they sampling from between calls to the weak learner: more weight is assigned to x's which are incorrectly classified by previous hypotheses. This is most commonly achieved by taking an initial (large) training sample (using $\mathcal{D}$), starting with a uniform distribution on this training sample, an subsequently altering this distribution. The boosting algorithm outputs a hypothesis with small empirical risk which, for a sufficiently large training sample, has small generalisation error. The precise way in which the distribution is updated depends on the boosting algorithm in use: common choices include `AdaBoost` [98] and `SmoothBoost` [99].

In quantum boosting, the underlying weak learner is assumed to be a quantum algorithm, which is fed copies of the example state $|\psi_c\rangle$. By changing the amplitudes in $|\psi_c\rangle$, one can run boosting algorithms. This has already been studied in Refs. [100, 101], where one finds a quadratic improvement in VC-dimension over classical algorithms, but slightly worse dependence on γ. There is a natural appeal to use quantum algorithms for boosting, due to the relationship between updating distributions and changing quantum amplitudes. For example, in `Adaboost`, the distribution updates in a simple way: every x for which the previous hypothesis h was wrong has $\mathcal{D}(x)$ increase by the same factor, the remaining x's have $\mathcal{D}(x)$ decrease by a different factor. The factor is chosen in such a way that the probability of the set $\{x \mid h(x) \neq c(x)\}$ (under $\mathcal{D}$) is exactly $1/2$. This is closely linked to amplitude amplification: we define a good subspace as being spanned by states $|x\ b\rangle$ where $h(x) \neq c(x)$ and $b \in \{0, 1\}$. The `Adaboost` distribution update corresponds to amplifying the magnitude of the projection of $|\psi_c\rangle$ onto the good subspace to be exactly $1/2$. If one has access to the unitary Q_c that prepares $|\psi_c\rangle$, and its inverse, this can be achieved by e.g. the quantum singular value transformation [102]. By repeating this procedure, one can run a boosting algorithm. Unfortunately, the recursive nature of this approach will lead to exponential scaling. Suppose that the amplification subroutine has a very simple form: $Q_c^\dagger U Q_c V$ for some unitaries U and V (that depend on the choice of good subspace). In the first round of boosting, we apply this circuit to produce modified example states $|\widetilde{\psi_c}\rangle$. In the second round of boosting, we apply the circuit $C = \widetilde{Q_c}^\dagger \widetilde{U} \widetilde{Q_c} \widetilde{V}$, where $\widetilde{Q_c}$ is a unitary that prepares $|\widetilde{\psi_c}\rangle$. But then $C = (V^\dagger Q_c^\dagger U^\dagger Q_c) \widetilde{U} (Q_c^\dagger U Q_c V) \widetilde{V}$ makes 4 calls to Q_c or $Q_c^\dagger$. Similarly, the next round of boosting will make 8, and so on. We note that although this leads to exponential scaling in γ, one can use this approach to construct an algorithm no dependence on VC dimension. If one can successfully modify this approach, then there is potential for a large quantum advantage.

Chapter 4

Exact Learning Within the Stabiliser Formalism

4.1 Background

4.1.1 The Stabiliser Formalism

In this Section we will introduce the stabiliser formalism, one of the cornerstones of quantum computation. We will proceed to introduce several (classical) computational primitives within the stabiliser formalism, and explain their importance. Our main results, in subsequent Sections, are new, faster algorithms for these primitives.

The computational importance of the stabiliser formalism was first studied in the context of the Gottesmann-Knill theorem [103, 104], which roughly states that the stabiliser formalism may be classically efficiently simulated. Since then, a plethora of related simulation results have been published, showing where the original simulation result may, and may not, be strengthened, see Ref. [105] for a review.

The stabiliser formalism has become central in the field of quantum error correction [106], which is essential for any realisation of a quantum computer. Almost all error correcting codes make use of the stabiliser formalism: quantum data is encoded in stabiliser "states". Moreover, in most of these codes, the only measurements and gates that can be implemented in a noise-tolerant way correspond to elements of the stabiliser formalism. Indeed, most modern proposals for quantum computers use elements of the stabiliser formalism supplemented with very specific non-stabiliser resources [107].

The full theory of the stabiliser formalism is far too rich to give a full review here - we will give a barebones description of what is required for our results, for a review see Refs. [106, 108]. We will focus on the (most commonly studied) case of a qubit ($\mathcal{H} = \mathbb{C}^2$), but our results can be easily generalised to higher dimensions. The fundamental objects in the stabiliser formalism are the Pauli operators, which have matrices

$$X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{4.1}$$

with respect to the computational basis. Note that $Y = iXZ$. On n-qubits, Paulis are given by the tensor product of single-qubit Paulis. With suitable choices of phases, they form a group (under multiplication), called the (n-qubit) Pauli group, $\mathcal{P}_n$:

$$\mathcal{P}_n = \{(\pm 1)(\pm i)A_1 \otimes \cdots \otimes A_n \mid A_i \in \{\mathbb{1}, X, Y, Z\}\}. \tag{4.2}$$

Paulis clearly have a classical, efficient representation: there are 4 choices for the phase, and 4 choices for the operator on each qubit. In order to leverage this representation, we briefly introduce some notation. Given a collection of operators $\{U_i\}_{i=1}^n$, and a vector $\vec{z} \in \mathbb{Z}_2^n$, we let

$$U^z = \prod_{i=1}^{n} U_i^{z_i}. \tag{4.3}$$

We define $X_i \in \mathcal{P}_n$ as X acting on the i-th qubit and identity on the rest, and Z_i similarly. Using that $Y = iXZ$, we may represent any Pauli as a tuple $(c, d, \vec{p}, \vec{q}) \in \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2^n \times \mathbb{Z}_2^n$ corresponding to the Pauli

$$(-1)^c(-i)^d X^{\vec{p}} Z^{\vec{q}}. \tag{4.4}$$

We often refer to c as the sign-bit of the tuple. Note that X_i, Z_i are Hermitian, and anticommute. A Pauli is Hermitian iff. $d = \vec{p} \cdot \vec{q}$, where $\cdot$ is the $\mathbb{Z}_2^n$ dot product, and thus Hermitian Paulis are be specified by $c, \vec{p}$ and $\vec{q}$. Two paulis $(c_1, d_1, \vec{p}_1, \vec{q}_1)$ and $(c_2, d_2, \vec{p}_2, \vec{q}_2)$ commute iff. $\vec{p}_1 \cdot \vec{q}_2 + \vec{q}_1 \cdot \vec{p}_2 = 0$[1]. Ignoring global phase, the Paulis form an orthonormal basis of $\mathcal{B}(\mathcal{H})$:

Proposition 4.1: The operators $\{X^{\vec{p}} Z^{\vec{q}} \mid \vec{p}, \vec{q} \in \mathbb{Z}_2^n\}$ are an orthonormal basis of $\mathcal{B}(\mathcal{H})$ (with respect to the Hilbert-Schmidt inner product)

Proof: Note that $\mathrm{Tr}(Z_i) = \mathrm{Tr}(X_i) = \mathrm{Tr}(X_i Z_i) = 0$. The result then follows from observing that $X^{\vec{p}} Z^{\vec{q}} X^{\vec{r}} Z^{\vec{s}} \propto X^{\vec{p}+\vec{r}} Z^{\vec{q}+\vec{s}}$, where $+$ is over $\mathbb{Z}_2$. □

Composition of Paulis induces a binary operation on tuples:

$$(c_1, d_1, \vec{q}_1, \vec{p}_1) \circ (c_1, d_1, \vec{q}_1, \vec{p}_1) \simeq (-1)^{c_1}(-i)^{d_1} X^{\vec{q}_1} Z^{\vec{p}_1} (-1)^{c_2}(-i)^{d_2} X^{\vec{q}_2} Z^{\vec{p}_2}, \tag{4.5}$$

$$\simeq (c_1 + c_2 + \vec{p}_1 \cdot \vec{q}_2 + d_1 d_2,\ d_1 + d_2,\ \vec{q}_1 + \vec{q}_2,\ \vec{p}_1 + \vec{p}_2). \tag{4.6}$$

$\circ$ is linear, except on the power of -1, which has a quadratic term $(d_1 d_2)$.

A stabiliser group is a maximal abelian subgroup of $\mathcal{P}_n$ that does not contain $-\mathbb{1}$. We show that a stabiliser subgroup always has size 2^n:

Proposition 4.2: Let $G \leq \mathcal{P}_n$ be a stabiliser subgroup, then $|G| = 2^n$. Moreover, there is a unique state $|\psi\rangle \in \mathcal{H}$ (up to global phase) that is stabilised by every member of G:

$$\forall g \in G,\ g\,|\psi\rangle = |\psi\rangle\,. \tag{4.7}$$

[1] This is precisely the symplectic inner product on the concatenated vectors $(\vec{p}_1, \vec{q}_1)$ and $(\vec{p}_2, \vec{q}_2)$. See Ref. [109] for more of a discussion

Proof: The Pauli subgroup $\langle Z_1, \dots Z_n \rangle$ is an abelian subgroup of size 2^n that does not contain $-\mathbb{1}$, and therefore (by maximality) $|G| \geq 2^n$.
Suppose $P \in G$ is anti-Hermitian. Then $\mathbb{1} = PP^\dagger = -P^2$, so $-\mathbb{1} \in G$, which is forbidden. Thus, every element of G is Hermitian and unitary, and therefore of order 2. Take a minimal generating set $P_1 \dots P_k$ of G. Since the P_i are independent, commute, and are of order 2, $|G| = 2^k$. Note that, since every P_i is Hermitian and unitary, it has eigenvalues ± 1. Hence the orthonormal projector onto the $+1$ eigenspace of P_i is given by $(\mathbb{1} + P_i)/2$. Since the P_i commute, the joint $+1$ eigenspace has a projector given by

$$\prod_{i=1}^{k} \frac{\mathbb{1} + P_i}{2}. \tag{4.8}$$

The dimension of this space can be found by taking the trace of the orthogonal projection:

$$\mathrm{Tr}\left(\prod_{i=1}^{k} \frac{\mathbb{1} + P_i}{2} \right) = \frac{1}{2^k} \sum_{s \in \mathbb{Z}_2^k} \mathrm{Tr}(P^s). \tag{4.9}$$

If $\pm i\mathbb{1} \in G$, then $(\pm i\mathbb{1})^2 = -\mathbb{1} \in G$. Moreover, by minimality of the generating set, no non-trivial product of the P_i can equal the identity. We deduce that all non-trivial products of the P_i are not proportional to the identity, and hence traceless. Therefore, the dimension of the mutual $+1$ eigenspace of the P_i is $2^n/2^k$. We deduce that $k \leq n$, and hence $k = n$. Therefore, $|G| = 2^n$, and there is a unique state (up to a global phase) stabilised by every member of G. □

A state that is stabilised by every member of some stabiliser group is called a stabiliser state. Note that a stabiliser group (or equivalently stabiliser state) has an efficient representation as a (non-unique) choice of generating set, i.e. n commuting, independent (no non-trivial product gives the identity) Paulis. Note that any Pauli in a stabiliser group must be Hermitian (or it would square to $-\mathbb{1}$), and thus may be represented by a sign bit and two vectors, as noted above.

Our work heavily relies on the following characterisation of stabiliser states, in terms of their amplitudes in the computational basis:

Theorem 4.1: A (normalised) state $|\psi\rangle \in \mathcal{H}$ is a stabiliser state iff. there exists a vector subspace $V \leq \mathbb{Z}_2^n$, a vector $\vec{s} \in \mathbb{Z}_2^n$, a linear map $\ell : V \to \mathbb{Z}_2$ and a quadratic form $Q : V \to \mathbb{Z}_2$ such that

$$|\psi\rangle = \frac{1}{\sqrt{|V|}} \sum_{\vec{z} \in V} (-1)^{Q(\vec{z})} i^{\ell(\vec{z})} \left| \vec{z} + \vec{s} \right\rangle. \tag{4.10}$$

For a proof, see Ref. [**dehaene2003Clifford**]. As a consequence of this theorem, the support of any stabiliser state (the indices $\vec{z} \in \mathbb{Z}_2^n$ such that $\langle \vec{z} | \psi \rangle$ is non-zero) is an affine space of $\mathbb{Z}_2^n$. Since we are working over $\mathbb{Z}_2$, one cannot represent quadratic forms by symmetric bilinear forms. We will let $R(-,-)$ denote some choice of bilinear form such that $Q(-) = R(-,-)$.

A unitary that conjugates Pauli operators to Pauli operators is called a Clifford. The collection of

Clifford operators forms the n-qubit Clifford group:

$$\mathcal{C}_n = \{C \in \mathcal{U}(2^n) \mid \forall P \in \mathcal{P}_n,\ CPC^\dagger \in \mathcal{P}_n\}. \tag{4.11}$$

Since $CP_1C^\dagger CP_2C^\dagger = CP_1P_2C^\dagger$, the action of a Clifford by conjugation on the Pauli group is uniquely specified by its action on the Paulis X_i, Z_i. In fact this uniquely determines the Clifford up to a phase:

Proposition 4.3: Suppose that $U, V \in \mathcal{U}(2^n)$ satisfy $\forall P \in \mathcal{P}_n, UPU^\dagger = VPV^\dagger$. Then $U = e^{i\theta}V$ for some $\theta \in \mathbb{R}$.

Proof: Let $W = V^\dagger U$, then W conjugates every Pauli to itself: $\forall P \in \mathcal{P}_n$, $WPW^\dagger = P$. Since the Pauli group spans $\mathcal{B}(\mathcal{H})$ (by proposition 4.1), its natural representation is irreducible. Thus, by Schur's Lemma, $W = \lambda\mathbb{1}$, for some $\lambda \in \mathbb{C}$. But W is also a unitary, so $|\lambda| = 1$ and the result follows. □

Thus, Cliffords have a classical efficient representation, given by recording the image of each X_i, Z_i under conjugation, and noting its global phase. Since global phases are irrelevant in quantum mechanics, the global phase of the Clifford is often ignored.

The stabiliser formalism, sometimes called stabiliser subtheory, considers quantum operations built using stabiliser states, Clifford gates and Pauli measurements (measuring in the eigenbasis of a Pauli operator). We have seen that each of these elements has an efficient classical representation. In fact, one can efficiently update these representations to efficiently clasically simulate the stabiliser formalism (the Gottesmann-Knill theorem), but we will not consider that here. Note that stabiliser states, Paulis and Cliffords also have *inefficient* classical representations. A stabiliser state may also be descried by its amplitudes in the computational basis, i.e. a vector in $\mathbb{C}^{2^n}$, called a statevector. A Clifford (or Pauli) may be described as a unitary matrix of size $2^n \times 2^n$ with respect to the computational basis. These inefficient representations fit into the framework of general quantum computation. Often, it is necessary to convert between efficient and inefficient representations at the interface of the stabiliser formalism with general quantum computation/information. In Sections 4.2 and 4.3 we give several new, fast algorithms for such conversions. We give a few specific use-cases below.

One can combine (generically exponentially) many efficient simulations of stabiliser operations to simulate universal quantum computation. When doing so, one aims to minimise the number of stabiliser simulations that need to be carried out. To achieve this, the input vector needs to be decomposed as a linear combination of stabiliser states, using as few stabiliser states as possible. This process requires expressing stabiliser states as vectors, or alternatively the ability to test whether a vector corresponds to a stabiliser state. Our algorithms have found application in this setting [110] (indeed, this was the original motivation for their creation).

Secondly, a recently proposed [111] approach to finding proofs of quantum advantage in boson sampling (using a specific type of photonic quantum computer [112]), involves translating photonics circuits into unitary matrices, and checking whether those matrices are Cliffords.

Finally, we note that our algorithms have already found applications in synthesising certain types of unitaries, known as Clifford isometries [113]. They use our algorithm to rapidly convert a stabiliser statevector into an efficient representation.

We note that the choice of efficient representation of stabiliser operations is not particularly important; converting between different efficient representations is usually itself efficient. Thus, in most algorithms, the time limiting step is transforming the inefficient representation into an efficient one, or vice-versa.

Finally, we note two notational conventions for this chapter. Firstly, we let $N = 2^n$ for notational brevity. Secondly, it will be useful to index elements of $\mathbb{Z}_2^n$ by elements of $\mathbb{Z}_2^n$. To be precise, let $I : \mathbb{Z}_2^n \to \mathbb{Z}$, $I(\vec{z}) = \sum_{i=0}^{n-1} 2^i z_i$ map the binary representation of an integer to that integer. Then, for $\vec{y}, \vec{z} \in \mathbb{Z}_2^n$, we write $\vec{y}_{\vec{z}}$ to mean $\vec{y}_{I(\vec{z})}$. Note we are zero-indexing our vectors.

4.1.2 Existing Conversion Algorithms

Given the extensive motivation for converting between efficient and inefficient representations of stabiliser states and Cliffords, it is unsurprising algorithms for these conversions have already been developed. We will consider existing implementations from the popular python libraries `Qiskit` [114] and `stim` [115].

Stabiliser state statevector → efficient description of stabiliser state
The `Qiskit` implementation of this conversion is based off the implementation in `stim`.
`stim` converts the amplitudes of a stabiliser state into a circuit of one- and two-qubit Clifford gates, which produces the stabiliser state when acting on the $|0\rangle^{\otimes n}$ computational basis state. The method is currently implemented in the `stabiliser_state_vector_to_circuit` function in the file `circuit_vs_amplitudes.cc`. On input $|\psi\rangle \in \mathbb{C}^N$, `stim` runs the following algorithm:

1. By applying a series of X gates, move the element with the largest amplitude to the first entry of the statevector.

2. Find a non-zero element in the vector which is not the first entry, suppose it has index $\vec{k} \in \mathbb{Z}_2^n$. If no such entry exists, terminate the algorithm.

3. Let $i \in [n]$ be the smallest index such that $\vec{k_i} = 1$. Apply a series of CX gates between qubit i and qubit j, for every $j \neq i$ such that $\vec{k_j} = 1$. Finally, depending on the value of the 2^i-th amplitude of the statevector, apply one of four different single-qubit Cliffords to qubit i. If the size of the support of the statevector has not halved, output that the original state was not a stabiliser state. Otherwise, return to step 2.

Their algorithm relies on the fact that if $|\psi\rangle$ is a stabiliser state, then step 3 halves the support of the statevector on every iteration. Conversely, if their algorithm terminates, they produce a Clifford circuit C such that $C|0\rangle = |\psi\rangle$ (up to a phase), in which case $|\psi\rangle$ is a stabiliser state (whose stabiliser group is generated by $CZ_iC^\dagger$). Thus, the algorithm can be used to extract an efficient representation, as well as verify whether $|\psi\rangle$ was a stabiliser state.

Since the support of the statevector halves after every iteration, in the worst case the algorithm will run step 3 n times. Each iteration of step 3 involves applying $\Omega(n)$ CX gates in the worse case, as well as a single qubit gate. Each application of a CX (or single qubit) gate to $|\psi\rangle$ requires $\Omega(N)$ steps. Thus, in the worst-case, `stim`'s algorithm runs in time $\Omega(N \log^2 N)$ and produces a circuit of size $\Omega(n^2)$.

n commuting, independent Paulis → stabiliser state statevector

As far as we can tell, this is not implemented in `Qiskit`.

`stim` currently implements this conversion in the `state_vector_from_stabilizers` method in `vector_simulator.h`. Given a list of suitable Paulis P_i, the algorithm generates a random intial vector in $\mathbb{C}^N$, and then applies the sequentially orthogonal projection matrix onto the $+1$ eigenspace of P_i, for $i = 1, \ldots, n$, to the statevector. In their implementation, applying the projector takes time $\Omega(N \log N)$ and thus the algorithm runs in time $\Omega(N \log^2 N)$.

Clifford Matrix → efficient representation of Clifford

`Qiskit` implements the conversion via the `from_matrix` method of the `Clifford` class. It is implemented by a brute-force approach - the matrix $CPC^\dagger$ is calculated for P every X_i and Z_i, taking time $\Omega(N^3 \log N)$. Note that this approach additionally verifies that C is a Clifford, by checking that $CPC^\dagger$ does correspond to a Pauli (which takes time $\Omega(N^2)$).

In `stim`, the conversion is implemented in the `unitary_to_tableau` function in `stabilizers_vs_amplitudes.inl`. The function returns a decomposition of the Clifford into one- and two-qubit Clifford gates. On input matrix C `stim` runs the following algorithm:

1. Find a Clifford U (and its circuit representation) such that $U\left|\vec{0}\right\rangle = C\left|\vec{0}\right\rangle$ (i.e. run the stabiliser state statevector algorithm above on the first column of C).
2. Calculate $M = U^\dagger C$. If C is a Clifford, then M will be a Clifford that permutes the computational basis vectors with phases.
3. Find a circuit implementing M. In this step, they verify that M is a Clifford, which in turn verifies that C is a Clifford.
4. Concatenate the circuits for U and M to find a circuit representation of C.

In particular, step 2 involves multiplying $U^\dagger$ and M. The circuit for U has size $\Omega(n^2)$ in the worst-case (see above), and multiplying C by a one or two qubit gate takes time $\Omega(N^2)$. Thus step 2, and hence `stim`'s algorithm, runs in time $\Omega(N^2 \log^2 N)$ in the worst case.

Efficient description of a Clifford → Clifford matrix

In `Qiskit`, the conversion is the `to_matrix` method of the `Clifford` class. The algorithm first decomposes the Clifford as a circuit of one- and two-qubit gates, using the method of [116], and then finds the matrix of that circuit. As noted above, if the circuit is size s, computing the matrix from the circuit by multiplication takes time $\Omega(sN^2)$. In the worst-case (as with `stim`'s stabiliser state algorithm), $s = \Omega(n^2)$ and thus `Qiskit`'s algorithm has a worst-case runtime of $\Omega(N^2 \log^2 N)$.

In `stim`, the conversion is implemented by `tableau_to_unitary` method in `stabilizers_vs_amplitudes.inl`. It relies on the fact that flattening a Clifford matrix (i.e. stacking the columns to produce

a long vector) produces a stabiliser state. In fact, if C is a Clifford such that $CZ_iC^\dagger = U_i$ and $CX_iC^\dagger = V_i$, then flattening the columns of C gives rise to a stabiliser state whose stabiliser group is generated by $Z_i \otimes U_i$, $X_i \otimes V_i$. This can be seen by writing the flattened vector as $|\psi\rangle = \sum_{\vec{z} \in \mathbb{Z}_2^n} |\vec{z}\rangle\, C\, |\vec{z}\rangle$. `stim` runs the stabiliser group to stabiliser statevector algorithm above, with stabilisers given by $Z_i \otimes U_i$, $X_i \otimes V_i$ and then reshapes the resulting statevector back into a matrix. Thus, the worst-case complexity is the determined by the stabiliser conversion algorithm, $\Omega(N^2 \log^2 N)$.

4.2 Converting between representations of stabiliser states

In this Section, we give our algorithms for converting between representations of stabiliser states. We will use 3 representations of a stabiliser state $|\psi\rangle$:

(S1) A statevector: an element of $\mathbb{C}^N$.

(S2) A compact description of the amplitudes of $|\psi\rangle$, using theorem 4.1: A basis for a vector space $V \leq \mathbb{Z}_2^n$, a constant vector $\vec{s} \in \mathbb{Z}_2^n$ and a linear and quadratic form on V (described by their action on the basis).

(S3) A list of n commuting, independent, Hermitian Paulis that generate the stabiliser group of $|\psi\rangle$, called a check matrix.

Note that (S2) and (S3) are efficient descriptions of $|\psi\rangle$, whereas (S1) is inefficient.

4.2.1 Converting from (S1) to (S2)

We are given a statevector of a stabiliser state $|\psi\rangle \in \mathbb{C}^N$ as input. Our algorithm can be split into two distinct stages

1. Find a compact description of the support of $|\psi\rangle$, i.e. a basis for the vector space $V \leq \mathbb{Z}_2^n$ and the constant "shift" vector $\vec{s}$.

2. Find a compact description of the amplitudes in $|\psi\rangle$, i.e. the linear and quadratic forms on V.

When finding the support of $|\psi\rangle$, we use the ordering on $\mathbb{Z}_2^n$ induced by I, i.e. we say $\vec{z} < \vec{y} \Leftrightarrow I(\vec{z}) < I(\vec{y})$. This is the same as the lexicographical ordering on $\mathbb{Z}_2^n$ (read from right to left). We will need the following two technical Lemmas:

Lemma 4.1: Suppose that $V \leq \mathbb{Z}_2^n$ is a dimension k vector subspace of $\mathbb{Z}_2^n$. Suppose that we order $V = \{\vec{v_0}, \ldots \vec{v}_{2^k-1}\}$, i.e. $\vec{v_i} < \vec{v_j}$ iff. $i < j$. Then the map $\alpha : \mathbb{Z}_2^k \to V$, $\alpha(\vec{a}) = \vec{v}_{I(\vec{a})}$ is a linear isomorphism. In particular, $\{\vec{v}_{2^j} \mid j = 0, \ldots, k-1\}$ is a basis of V.

Proof: Take some basis $\vec{w}_0, \ldots, \vec{w}_{k-1}$ of V. Take a matrix W whose rows are given by the $\vec{w}_i$ (as elements of $\mathbb{Z}_2^n$). Let W' be the reduced row-echelon form of W; the rows of W' still form a basis of W. Let $\vec{u}_j$ be the $n-j$th row of W' for $j = 0, \ldots, n-1$. Suppose that the leftmost 1 of $\vec{u}_j$ is in position h_j. As W' is in reduced row-echelon form, the h_i'th element of $\vec{u}_j$ is δ_{ij}. Thus, one can tell whether $\vec{u}_j$ is present in the linear decomposition of $\vec{v} \in V$ by checking whether the h_j'th element is 0 or 1. Since the ordering $<$ on $\mathbb{Z}_2^n$ is decided by the rightmost position at which two vectors disagree, we deduce that $\sum_{i=0}^n z_i\vec{u}_i < \sum_{i=0}^n y_i\vec{u}_i$ iff. $\vec{z} < \vec{y}$. But then $\alpha(\vec{a}) = \sum_{i=0}^{n-1} a_i\vec{u}_i = \vec{v}_{I(\vec{a})}$ is a linear isomorphism as claimed. □

Lemma 4.2: Let $\mathcal{A} \subseteq \mathbb{Z}_2^n$ be an affine space. Let $\vec{s} \in \mathcal{A}$ be the minimal element of $\mathcal{A}$ (with respect to $<$). Let $V = \vec{s} + \mathcal{A}$ and take $\vec{v}, \vec{w} \in V$. Then $\vec{v} < \vec{w}$ iff. $\vec{v} + \vec{s} < \vec{w} + \vec{s}$.

Proof: Suppose $\vec{v} < \vec{w}$, then $\vec{v} = (\vec{a},\ 0,\ \vec{b})$, $\vec{w} = (\vec{d},\ 1,\ \vec{b})$ for some bitstrings $\vec{a}, \vec{b}$ and $\vec{d}$. Let $\vec{s} = (\vec{f},\ x,\ \vec{g})$, where the length of $\vec{g}$ is the same as $\vec{b}$. Then, we see that $\vec{v} + \vec{s} < \vec{w} + \vec{s}$ iff. $x = 0$. If, however, $x = 1$ then $\vec{v} + \vec{w} + \vec{s} = (\vec{a} + \vec{d} + \vec{g},\ 0,\ \vec{g}) < \vec{s}$ is a smaller element of $\mathcal{A}$ than $\vec{s}$, contradicting our assumption. Thus, $\vec{v} + \vec{s} < \vec{w} + \vec{s}$. The converse direction follows similarly. □

Lemmas 4.1 and 4.2 give a method for extracting an efficient description of the support of $|\psi\rangle$. We find the smallest index of $|\psi\rangle$ which has a non-zero amplitude, and we take $\vec{s}$ to be (the binary representation of) this index. We iterate through the remaining entries of $|\psi\rangle$, any time we find a non-zero entry at index $\vec{z}$, we add $\vec{z} + \vec{s}$ to a list. This list will contain all of the vectors in V; it will have size 2^k, where k is the dimension of V. By Lemma 4.2, the entries of this list will be sorted (in increasing order) as $\vec{s}$ was chosen as to be minimal. By Lemma 4.1, we may take the 2^jth element of our list, which we call $\vec{u}_j$, for $j = 0, \ldots, k-1$, as a basis for V.

It remains to find a suitable linear map ℓ and quadratic form Q on V. First, we divide all amplitudes of the statevector by $\langle \vec{s}|\psi\rangle$ so that they take values in $\{\pm 1, \pm i\}$. Second, by looking at $\langle \vec{s} + \vec{u}_j|\psi\rangle$ for each j we deduce the value of $\ell(\vec{u}_j)$ and $R(\vec{u}_j, \vec{u}_j)$. Finally, by considering the values of $\langle \vec{s} + \vec{u}_j + \vec{u}_k|\psi\rangle$ for $k \neq j$, we deduce the value of $R(\vec{u}_j, \vec{u}_k) + R(\vec{u}_j, \vec{u}_k)$ for each $j \neq k$. But these values fully specify Q and ℓ, and the algorithm can be terminated.

We give a brief summary of our algorithm below:

Algorithm 4.1:

Input: An (S1) description of a stabiliser state, i.e. a statevector $|\psi\rangle \in \mathbb{C}^N$

Output: An (S2) description of $|\psi\rangle$, i.e. a basis for a vector space V, a constant vector $\vec{s} \in \mathbb{Z}_2^n$, and a linear map ℓ and quadratic form Q on V (specified by their action on the basis)

1. Find the smallest (binary) index $\vec{s}$ of a non-zero element of $|\psi\rangle$
2. Add $\vec{s}$ to all the indices of non-zero elements of $|\psi\rangle$. This preserves their order. Suppose there are 2^k such indices (so that V has dimension k). Take the non-zero indicies in positions 2^j, for $j = 0, \ldots, k-1$ as a basis of V, which we call $\{\vec{u}_j\}$
3. Find the action of ℓ, Q on V by considering the amplitudes of the form $\langle \vec{s} + \vec{u}_j|\psi\rangle$ and $\langle \vec{s} + \vec{u}_j + \vec{u}_k|\psi\rangle$.

Finally, we analyse the complexity of our algorithm:

1. Finding the smallest nonzero amplitude takes $O(N)$ time.
2. Adding $\vec{s}$ to each non-zero index takes time $O(Nn)$[2].
3. Extracting ℓ takes time $O(n)$. Extracting Q takes time $O(n^2)$.

We see that step 2 dominates the algorithm, and thus it runs in time $O(Nn)$. This is a factor of n faster than the best existing algorithm.

[2]Since this addition is over Z_2^n, it corresponds to XORing two integers, which is practically very fast. Depending on implementation, it could even be done in constant time, reducing the worst-case runtime of our algorithm.

4.2.2 Converting from (S2) to (S1) and Verifying (S1)

Given an (S2) representation of a stabiliser state, there is an obvious method for producing a statevector: one can iterate through all possible linear combinations of the basis vectors in order, i.e. iterate through the vectors $\vec{a} \in \mathbb{Z}_2^k$, with corresponding expansions $\sum_{i=0}^{n-1} a_i \vec{u}_i$. To find the amplitude for a particular $\vec{a}$, one evaluates ℓ and Q. The remaining entries of the statevector are all zero. Evaluating $\sum_{i=0}^{n-1} a_i \vec{u}_i + \vec{s}$ takes time $\Theta(n^2)$, evaluating ℓ and Q also takes time $\Theta(n^2)$. Thus, this approach would take time $\Theta(2^k n^2)$, where k is the dimension of V. In the worst-case, this is $\Theta(Nn^2)$ operations.

This algorithm can be improved by iterating through the elements of V in a more efficient manner. In particular, if a single a_j changes in each step, then one only needs to add $\vec{u}_j$ to the current value of $\sum_{i=0}^{n-1} a_i \vec{u}_i + \vec{s}$ to find the new value. Moreover,

$$Q(\vec{v} + \vec{u}_i) + Q(\vec{v}) = R(\vec{u}_i, \vec{u}_i) + \sum_{j=0}^{n-1} v_j [R(\vec{u}_i, \vec{u}_j) + R(\vec{u}_j, \vec{u}_i)] \tag{4.12}$$

can be evaluated in $O(n)$ time. Similarly, $\ell(\vec{v}+\vec{u}_i)+\ell(\vec{v}) = \ell(\vec{u}_i)$ takes constant time to evaluate. Thus, if we iterate through the elements of V in such a manner, keeping track of the values of $\sum_{i=0}^{n-1} v_i \vec{u}_i + \vec{s}$ and $i^{\ell(\vec{v})}(-1)^{Q(\vec{v})}$, we only need $O(n)$ to find each new amplitude.

The problem of iterating through $\mathbb{Z}_2^k$, flipping a single bit in each step, is well-studied. One of the simplest such schemes is known as the Gray code [117]. The ith iterate of the Gray code can be computed in time $O(k)$, which allows for efficient iteration. Thus, using the Gray code, one can find the statevector of a stabiliser state in time $O(Nn)$. We give a brief overview of our algorithm below:

Algorithm 4.2:

Input: An (S2) description of $|\psi\rangle$, i.e. a basis for a vector space V, a constant vector $\vec{s} \in \mathbb{Z}_2^n$, and a linear map ℓ and quadratic form Q on V (specified by their action on the basis).

Output: An (S1) description of a stabiliser state, i.e. a statevector $|\psi\rangle \in \mathbb{C}^N$.

1. Initialise a statevector of length N of all zeros.
2. Initialise $\vec{a} = (0, \ldots 0) \in \mathbb{Z}_2^k$, keep track of the values of $E[\vec{a}] = \sum_{i=0}^{n-1} v_i \vec{u}_i + \vec{s}$ and $\theta(\vec{a}) = i^{\ell(\vec{v})}(-1)^{Q(\vec{v})}$
3. Iterate $\vec{a}$ through the Gray code, so that a single bit changes in every iteration. At each step, update E and θ (using equation (4.12)), to set the non-zero amplitudes of the statevector.

One can use algorithms 4.1 and 4.2 in sequence to test whether a statevector is a stabiliser state:

Algorithm 4.3:

Input: A statevector, i.e. $|\psi\rangle \in \mathbb{C}^n$

Output: Yes if $|\psi\rangle$ is a stabiliser state, No otherwise

1. Run algorithm 4.1, if at any point it fails (e.g. an entry is not in $\{\pm 1, \pm i\}$ when it should be), output No

2. Run algorithm 4.2 on the output of step 1, giving $|\phi\rangle$.

3. If $|\psi\rangle = |\phi\rangle$ (up to a global phase), ouput Yes, otherwise output No.

The correctness of algorithm 4.3 follows from the iff. in theorem 4.1. Steps 1 and 2 have complexity $O(Nn)$, step 3 has complexity $O(N)$. Thus, the algorithm runs in time $O(Nn)$. This is a factor of n faster than the best existing algorithm. We note that in practice, one can slightly improve the run time of this algorithm by removing duplicate steps. For example, one only needs to check that the non-zero amplitudes are equal in step 3. However, the complexity will not be reduced by such optimisations.

4.2.3 Converting between (S2) and (S3)

We begin by establishing some required relations between a stabiliser state expressed in (S2) form and the Pauli gates that stabilise it. Recall the expression of the amplitudes of a stabiliser state from Theorem 4.1. An arbitrary stabiliser state $|\psi\rangle$ can be expressed in the form:

$$|\psi\rangle \propto \sum_{\vec{z}\in V} (-1)^{Q(\vec{z})} i^{\ell(\vec{z})} |\vec{z}+\vec{s}\rangle\,, \tag{4.13}$$

where $V \leq \mathbb{Z}_2^n$ is a vector subspace of dimension $k \leq n$, $\vec{s} \in \mathbb{Z}_2^n$, and $Q, \ell : V \to \mathbb{Z}_2$.

Suppose a Pauli P stabilises $|s\rangle$. Since P has a $+1$ eigenvalue, it must be Hermitian, and thus takes the form $P = (-1)^c (-i)^{\vec{p}\cdot\vec{q}} X^{\vec{q}} Z^{\vec{p}}$. We can see immediately that $\vec{p} \in V$ or else P changes the support of $|s\rangle$. Moreover, by comparing the amplitudes of $|\psi\rangle$ and $P|\psi\rangle$, we find that P stabilises $|\psi\rangle$ iff.

$$\ell(\vec{q}) = \vec{p}\cdot\vec{q}, \tag{4.14}$$

$$\text{For all } \vec{z} \in V: \quad Q(\vec{z}) = c + Q(\vec{z}+\vec{q}) + \vec{p}\cdot(\vec{z}+\vec{q}+\vec{s}) + (\vec{p}\cdot\vec{q})\ell(\vec{z}) \tag{4.15}$$

We can then express equation (4.15) as

$$[R(\vec{q},\vec{z}) + R(\vec{z},\vec{q}) + (\vec{p}\cdot\vec{q})\ell(\vec{z}) + \vec{p}\cdot\vec{z}] + [R(\vec{q},\vec{q}) + \vec{p}\cdot(\vec{q}+\vec{s}) + c] = 0, \tag{4.16}$$

where the first term is a linear function of $\vec{z} \in V$ and the second term is a constant. This holds iff. both terms are identically zero. Therefore, P stabilises $|\psi\rangle$ iff.

$$\vec{q} \in V, \tag{4.17}$$

$$\ell(\vec{q}) = \vec{p}\cdot\vec{q}, \tag{4.18}$$

$$c = \vec{p}\cdot(\vec{q}+\vec{s}) + R(\vec{q},\vec{q}), \tag{4.19}$$

$$\vec{p}\cdot\vec{z} = R(\vec{q},\vec{z}) + R(\vec{z},\vec{q}) + (\vec{p}\cdot\vec{q})\ell(\vec{z}), \tag{4.20}$$

where the latter equality is as linear functions from V to $\mathbb{Z}_2$. Equations (4.17)-(4.20) can be read in two different ways; firstly if ℓ, R, V and $\vec{s}$ are known and we wish to find suitable $\vec{p}, \vec{q}$ and c, or vice-versa. This allows us to convert between (S2) and (S3).

Firstly, suppose we are given a list of n commuting Paulis that generate the stabiliser group of the stabiliser state: $(c_i, \vec{a}_i, \vec{b}_i)$ for $i = 1, \ldots n$ (they are Hermitian, so we do not need to specifiy the power

of i). We collect the Paulis into the rows of a matrix:

$$M = \begin{bmatrix} \vec{a}_1 & \vec{b}_1 & c_1 \\ \vdots & \vdots & \vdots \\ \vec{a}_n & \vec{b}_n & c_n \end{bmatrix}. \tag{4.21}$$

Swapping two rows of M corresponds to permuting the generators of the stabiliser group. Adding two rows together, and then adjusting the sign bit according to equation (4.5), corresponds to composing two generators, which still gives a valid list of generators of the stabiliser group. Thus, we may perform row reduction on M until the first $2n$ columns are in reduced row-echelon form:

$$M' = \begin{bmatrix} \vec{q}_1 & \vec{p}_1 & c_1 \\ \vdots & \vdots & \vdots \\ \vec{q}_k & \vec{p}_k & c_k \\ \vec{0} & \vec{\rho}_1 & \gamma_1 \\ \vdots & \vdots & \vdots \\ \vec{0} & \vec{\rho}_{n-k} & \gamma_{n-k} \end{bmatrix}. \tag{4.22}$$

The rows of M' still form a generating set of the stabiliser group.

We can take the $\vec{q}_i$ as a basis for V. The shift $\vec{s}$ may be taken as any solution to the system of $n-k$ equations $\vec{\rho}_i \cdot \vec{s} = \gamma_i$. By equation (4.18) and the fact that all the stabilising Pauli gates must have order 2, we can extract ℓ and define it on the basis for V via the equations $\ell(\vec{q}_i) = \vec{p}_i \cdot \vec{q}_i$.

We can extract the diagonal entries of R by rearranging equation (4.20):

$$R(\vec{q}_i, \vec{q}_i) = c_i + \vec{p}_i \cdot (\vec{q}_i + \vec{z}_0). \tag{4.23}$$

Finally, we can extract the remaining entries of R by rearranging equation (4.20):

$$R(\vec{q}_i, \vec{q}_j) + R(\vec{q}_j, \vec{q}_i) = \vec{p}_i \cdot \vec{q}_j + (\vec{p}_i \cdot \vec{q}_i)(\vec{p}_j \cdot \vec{q}_j). \tag{4.24}$$

The resulting (S2) state is stabilised by all the rows of M', and thus must correspond to the correct stabiliser state. We summarise our conversion algorithm below:

Algorithm 4.4:

Input: An (S3) description of $|\psi\rangle$, i.e. a list of n-commuting, independent, Hermitian Pauli gates

Output: An (S2) description of $|\psi\rangle$, i.e. a basis for a vector space V, a constant vector $\vec{s} \in \mathbb{Z}_2^n$, and a linear map ℓ and quadratic form Q on V (specified by their action on the basis)

1. Collect the Paulis into a matrix, as in equation (4.21). Row reduce the first $2n$ columns whilst updating the sign bits consistently (to ensure that the resulting check matrix represents the same stabiliser state), until it is in the form of equation (4.22).

2. Take the $\vec{q}_i$ as a basis for V and a solution to $\vec{\rho}_i \cdot \vec{s} = \gamma_i$ for the shift vector $\vec{s}$.
3. Compute $\ell(\vec{q}_i) = \vec{q}_i \cdot \vec{p}_i$ and store these as a vector representing ℓ (with respect to the basis $\{\vec{q}_1, \ldots, \vec{q}_k\}$ of V).
4. Compute a matrix representing R relative to the same basis using equations (4.19) and (4.20).

We give a complexity analysis of algorithm 4.4:

1. Gaussian elimination for row reduction takes time $O(n^3)$.
2. Finding a solution to the row reduced equation takes time $O(n)$.
3. Computing ℓ takes time $O(k)$, where k is the dimension of the affine subspace.
4. Computing Q takes time $O(k^2)$.

Thus, algorithm 4.4 runs in time $O(n^3)$.

Conversely, suppose we have $\vec{s}$, a basis for V and ℓ, R with respect to this basis. We can generate the $\vec{q}_i$ by row reducing the basis $\vec{v}_i$. As we perform the row reduction, we update the matrices of ℓ and R such that they are valid with respect to the new basis.

Applying equation (4.20) to $(\vec{p}, \vec{q}) = (\vec{\rho}_i, \vec{0})$, we see that the $\vec{\rho}_i$ can be taken to be a basis of the null space of matrix whose rows are $\vec{q}_i$, i.e. a basis of $V^\perp$. We can then extract γ_i using equation (4.19): $\gamma_i = \vec{\rho}_i \cdot \vec{s}$.

Using equations (4.18) and (4.20) we can use our R and ℓ to give linear systems for each $\vec{p}_i$:

$$\vec{p}_i \cdot \vec{q}_j = R(\vec{q}_i, \vec{q}_j) + R(\vec{q}_j, \vec{q}_i) + \ell(\vec{q}_i)\ell(\vec{q}_j). \tag{4.25}$$

These determine each $\vec{p}_i$ up to the addition of an element of $V^\perp$. Finally, we compute the c_i using equation (4.19):

$$c_i = \vec{p}_i \cdot (\vec{q}_i + \vec{s}) + R(\vec{q}_i, \vec{q}_i). \tag{4.26}$$

We note that these the generated Paulis commute, are Hermitian, and are independent and thus must generate the correct stabiliser group. We summarise our conversion algorithm below:

Algorithm 4.5:

Input: An (S2) description of $|\psi\rangle$, i.e. a basis for a vector space V, a constant vector $\vec{s} \in \mathbb{Z}_2^n$, and a linear map ℓ and quadratic form Q on V (specified by their action on the basis).

Output: An (S3) description of $|\psi\rangle$, i.e. a list of n-commuting, independent, Hermitian Pauli gates.

1. Row reduce the basis $\{\vec{u}_1, \ldots, \vec{u}_k\}$ of V to find $\{\vec{q}_1, \ldots, \vec{q}_k\}$ and a basis $\{\vec{\rho}_1, \ldots, \vec{\rho}_{n-k}\}$ of $V^\perp$. Update the representations of ℓ and R to this new basis.
2. Using the notation of equation (4.22), compute $\gamma_i = \vec{\rho}_i \cdot \vec{s}$.

3. Find a solution for each of the $\vec{p}_i$ from the (underdetermined) linear systems of equation (4.25).

4. Compute $c_i = \vec{p}_i \cdot (\vec{q}_i + \vec{s}) + Q(\vec{q}_i)$.

We give a complexity analysis of algorithm 4.5:

1. Row reduction takes time $O(nk^2)$, where k is the dimension of V. Finding a basis for the null space then takes time $O((n-k)k)$. Computing the new matrices for R and ℓ can be done in parallel to this step, altering after each row operation. Each update takes constant time (as it corresponds to swapping a row or adding two rows together) and thus this does not affect the complexity.

2. This takes time $O((n-k)n)$.

3. This takes time $O(kn)$.

In general, when $k = \Omega(n)$, Algorithm 4.5 runs in time $O(n^3)$. We note that if one wishes to convert (S1) to (S3), one can run algorithms 4.1, 4.5 sequentially. The basis of V from algorithm 4.1 is already row reduced, and thus step 1 can be skipped, and algorithm 4.5 runs in time $O(n^2)$.

4.3 Converting between representations of Clifford operators

In this Section, we give our algorithms for converting between representations of Clifford operators. We will use 2 representations of a Clifford operator C:

(C1) A matrix: an element of $\mathcal{M}_N(\mathbb{C})$.

(C2) A compact description of the image of the basic Pauli gates under conjugation: $U_i = CZ_iC^\dagger$, $V_i = CX_iC^\dagger$. Each U_i and V_i will be Hermitian, so admits a more compact description (as noted above).

4.3.1 Converting from (C1) to (C2)

We are given the matrix representing a Clifford C. We wish to extract $U_i = CZ_iC^\dagger$ and $V_i = CX_iC^\dagger$ for $i = 1, \ldots, n$. We begin by examining the stabiliser group of the first column of C.

Lemma 4.3: For a Clifford gate C, the stabiliser group of $C\left|\vec{0}\right\rangle$ (the first column of C) is generated by $U_i = CZ_iC^\dagger$.

Proof: $U_iC\left|\vec{0}\right\rangle = CZ_i\left|\vec{0}\right\rangle = C\left|\vec{0}\right\rangle$. The U_i are independent since the Z_i are. □

We can rapidly extract a set of generators P_i of the stabiliser subgroup of $C\left|0\right\rangle$ using algorithms 4.1 and 4.5 in series. Since P_i and U_i generate the same stabiliser subgroup, we have that $P_i = U^{\vec{\rho}_i}$ for some $\vec{\rho}_i \in \mathbb{Z}_2^n$. Thus,

$$P_i\, C\left|\vec{z}\right\rangle = U^{\vec{\rho}_i} C\left|\vec{z}\right\rangle = CZ^{\vec{\rho}_i}\left|\vec{z}\right\rangle = (-1)^{\vec{\rho}_i \cdot \vec{z}} C\left|\vec{z}\right\rangle. \tag{4.27}$$

Suppose that P_i corresponds to $(c_i, \vec{q}_i, \vec{p}_i)$. Fix $j \in [n]$. We find a non-zero entry of the $\vec{e}_j$'th column of C: $C\left|\vec{e}_j\right\rangle$. Suppose it has index $\vec{z}$. Then, by comparing $\langle \vec{z} + \vec{q}_i | C | \vec{e}_j \rangle$ and $(-1)^{c_i + \vec{p}_i \cdot \vec{z}} i^{\vec{p}_i \cdot \vec{q}_i} \langle \vec{z} | C | \vec{e}_j \rangle$

we can find $(\vec{\rho}_i)_j$. By repeating this for $i, j = 1, \ldots, n$, we can find all of the $\vec{\rho}_i$.

As the P_i are independent, the $\vec{\rho}_i$ are also linearly independent. Thus, we may invert the matrix whose columns are $\vec{\rho}_i$; we call the rows of this inverse $\vec{\mu}_i$. We deduce that $P^{\vec{\mu}} = U$, and thus we find the U_i.

Having found the U_i, we must now find the V_i. First, we find Pauli gates W_i of order 2 such that U_i and W_j anticommute if and only if $i = j$. Constructing such W_i is straightforward. If $U_i \propto X^{\vec{q}_i} Z^{\vec{p}_i}$, define U to be the $n \times 2n$ matrix over $\mathbb{Z}_2$ whose i-th row is $(\vec{q}_i, \vec{p}_i)$. As these rows are linearly independent, $U^+ = U^T(UU^T)^{-1}$ is a right inverse: $UU^+ = \mathbb{1}$. We may thus take W_i to be $(-i)^{\vec{\alpha}_i \cdot \vec{\beta}_i} X^{\vec{\beta}_i} Z^{\vec{\alpha}_i}$ where $(\vec{\alpha}_i, \vec{\beta}_i)$ is the i-th column of U^+. We make use of the following Lemma:

Lemma 4.4: Suppose the n Pauli gates U_i generate a stabiliser group, the Pauli gates V_i, W_i are of order 2, and that V_i commutes (or anticommutes) with U_i if and only if W_i does. Then $V_i = (-1)^{c_i}(-i)^{d_i} W_i U^{\vec{v}_i}$ for some $\vec{v}_i \in \mathbb{Z}_2^n$ and $c_i, d_i \in \mathbb{Z}_2$.

Proof: Define the Pauli gates $T_i = i^{d_i} W_i V_i$ where d_i is 0 if W_i and V_i commute and 1 if they anticommute. They commute with every U_j: $T_i U_j = i^{d_i} W_i V_i U_j = i^{d_i} U_j W_i V_i = U_j T_i$. Further, the T_i are of order 2. Therefore, for each i, either T_i or $-T_i$ is a member of the stabiliser group generated by U_i. □

Therefore, in the notation of Lemma 4.4, our final task is to find the V_i is to determine the appropriate $\vec{v}_i \in \mathbb{Z}_2^n$ and $c_i, d_i \in \mathbb{Z}_2$. We note that

$$\begin{aligned} C\left|\vec{z} + \vec{e}_i\right\rangle &= CX_i \left|\vec{z}\right\rangle, && (4.28)\\ &= V_i C \left|\vec{z}\right\rangle, && (4.29)\\ &= (-1)^{c_i}(-i)^{d_i} W_i U^{\vec{v}_i} C \left|\vec{z}\right\rangle, && (4.30)\\ &= (-1)^{c_i}(-i)^{d_i} W_i C Z^{\vec{v}_i} \left|\vec{z}\right\rangle, && (4.31)\\ &= (-1)^{\vec{v}_i \cdot \vec{z} + c_i}(-i)^{d_i} W_i \, C \left|\vec{z}\right\rangle, && (4.32) \end{aligned}$$

where $\vec{e}_i$ is the i-th unit vector of $\mathbb{Z}_2^n$.

By comparing the relative phases of the columns of C corresponding to $\vec{z}$ and $\vec{z} + \vec{e}_i$, we gain information about the required variables $\vec{v}_i, c_i, d_i$. Choosing $\vec{z} = \vec{0}$, we determine the c_i, d_i. Choosing $\vec{z}$ to be of Hamming weight 1, we determine the components of the $\vec{v}_i$. We can then multiply W_i by $(-1)^{c_i} i^{d_i} U^{\vec{v}_i}$ to find V_i.

To determine the relative phases of two columns, we need only check two nonzero entries: one from each column. To avoid potentially having to perform a high number of checks for columns that contain many zeros, we can compute the support of each column first. We do this using that the fact that the support of $C\left|\vec{z}\right\rangle$ is the same as the support of $W^{\vec{z}} C \left|\vec{0}\right\rangle$, as each U_i either stabilises or antistabilises every column. We give a summary of our algorithm below:

Algorithm 4.6:

Input: A (C1) description of a Clifford C, i.e. a unitary matrix in $\mathcal{M}_N(\mathbb{C})$.

Output: A (C2) description of C, i.e. a compact description of $U_i = CZ_iZ^\dagger$ and $V_i = CX_iC^\dagger$ for every $i = 1, \ldots n$.

1. Compose algorithms 4.1 and 4.5 to extract generators P_i for the stabiliser group of $C\left|\vec{0}\right\rangle$.
2. For each $\vec{z}$ with Hamming weight 1, find a nonzero entry of $C\left|\vec{z}\right\rangle$. For each i, use these nonzero entries to determine $\vec{\rho}_i$ satisfying $P_iC\left|\vec{z}\right\rangle = (-1)^{\vec{\rho}_i\cdot\vec{z}}\left|\vec{z}\right\rangle$.
3. Invert the $n \times n$ matrix over $\mathbb{Z}_2$ whose columns are $\vec{\rho}_i$ to find one with rows $\vec{\mu}_i$. Conclude that $U_i = P^{\vec{\mu}_i}$.
4. Compute the bitstrings specifying the W_i (up to sign) by taking the pseudoinverse of the matrix whose rows are the bitstrings specifying the U_i.
5. Correct the W_i to V_i by comparing two nonzero entries of the $\vec{z}$-th and $\vec{z} + \vec{e}_i$-th columns of C for $\vec{z}$ being $\vec{0}$ or having Hamming weight 1.

We give a complexity analysis of algorithm 4.6:

1. The conversion (S1) to (S3) takes $O(Nn)$ time as discussed above.
2. In the worst-case, we search for time $O(N)$ to find a nonzero element of a given column. Once we have found one for each column of Hamming weight 1, taking time $O(Nn)$, we can extract all the $\vec{\rho}_i$ vectors in time $O(n^2)$.
3. It takes time $O(n^3)$ to find the pseudoinverse of the U. In fact, the output of our (S1) to (S3) conversion gave us the row-reduced version of U; by keeping track of the row operations when constructing the U_i from the P_i, we can construct the pseudoinverse simultaneously. We then require time $O(n^2)$ to find each of the W_i.
4. Comparing the relative phase of the columns corresponding to Hamming weight 1 and 2 takes time $O(n^3)$ as there are n^2 such columns and finding a nonzero element takes time $O(n)$. Then, finding the V_i involves multiplying each W_i by $O(n)$ Pauli gates, where each Pauli multiplication takes time $O(n)$. Thus, this step also runs in time $O(n^3)$.

The algorithm is dominated by Steps 1 and 2 and thus takes time $O(Nn)$. This is a factor of Nn faster than the best existing algorithm. Since $N = 2^n$, this is an exponential advantage in runtime.

4.3.2 Converting from (C2) to (C1) and Verifying (C1)

We now suppose that we are given the $U_i = CZ_iC^\dagger, V_i = CX_iC^\dagger$, and wish to find the matrix of C (up to a global phase). By Lemma 4.3, the first column of C is stabilised by all of the U_i. Thus, the first column can be quickly computed by combining algorithms 4.4 and 4.2. Then, to find the remaining columns, we iterate through the Gray code (see above). By equation (4.29), columns corresponding to consecutive iterates of the Gray code will differ by the application of a single V_i (corresponding to the bit at which the iterates differ). Thus, to produce one column from the previous one, we

need to multiply a vector by a single Pauli. We can perform this multiplication in time $O(Nn)$: If $P \simeq (c, d, \vec{q}, \vec{p})$ then $P|\vec{z}\rangle = (-1)^{c+\vec{p}\cdot\vec{z}} i^d |\vec{z}+\vec{q}\rangle$ takes time $o(n)$ to evaluate. We summarise our algorithm below:

Algorithm 4.7:

Input: A (C2) description of C, i.e. a compact description of $U_i = CZ_iZ^\dagger$ and $V_i = CX_iC^\dagger$ for every $i = 1, \dots n$.

Output: A (C1) description of a Clifford C, i.e. a unitary matrix in $\mathcal{M}_N(\mathbb{C})$.

1. Find the first column of C by running algorithms 4.4 and 4.2 sequentially, with input $U_1, \dots, U_n$.
2. Find the remaining columns by iterating through the Gray code, multiplying the current column by a single V_i in every step

We give a complexity analysis of algorithm 4.7:

1. Running algorithms 4.4 and 4.2 takes time $O(Nn)$, as described above.
2. As discussed, finding the column corresponding to the next iterate of the Gray code from the previous takes time $O(Nn)$. There are N such columns and thus this step takes time $O(N^2n)$.

The algorithm is dominated by step 2, giving a total time of $O(N^2n)$. Again, this is a factor of n faster than existing implementations.

As with the stabiliser state case, one can use algorithms 4.6 and 4.7 in sequence to test whether a matrix is a Clifford:

Algorithm 4.8:

Input: A matrix $C \in \mathcal{M}_N(\mathbb{C})$.

Output: Yes if C is a Clifford, No otherwise.

1. Run algorithm 4.6, if at any point it fails (e.g. an entry is not in $\{\pm 1, \pm i\}$ when it should be), output No.
2. Run algorithm 4.7 on the output of step 1, giving a matrix M.
3. If $M = C$ (up to a global phase), ouput Yes, otherwise output No.

If the algorithm outputs yes, then U is a Clifford matrix (as algorithm 4.7 only outputs Cliffords), and thus the algorithm is correct. Step 1 has complexity $O(Nn)$, step 2 has complexity $O(N^2n)$, step 3 has complexity $O(N^2)$. Thus, the algorithm runs in time $O(N^2n)$. This is a factor of n faster than the best existing algorithm. As in the stabiliser state case, one can reduce redundancy between steps 1 and two of the algorithm (e.g. one does not need to find the first column again in step 2).

4.4 Benchmarking our Algorithms

Implementations of our algorithms are available at `https://github.com/WilfredSalmon/Stabiliser`. In this Section, we benchmark the implementation of our algorithms against the existing implementations provided by `stim` and `Qiskit`. We find that in almost all cases, the asymptotic advantage of

our algorithm translates into a practical speedup. We also consider theoretically worse-case inputs to our algorithms. We find that our algorithms still have fast runtimes on these extremal cases.

4.4.1 Stabiliser Algorithms

We begin by benchmarking the conversion (S1) → an efficient representation. Our algorithm converts (S1) to (S2), `stim` converts (S1) to a circuit that synthesises the stabiliser state (see Section 4.1.2). In figure 4.1 we compare (log) the execution time of our algorithm to `stim`'s, for different numbers of qubits. We consider the average case input to the algorithm - i.e. a uniformly sampled stabiliser state. We consider two paradigms for our algorithm, one where we assume the input is a stabiliser state (labelled as "with assumption"), and one where we do not guarantee the input is a stabiliser state (labelled as "without assumption"). In the latter case, we must run our (S1) → (S2) conversion, and additionally verify that the input is a stabiliser state (`stim`'s algorithm is the same in both cases). We shade the 33-66th percetile range of execution times (across 1000 different runs of the algorithm on each input), to observe the variance in runtimes. We see that our algorithm is between a half and two orders of magnitude faster than `stim`'s implementation, in average case runtime, and on the majority of stabiliser states.

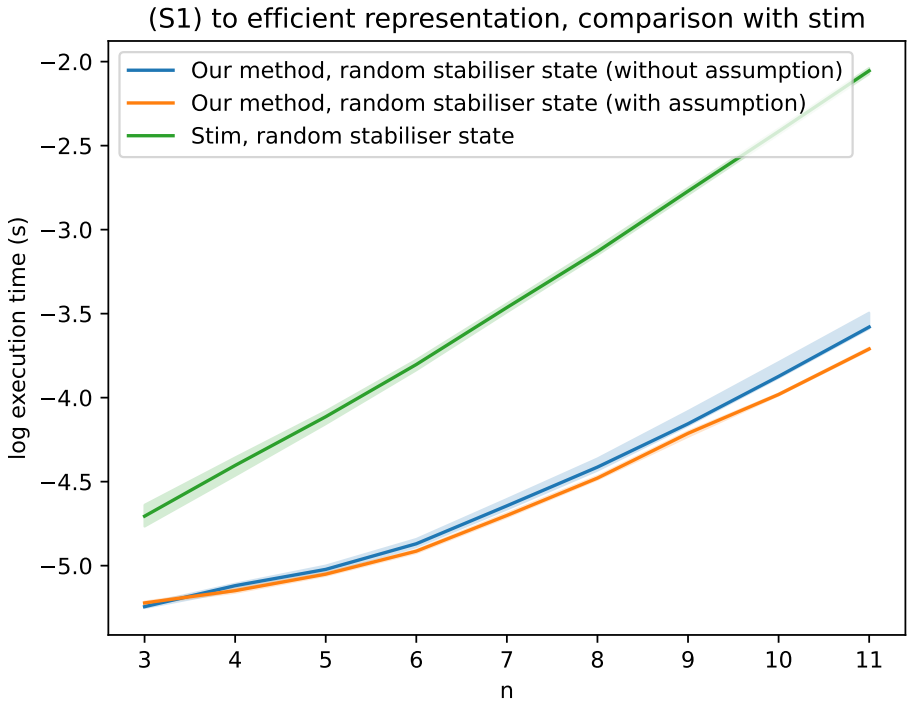


Figure 4.1: Comparison of our (S1) → efficient representation algorithm (i.e. (S1) → (S2)) with `stim`'s. The algorithms were run 1000 times for each qubit number, with a uniformly random stabiliser state as input. For each input, we plot (log base 10) the average execution time (solid line), and shade the 33-66th percentile range. We test our algorithm in two paradigms: when its input is declared to be stabiliser state (labelled as "with asssumption"), and when there are no guarantees on its input (labelled as "without asssumption").

We also consider the running time of our algorithm in two theoretically extremal cases. Firstly, we consider the computational basis state $|0\rangle^{\otimes n}$, which has a single non-zero amplitude. Secondly, we consider a random full support stabiliser state, i.e. we fix $V = \mathbb{Z}_2^n$, and sample ℓ and Q uniformly

at random. In both cases, we do not assume that the input to our algorithm is a stabiliser state. Theoretically, these correspond to the best and worst-case inputs to our algorithm, respectively. The computational basis state $|0\rangle^{\otimes n}$ requires checking every entry of the vector exactly once, whereas the full support state requires the maximal number of steps in our algorithm. We plot the (average) execution times for these inputs in figure 4.2. Unsurprisingly, we see that our algorithm is much faster with input $|0\rangle^{\otimes n}$, than the average case. Surprisingly, our algorithm also executes much faster than average on a random full support stabiliser state, showing a disconnect between theory and practical implementations. This is likely due to overheads in implementation (such as speculative code execution and memory layout).

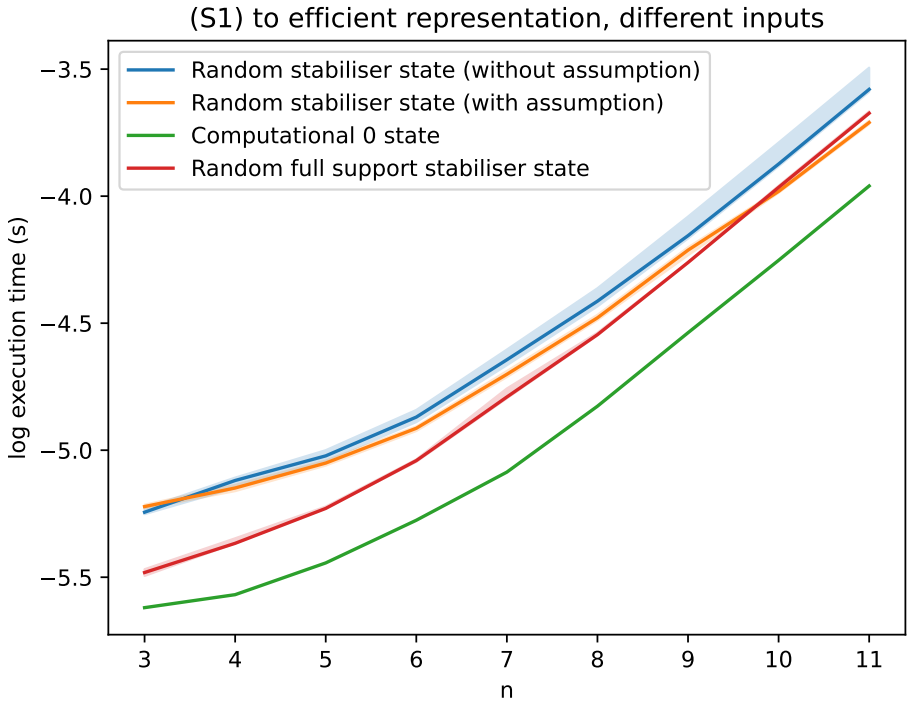


Figure 4.2: Comparison of the execution time of our (S1) → (S2) algorithm, for a variety of inputs. The algorithms were run 1000 times for each qubit number and type of input. For each type of input, we plot (log base 10) the average execution time (solid line), and if the input is inherently random, we additionally shade the 33-66th percentile range. We test our algorithm on a random stabiliser state, both with and without a guarantee on the input, the computational basis state $|0\rangle^{\otimes n}$, and a uniformly random stabiliser state with full support. For the latter two inputs, we do not assume that the input is a valid stabiliser state.

Finally, we benchmark our algorithm for verifying whether a statevector $|\psi\rangle \in \mathbb{C}^N$ is a stabiliser state. We note that the (S1) → (S2) conversion algorithm, without an assumption on its input, is identical to the verification algorithm. Thus, we have already tested the worst-case for acceptance. It remains to benchmark the worst-case for rejecting. This corresponds to a statevector whose amplitudes are almost identical to those of a full-support stabiliser state, but differ in a single entry. We call such a state an "almost" stabiliser state. We benchmark our algorithm by uniformly sampling from "almost" stabiliser states - we sample a full-support stabiliser state uniformly at random, and then modify a single entry uniformly at random (such that the resulting state is not a stabiliser state). The execution times for our algorithm, as well as `stim`'s algorithm are plotted in figure 4.3. As before, we see that

our algorithm is significantly faster than `stim`'s on a random stabiliser state input. Our algorithm is faster on a uniformly random "almost" stabiliser state than a uniformly random stabiliser state, since the algorithm can terminate as soon as an inconsistent entry is found. Furthermore, it is slightly faster than `stim`'s, on average. We remark that "almost" stabiliser states are a "good" case for `stim` - they are usually rejected within one iteration (see Section 4.1.2 for a description of their algorithm), and despite this, our algorithm is still faster than `stim`'s. We deduce that our algorithm has good performance across a wide range of inputs.

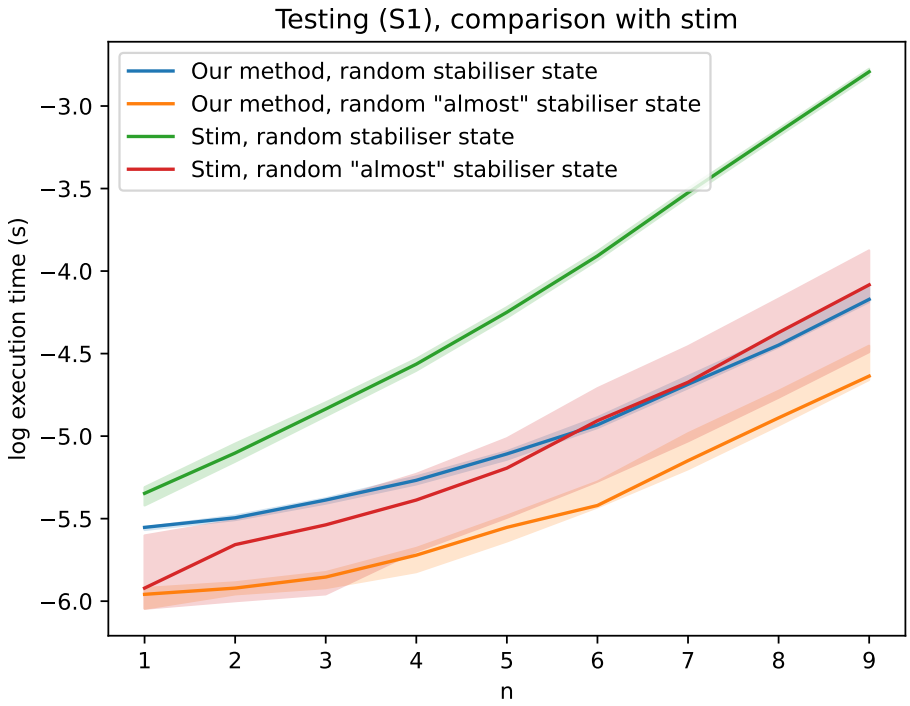


Figure 4.3: Comparison of our algorithm for testing whether a statevector is a stabiliser state with `stim`'s. The algorithms were run 1000 times for each qubit number and type of input. For each type of input, we plot (log base 10) the average execution time (solid line), and shade the 33-66th percentile range. We test the algorithms with a uniformly random stabiliser state, and a uniformly random "almost" stabiliser state (a statevector that differs from a full rank stabiliser state in a single entry).

4.4.2 Clifford Algorithms

In this Section, we benchmark our algorithms for dealing with Clifford conversions against those in `stim` and `Qiskit`. We begin by benchmarking the conversion (C1) $\rightarrow$ an efficient representation. Our algorithm, `stim` and `Qiskit` (see Section 4.1.2) convert (C1) to (C2). In figure 4.4 we compare (log) the execution time of our algorithm to `stim`'s and `Qiskit`'s, for different numbers of qubits. We consider the average case input to the algorithm, i.e. a uniformly sampled Clifford matrix (using `Qiskit`'s implementation of the algorithm in Ref. [116]). As in the stabiliser state case, we run our algorithm twice, once with the assumption that the input is a Clifford matrix, and once without this assumption, requiring us to additionally verify the input (`stim` and `Qiskit` run the same algorihtm in both cases). The 33-66th percentile range is too small to be visible on the graph; it is on the order of 0.01 for each algorithm. We see that our algorithm is roughly one and a half orders of magnitudes faster than `stim`'s and `Qiskits`'s, for the considered qubit numbers. `Qiskit` has a theoertically much

slower running time that `stim`, but, surprisingly, has a faster average run time, for large numbers of qubits. This is due to the high level of optimisation of matrix multiplication in scientific python packages.

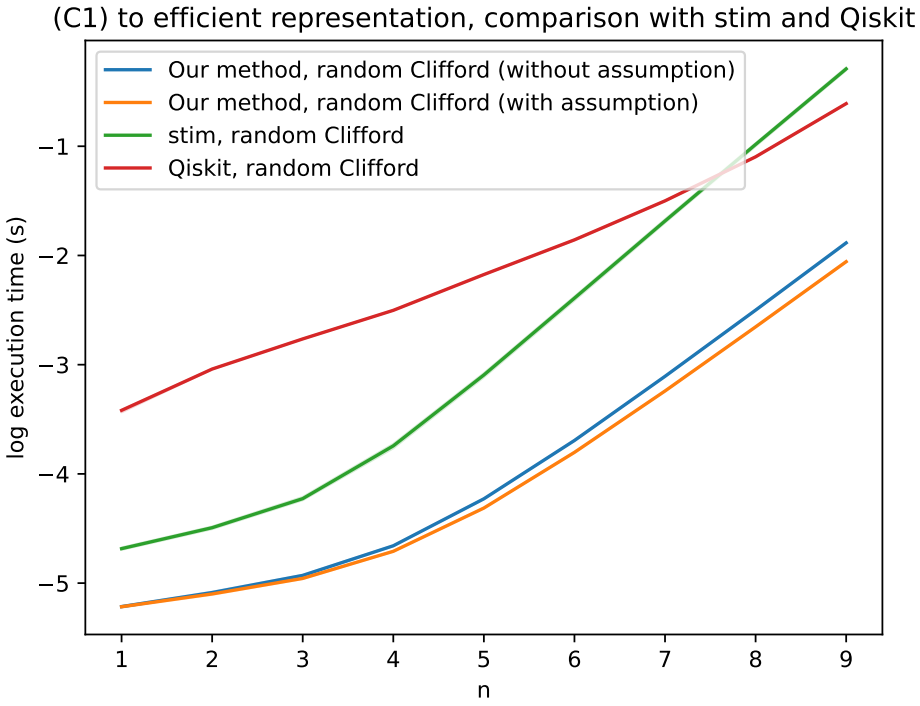


Figure 4.4: Comparison of our (C1) $\rightarrow$ efficient representation algorithm (i.e. (C1) $\rightarrow$ (C2)) with `stim`'s and `Qiskits`'s. The algorithms were run 1000 times for each qubit number, with a uniformly random Clifford matrix as input. For each input, we plot (log base 10) the average execution time (solid line), and shade the 33-66th percentile range (though these are too small to see). We test our algorithm in two paradigms: when its input is declared to be Clifford (labelled as "with assumption"), and when there are no guarantees on its input (labelled as "without assumption").

We also consider the running time of our algorithm in several theoretically extremal cases. Firstly, we consider two sparse matrices; the identity matrix and a matrix with ones on the leading anti-diagonal, and zeros elsewhere, which we call the anti-identity matrix. Secondly, we consider the n-fold tensor product of the Hadamard matrix: $H^{\otimes n}$, which has a non-zero amplitude in every entry. In all cases, we do not assume the input to our algorithm is a Clifford matrix (so that we must run the conversion, and additionally verify the input). These inputs correspond to the theoretical extremes of our algorithm: in the sparse case, the algorithm spends a significant amount of time searching for non-zero entries. In the Hadamard matrix case, it spends a significant time verifying the matrix is indeed a Clifford. We plot the (average) execution times in figure 4.2. As expected, for large input numbers, we see that the extremal inputs are worst-cases for our algorithm. However, the execution time is very similar in all cases, showing that our algorithm has good performance on all Clifford matrices.

We conclude by benchmarking our algorithm for verifying whether a matrix $M \in \mathcal{M}_N(\mathbb{C})$ is a Clifford. Similarly to the stabiliser state case, we note that the (C1)$\rightarrow$ (C2) conversion algorithm, without an assumption on its input, is identical to the verification algorithm. Thus, we have already tested the worst-case for acceptance. It remains to benchmark the worst-case for rejecting. This corresponds to a

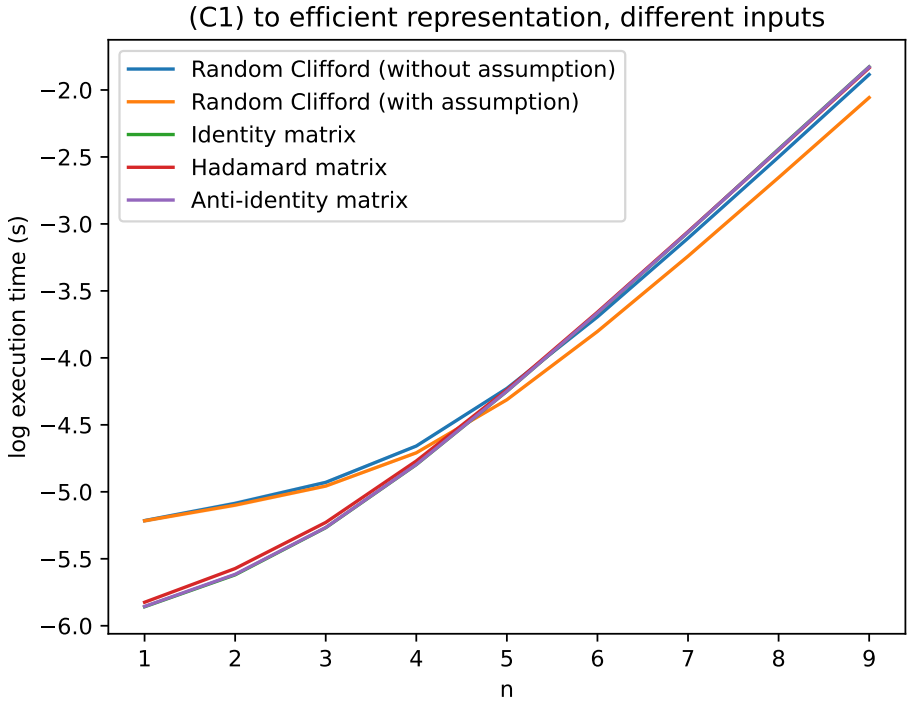


Figure 4.5: Comparison of the execution time of our (C1) → (C2) algorithm, for a variety of inputs. The algorithms were run 1000 times for each qubit number and type of input. For each type of input, we plot (log base 10) the average execution time (solid line), and if the input is inherently random, we additionally shade the 33-66th percentile range (though these are too small to see). We test our algorithm on a random Clifford matrix, both with and without a guarantee that on the input, the identity matrix $\mathbb{1}$, the anti-identity matrix, and the n-fold tensor product of the Hadamard matrix (which we label "Hadamard"). For the latter three inputs, we do not assume that the input is a valid stabiliser state.

matrix that differs from a Clifford in a single entry, or has an incorrect relative phase between columns. we call such a matrix an "almost" Clifford matrix. We benchmark our algorithm by sampling from "almost" Clifford matrices - we sample a random Clifford matrix uniformly at random, then with equal probability, we modify a single entry uniformly at random or change the relative phase of a uniformly random column, such that the resulting matrix is no longer a Clifford. The execution times for our algorithm, as well as `stim`'s and `Qiskit`'s are plotted in figure 4.6. As before, we see that our algorithm is significantly faster than `stim` and `Qiskit` on a random Clifford matrix input, and `Qiskit` outperforms `stim` for large qubit numbers. Our algorithm has similar performance on "almost" Clifford matrices to actual Clifford matrices, despite the early termination of the algorithm. This is due to the large initial overhead in finding the U_i, which is not usually terminated early by an "almost" Clifford input. `stim` shows similar behaviour, albeit significantly slower than our algorithm. `Qiskit` is slightly faster on an "almost" Clifford, since this is usually a best-case scenario - with high probability the "almost" Clifford M will not conjugate Z_1 to a Pauli, and thus can be rejected after a single matrix multiplication $MZ_1M^\dagger$. We emphasise that this single matrix multiplication is theoretically asymptotically slower than our entire algorithm, but is highly optimised in practice.

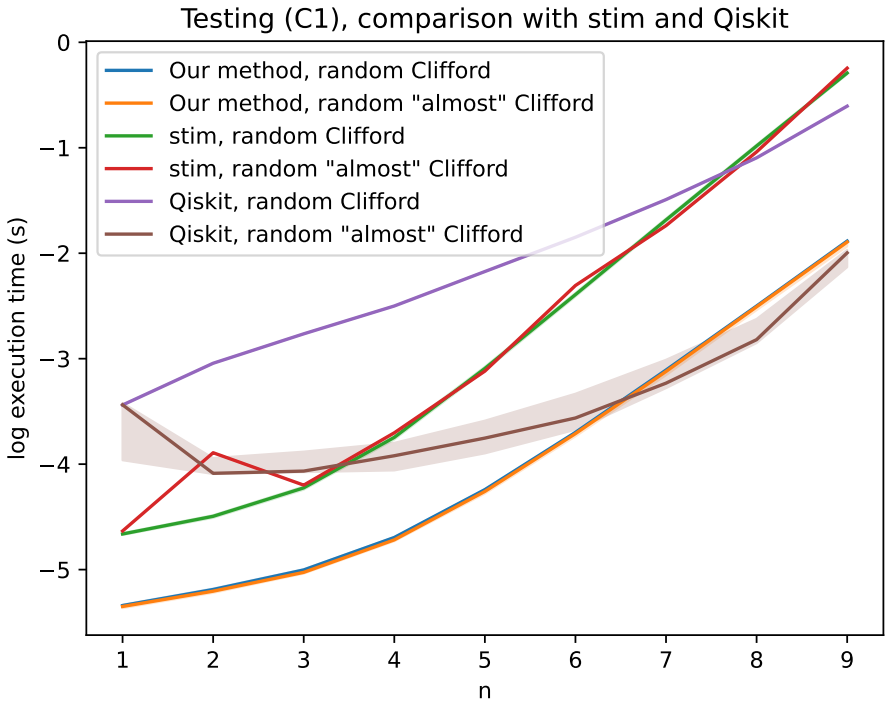


Figure 4.6: Comparison of our algorithm for testing whether a matrix is a Clifford with `stim`'s and `Qiskit`'s. The algorithms were run 1000 times for each qubit number and type of input. For each type of input input, we plot (log base 10) the average execution time (solid line), and shade the 33-66th percentile range. We test the algorithms with a uniformly random Clifford matrix, and a uniformly random "almost" Clifford matrix (a matrix that differs from a Clifford state in a single entry, or by the relative phase of two columns).

Bibliography

[1] Wilfred Salmon, Sergii Strelchuk, and David Arvidsson-Shukur. "Only classical parameterised states have optimal measurements under least squares loss". In: *Quantum* 7 (2023), p. 998.

[2] Flavio Salvati, Wilfred Salmon, Crispin HW Barnes, and David RM Arvidsson-Shukur. "Compression of metrological quantum information in the presence of noise". In: *arXiv preprint arXiv:2307.08648* (2023).

[3] Wilfred Salmon, Sergii Strelchuk, and David Arvidsson-Shukur. "James-Stein Estimation in Gaussian Metrology: V2". In: *arXiv preprint arXiv:2404.02203* (2024).

[4] Wilfred Salmon, Sergii Strelchuk, and Tom Gur. "Provable advantage in quantum PAC learning". In: *arXiv preprint arXiv:2309.10887* (2023).

[5] Nadish de Silva, Wilfred Salmon, and Ming Yin. "Fast algorithms for classical specifications of stabiliser states and Clifford gates". In: *arXiv preprint arXiv:2311.10357* (2023).

[6] Werner Heisenberg. "Physics and beyond, New York (Harper & Row) 1971." In: (1971).

[7] Maximilian Schlosshauer. "Decoherence, the measurement problem, and interpretations of quantum mechanics". In: *Reviews of Modern physics* 76.4 (2004), pp. 1267–1305.

[8] David Hanneke, S Fogwell, and Gerald Gabrielse. "New measurement of the electron magnetic moment and the fine structure constant". In: *Physical review letters* 100.12 (2008), p. 120801.

[9] Gordon Moore. "Cramming more components onto integrated circuits (1965)". In: (2021).

[10] Mohsen Razavy. *Quantum theory of tunneling*. World Scientific, 2013.

[11] Richard P Feynman. "Simulating physics with computers". In: *Feynman and computation*. CRC Press, 2018, pp. 133–153.

[12] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[13] Solomon Kullback and Richard A Leibler. "On information and sufficiency". In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86.

[14] Rajeev Motwani and Prabhakar Raghavan. *Randomized algorithms*. Cambridge university press, 1995.

[15] Stephen M Stigler. "The epic story of maximum likelihood". In: *Statistical Science* (2007), pp. 598–620.

[16] Charles Stein. "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution". In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Vol. 3. University of California Press. 1956, pp. 197–207.

[17] William James and Charles Stein. "Estimation with quadratic loss". In: *Breakthroughs in statistics: Foundations and basic theory*. Springer, 1992, pp. 443–460.

[18] Tze Fen Li and Dinesh S Bhoj. "A modified James-Stein estimator with application to multiple regression analysis". In: *Scandinavian journal of statistics* (1988), pp. 33–37.

[19] Adam La Caze. "Frequentism". In: (2016).

[20] Dennis Victor Lindley. *Bayesian statistics: A review*. SIAM, 1972.

[21] Robert Leslie Ellis. "On the foundations of the theory of probabilities". In: *(No Title)* (1843).

[22] Stephen M Stigler. "The history of statistics, Chapter 3". In: (1986).

[23] James V Stone. "Bayes' rule: a tutorial introduction to Bayesian analysis". In: (2013).

[24] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[25] Aad W Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000.

[26] Whitney K Newey and Daniel McFadden. "Large sample estimation and hypothesis testing". In: *Handbook of econometrics* 4 (1994), pp. 2111–2245.

[27] Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. "Advances in quantum metrology". In: *Nature photonics* 5.4 (2011), pp. 222–229.

[28] Carl W Helstrom. "Quantum detection and estimation theory". In: *Journal of Statistical Physics* 1 (1969), pp. 231–252.

[29] Rafał Demkowicz-Dobrzański, Wojciech Górecki, and Mădălin Guţă. "Multi-parameter estimation beyond quantum Fisher information". In: *Journal of Physics A: Mathematical and Theoretical* 53.36 (2020), p. 363001.

[30] Johannes Jakob Meyer, Sumeet Khatri, Daniel Stilck França, Jens Eisert, and Philippe Faist. "Quantum metrology in the finite-sample regime". In: *arXiv preprint arXiv:2307.06370* (2023).

[31] Samuel L Braunstein and Carlton M Caves. "Statistical distance and the geometry of quantum states". In: *Physical Review Letters* 72.22 (1994), p. 3439.

[32] Francesco Albarelli, Marco Barbieri, Marco G Genoni, and Ilaria Gianani. "A perspective on multiparameter quantum metrology: From theoretical tools to applications in quantum imaging". In: *Physics Letters A* 384.12 (2020), p. 126311.

[33] Carl W Helstrom. "Minimum mean-squared error of estimates in quantum statistics". In: *Physics letters A* 25.2 (1967), pp. 101–102.

[34] Horace Yuen and Melvin Lax. "Multiple-parameter quantum estimation and measurement of nonselfadjoint observables". In: *IEEE Transactions on Information Theory* 19.6 (1973), pp. 740–750.

[35] Vyacheslav P Belavkin. "Generalized uncertainty relations and efficient measurements in quantum systems". In: *arXiv preprint quant-ph/0412030* (2004).

[36] Hiroshi Nagaoka. "A new approach to Cramér-Rao bounds for quantum state estimation". In: *Asymptotic Theory of Quantum Statistical Inference: Selected Papers*. 2005, pp. 100–112.

[37] Keiji Matsumoto. "A new approach to the Cramér-Rao-type bound of the pure-state model". In: *Journal of Physics A: Mathematical and General* 35.13 (2002), p. 3111.

[38] Masahito Hayashi. "Comparison between the Cramer-Rao and the mini-max approaches in quantum channel estimation". In: *Communications in mathematical physics* 304.3 (2011), pp. 689–709.

[39] Akio Fujiwara. "Strong consistency and asymptotic efficiency for adaptive quantum estimation problems". In: *JOURNAL OF PHYSICS-LONDON-A MATHEMATICAL AND GENERAL* 39.40 (2006), p. 12489.

[40] Yuxiang Yang, Giulio Chiribella, and Masahito Hayashi. "Attaining the ultimate precision limit in quantum state estimation". In: *Communications in Mathematical Physics* 368 (2019), pp. 223–293.

[41] Zihao Gong and Boulat A Bash. "Two-stage Quantum Estimation and the Asymptotics of Quantum-enhanced Transmittance Sensing". In: *arXiv preprint arXiv:2402.17922* (2024).

[42] Koichi Yamagata, Akio Fujiwara, and Richard D Gill. "Quantum local asymptotic normality based on a new quantum likelihood ratio". In: (2013).

[43] Noah Lupu-Gladstein, Y Batuhan Yilmaz, David RM Arvidsson-Shukur, Aharon Brodutch, Arthur OT Pang, Aephraim M Steinberg, and Nicole Yunger Halpern. "Negative quasiprobabilities enhance phase estimation in quantum-optics experiment". In: *Physical Review Letters* 128.22 (2022), p. 220504.

[44] Hiwa Mahmoudi, Michael Hofbauer, Bernhard Goll, and Horst Zimmermann. "Noise and breakdown characterization of SPAD detectors with time-gated photon-counting operation". In: *Sensors* 21.16 (2021), p. 5287.

[45] Christophe Couteau, Stefanie Barz, Thomas Durt, Thomas Gerrits, Jan Huwer, Robert Prevedel, John Rarity, Andrew Shields, and Gregor Weihs. "Applications of single photons to quantum communication and computing". In: *Nature Reviews Physics* 5.6 (2023), pp. 326–338.

[46] David RM Arvidsson-Shukur, Nicole Yunger Halpern, Hugo V Lepage, Aleksander A Lasek, Crispin HW Barnes, and Seth Lloyd. "Quantum advantage in postselected metrology". In: *Nature communications* 11.1 (2020), p. 3775.

[47] Joe H Jenne and David RM Arvidsson-Shukur. "Unbounded and lossless compression of multiparameter quantum information". In: *Physical Review A* 106.4 (2022), p. 042404.

[48] Sahar Alipour and Ali T Rezakhani. "Extended convexity of quantum fisher information in quantum metrology". In: *Physical Review A* 91.4 (2015), p. 042104.

[49] Milán Mosonyi and Tomohiro Ogawa. "Quantum hypothesis testing and the operational interpretation of the quantum Rényi relative entropies". In: *Communications in Mathematical Physics* 334 (2015), pp. 1617–1648.

[50] Bruce McK. "On the admissible estimators for certain fixed sample binomial problems". In: *The Annals of Mathematical Statistics* (1971), pp. 1579–1587.

[51] Lawrence D Brown. "Admissible estimators, recurrent diffusions, and insoluble boundary value problems". In: *The Annals of Mathematical Statistics* 42.3 (1971), pp. 855–903.

[52] Samuel Karlin. "Admissibility for estimation with quadratic loss". In: *The Annals of Mathematical Statistics* 29.2 (1958), pp. 406–436.

[53] Witold Klonecki and Stefan Zontek. "On the structure of admissible linear estimators". In: *Journal of multivariate analysis* 24.1 (1988), pp. 11–30.

[54] Arindam Banerjee, Xin Guo, and Hui Wang. "On the optimality of conditional expectation as a Bregman predictor". In: *IEEE Transactions on Information Theory* 51.7 (2005), pp. 2664–2669.

[55] David RM Arvidsson-Shukur, William F Braasch Jr, Stephan De Bievre, Justin Dressel, Andrew N Jordan, Christopher Langrenez, Matteo Lostaglio, Jeff S Lundeen, and Nicole Yunger Halpern. "Properties and Applications of the Kirkwood-Dirac Distribution". In: *arXiv preprint arXiv:2403.18899* (2024).

[56] Walter Rudin. "Principles of mathematical analysis". In: (2021).

[57] Jing Yang. "Pure State Inspired Lossless Post-selected Quantum Metrology of Mixed States". In: *arXiv preprint arXiv:2405.00405* (2024).

[58] Konstantinos I Roumeliotis and Nikolaos D Tselikas. "Chatgpt and open-ai models: A preliminary review". In: *Future Internet* 15.6 (2023), p. 192.

[59] Chaoyang Li, Xiaohan Li, Manni Chen, and Xinyao Sun. "Deep learning and image recognition". In: *2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT)*. IEEE. 2023, pp. 557–562.

[60] Iqbal H Sarker, Md Hasan Furhad, and Raza Nowrozy. "Ai-driven cybersecurity: an overview, security intelligence modeling and research directions". In: *SN Computer Science* 2.3 (2021), p. 173.

[61] Mohammed Yousef Shaheen. "Applications of Artificial Intelligence (AI) in healthcare: A review". In: *ScienceOpen Preprints* (2021).

[62] Ewen Callaway. "What's next for the AI protein-folding revolution". In: *Nature* 604 (2022), pp. 234–238.

[63] Leslie G Valiant. "A theory of the learnable". In: *Communications of the ACM* 27.11 (1984), pp. 1134–1142.

[64] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[65] Michael J Kearns and Umesh Vazirani. *An introduction to computational learning theory*. MIT press, 1994.

[66] Vladimir N Vapnik and A Ya Chervonenkis. "On the uniform convergence of relative frequencies of events to their probabilities". In: *Measures of complexity: festschrift for alexey chervonenkis*. Springer, 2015, pp. 11–30.

[67] Richard P Anstee, Lajos Rónyai, and Attila Sali. "Shattering news". In: *Graphs and Combinatorics* 18 (2002), pp. 59–73.

[68] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[69] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. "Learnability and the Vapnik-Chervonenkis dimension". In: *Journal of the ACM (JACM)* 36.4 (1989), pp. 929–965.

[70] Steve Hanneke. “The optimal sample complexity of PAC learning”. In: *Journal of Machine Learning Research* 17.38 (2016), pp. 1–15.

[71] Vladimir Vapnik and Alexey Chervonenkis. “Theory of pattern recognition”. In: (1974).

[72] Michel Talagrand. “Sharper bounds for Gaussian and empirical processes”. In: *The Annals of Probability* (1994), pp. 28–76.

[73] Shai Ben-David and Ruth Urner. “The sample complexity of agnostic learning under deterministic labels”. In: *Conference on Learning Theory*. PMLR. 2014, pp. 527–542.

[74] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. “An introduction to quantum machine learning”. In: *Contemporary Physics* 56.2 (2015), pp. 172–185.

[75] Nader H Bshouty and Jeffrey C Jackson. “Learning DNF over the uniform distribution using a quantum example oracle”. In: *Proceedings of the eighth annual conference on Computational learning theory*. 1995, pp. 118–127.

[76] Ethan Bernstein and Umesh Vazirani. “Quantum complexity theory”. In: *Proceedings of the twenty-fifth annual ACM symposium on Theory of computing*. 1993, pp. 11–20.

[77] Alp Atıcı and Rocco A Servedio. “Quantum algorithms for learning and testing juntas”. In: *Quantum Information Processing* 6.5 (2007), pp. 323–348.

[78] SAM SPIRO. “FOURIER ANALYSIS OF BOOLEAN FUNCTIONS”. In: ().

[79] Rocco A Servedio and Steven J Gortler. “Equivalences and separations between quantum and classical learnability”. In: *SIAM Journal on Computing* 33.5 (2004), pp. 1067–1092.

[80] Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme. “A rigorous and robust quantum speed-up in supervised machine learning”. In: *Nature Physics* 17.9 (2021), pp. 1013–1017.

[81] Ryan Sweke, Jean-Pierre Seifert, Dominik Hangleiter, and Jens Eisert. “On the quantum versus classical learnability of discrete distributions”. In: *Quantum* 5 (2021), p. 417.

[82] Scott Aaronson. “Read the fine print”. In: *Nature Physics* 11.4 (2015), pp. 291–293.

[83] Srinivasan Arunachalam and Ronald de Wolf. “Guest column: A survey of quantum learning theory”. In: *ACM Sigact News* 48.2 (2017), pp. 41–67.

[84] Srinivasan Arunachalam and Ronald De Wolf. “Optimal quantum sample complexity of learning algorithms”. In: *Journal of Machine Learning Research* 19.71 (2018), pp. 1–36.

[85] Joran van Apeldoorn, Arjan Cornelissen, András Gilyén, and Giacomo Nannicini. “Quantum tomography using state-preparation unitaries”. In: *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM. 2023, pp. 1265–1318.

[86] Jeongwan Haah, Robin Kothari, Ryan O’Donnell, and Ewin Tang. “Query-optimal estimation of unitary channels in diamond distance”. In: *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2023, pp. 363–390.

[87] Robin Kothari and Ryan O’Donnell. “Mean estimation when you have the source code; or, quantum Monte Carlo methods”. In: *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM. 2023, pp. 1186–1215.

[88] Leonard Pitt and Leslie G Valiant. “Computational limitations on learning from examples”. In: *Journal of the ACM (JACM)* 35.4 (1988), pp. 965–984.

[89] Dana Angluin and Michael Kharitonov. "When won't membership queries help?" In: *Proceedings of the twenty-third annual ACM symposium on Theory of computing*. 1991, pp. 444–454.

[90] Michael A Nielsen and Isaac L Chuang. *Quantum computation and quantum information*. Vol. 2. Cambridge university press Cambridge, 2001.

[91] 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.

[92] Michel Boyer, Gilles Brassard, Peter Høyer, and Alain Tapp. "Tight bounds on quantum searching". In: *Fortschritte der Physik: Progress of Physics* 46.4-5 (1998), pp. 493–505.

[93] Grzegorz Gluch and Ruediger Urbanke. "Exponential separation between two learning models and adversarial robustness". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 20785–20797.

[94] Andrés Masegosa, Stephan Lorenzen, Christian Igel, and Yevgeny Seldin. "Second order PAC-Bayesian bounds for the weighted majority vote". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 5263–5273.

[95] Wassily Hoeffding. "Probability inequalities for sums of bounded random variables". In: *The collected works of Wassily Hoeffding* (1994), pp. 409–426.

[96] Ashley Montanaro. "Quantum speedup of Monte Carlo methods". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 471.2181 (2015), p. 20150301.

[97] Artur J Ferreira and Mário AT Figueiredo. "Boosting algorithms: A review of methods, theory, and applications". In: *Ensemble machine learning: Methods and applications* (2012), pp. 35–85.

[98] Yoav Freund and Robert E Schapire. "A desicion-theoretic generalization of on-line learning and an application to boosting". In: *European conference on computational learning theory*. Springer. 1995, pp. 23–37.

[99] Rocco A Servedio. "Smooth boosting and learning with malicious noise". In: *The Journal of Machine Learning Research* 4 (2003), pp. 633–648.

[100] Adam Izdebski and Ronald de Wolf. "Improved quantum boosting". In: *arXiv preprint arXiv:2009.08360* (2020).

[101] Srinivasan Arunachalam and Reevu Maity. "Quantum boosting". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 377–387.

[102] John M Martyn, Zane M Rossi, Andrew K Tan, and Isaac L Chuang. "Grand unification of quantum algorithms". In: *PRX quantum* 2.4 (2021), p. 040203.

[103] Daniel Gottesman. "The Heisenberg representation of quantum computers". In: *arXiv preprint quant-ph/9807006* (1998).

[104] Scott Aaronson and Daniel Gottesman. "Improved simulation of stabilizer circuits". In: *Physical Review A—Atomic, Molecular, and Optical Physics* 70.5 (2004), p. 052328.

[105] Richard Jozsa and Maarten Van den Nest. "Classical simulation complexity of extended Clifford circuits". In: *arXiv preprint arXiv:1305.6190* (2013).

[106] Joschka Roffe. "Quantum error correction: an introductory guide". In: *Contemporary Physics* 60.3 (2019), pp. 226–245.

[107] Sergey Bravyi and Jeongwan Haah. "Magic-state distillation with low overhead". In: *Physical Review A—Atomic, Molecular, and Optical Physics* 86.5 (2012), p. 052329.

[108] David Poulin. "Stabilizer formalism for operator quantum error correction". In: *Physical review letters* 95.23 (2005), p. 230504.

[109] Michael Zurel. "Classical descriptions of quantum computations: foundations of quantum computation via hidden variable models, quasiprobability representations, and classical simulation algorithms". PhD thesis. University of British Columbia, 2024.

[110] Nadish De Silva, Ming Yin, and Sergii Strelchuk. "Bases for optimising stabiliser decompositions of quantum states". In: *Quantum Science and Technology* (2023).

[111] Shane Mansfield. Personal communication. 2023.

[112] Scott Aaronson and Alex Arkhipov. "The computational complexity of linear optics". In: *Proceedings of the forty-third annual ACM symposium on Theory of computing*. 2011, pp. 333–342.

[113] Vadym Kliuchnikov and Sebastian Schonnenbeck. "Stabilizer operators and Barnes-Wall lattices". In: *arXiv preprint arXiv:2404.17677* (2024).

[114] Qiskit contributors. *Qiskit: An Open-source Framework for Quantum Computing*. 2023.

[115] Craig Gidney. "Stim: a fast stabilizer circuit simulator". In: *Quantum* 5 (July 2021), p. 497.

[116] Sergey Bravyi and Dmitri Maslov. "Hadamard-free circuits expose the structure of the Clifford group". In: *IEEE Transactions on Information Theory* 67.7 (2021), pp. 4546–4563.

[117] James R. Bitner, Gideon Ehrlich, and Edward M. Reingold. "Efficient generation of the binary reflected Gray code and its applications". In: *Commun. ACM* 19.9 (Sept. 1976), pp. 517–521.