Some Theory and Applications of Probability in Quantum Mechanics

by

Christopher Ferrie

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Applied Mathematics

Waterloo, Ontario, Canada, 2012

Abstract

This thesis investigates three distinct facets of the theory of quantum information. The first two, quantum state estimation and quantum process estimation, are closely related and deal with the question of how to estimate the classical parameters in a quantum mechanical model. The third attempts to bring quantum theory as close as possible to classical theory through the formalism of quasi-probability.

Building a large scale quantum information processor is a significant challenge. First, we require an accurate characterization of the dynamics experienced by the device to allow for the application of error correcting codes and other tools for implementing useful quantum algorithms. The necessary scaling of computational resources needed to characterize a quantum system as a function of the number of subsystems is by now a well studied problem (the scaling is generally exponential). However, irrespective of the computational resources necessary to just write-down a classical description of a quantum state, we can ask about the experimental resources necessary to obtain data (measurement complexity) and the computational resources necessary to generate such a characterization (estimation complexity). These problems are studied here and approached from two directions.

The first problem we address is that of quantum state estimation. We apply high-level decision theoretic principles (applied in classical problems such as, for example, universal data compression) to the estimation of a qubit state. We prove that quantum states are more difficult to estimate than their classical counterparts by finding optimal estimation strategies. These strategies, requiring the solution to a difficult optimization problem, are difficult to implement in practise. Fortunately, we find estimation algorithms which come close to optimal but require far fewer resources to compute. Finally, we provide a classical analog of this quantum mechanical problem which reproduces, and gives intuitive explanations for, many of its features, such as why adaptive tomography can quadratically reduce its difficulty.

The second method for practical characterization of quantum devices takes is applied to the problem of quantum process estimation. This differs from the above analysis in two ways: (1) we apply strong restrictions on knowledge of various estimation and control parameters (the former making the problem easier, the latter making it harder); and (2) we consider the problem of designing future experiments based on the outcomes of past experiments. We show in test cases that adaptive protocols can exponentially outperform their off-line counterparts. Moreover, we adapt machine learning algorithms to the problem which bring these experimental design methodologies to realm of experimental feasibility.

In the final chapter we move away from estimation problems to show formally that a classical representation of quantum theory is not tenable. This intuitive conclusion is formally borne out through the connection to *quasi-probability* – where it is equivalent to the necessity of negative probability in all such representations of quantum theory. In particular, we generalize previous *no-go* theorems to arbitrary classical representations of quantum systems of arbitrary dimension. We also discuss recent progress in the program to identify quantum *resources* for subtheories of quantum theory and operational restrictions motivated by quantum computation.

This thesis is based on the following publications:

- Ferrie, C., Granade, C.E. and Cory, D.G. (2012) Quantum Information Processing, Online First. [72]
- Ferrie, C. and Blume-Kohout, R. (2012) AIP Conference Proceedings 1443, 14. [68]
- Ferrie, C., Granade, C. and Cory, D.G. (2012) AIP Conference Proceedings 1443, 165. [71]
- Ferrie, C. (2011) Reports on Progress in Physics 74, 116001. [67]
- Ferrie, C., Morris, R. and Emerson J. (2010) Physical Review A 82, 044103. [73]

Portions of the appendix are based on work lead by Victor Veitch (and contained in reference [215]) as well as unpublished work with Victor Veitch and Joseph Emerson.

The latter portions of Chapter 2 are based on completed work (with a manuscript in preparation) with Robin Blume-Kohout. Latter portions of Chapter 3 are based on completed work (now, with a near completed draft in the final stages of preparation) with Chris Granade, Nathan Wiebe and David Cory.

Acknowledgements

First of all I thank my advisor and friend, Joseph Emerson, for his intellectual and financial support over the years. I am grateful for the freedom he has given me but also for making sure I didn't stray too far off the beaten path. Thanks goes as well to my other mentors: David Cory for his patience and encouragement and Robin Blume-Kohout for his guidance. I thank also my senior colleagues at the Perimeter Institute for much advice: Chris Fuchs and Rob Spekkens. I feel extremely fortunate to have met my good friend and colleague Chris Granade. His knowledge and ability to instantly understand and formalize my half-baked ideas will always amaze me.

I also thank my friends at the Institute for Quantum Computing that I've had the pleasure of sharing the past few years with: Victor Veitch, Ian Hincks, Osama Moussa, Cozmin Ududec, Easwar Magesan and Nathan Wiebe. Thanks goes as well to my conference-friend Josh Combes for many helpful discussions.

I am indebted to Helen Warren in the Applied Mathematics Department and the administrative staff the the Institute for Quantum Computing for their assistance in the many administrative challenges I've faced over the years.

Financial support for my doctoral studies has been provided by the government of Canada through NSERC, the University of Waterloo, the Institute for Quantum Computing and the Waterloo Institute for Nanotechnology.

Finally, I thank my mom and dad for 30 years (and counting!) of love and support.

Dedication

To Lindsay, Dylan and Max.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

After all our physical needs are met – food, water, shelter, internet – our base intangible need is predictive information. We ask “why” because when we know the “reason” for something, we can (or, at least, we think we can) predict when it will happen again or perhaps even control it. For the last hundred years we have known that our physical world has promised more information than we have been able to ascertain. This theoretical limit of information we can have about a physical system is its quantum state and it is only recently that we have been able to exact enough experimental precision to extract this information. The problem is that the extraction process is a quantum measurement and does not have output of the form “here is the quantum state of your physical system”. The output is a string of seemingly random bits which we must use to estimate the quantum state. That is problem studied in Chapter 2. Namely, given a set of data, how and what should we conclude about the physical system that produced it.

A quantum state is deceptively similar to a classical probability distribution for coins or dice; it give the probabilities for the outcomes of future “tosses”, or measurements. Thus we might expect that the classical theory of estimation [56] will be of use – and it is. However, its use extends only so far and new techniques are needed. For this we propose a better classical analogue which reflects the key conceptual difference between classical and quantum measurements – classical measurement reveals the state of the system (the coin lands “heads”, for example). A quantum measurement, on the other hand, produces outcomes which are intrinsically noisy. We call this sampling *mismatch* and propose the *noisy coin* as a classical system with the same feature. The noisy coin is a tossed coin whose outcome we do not get to directly see – but we are told the correct outcome with some known probability. In other words, some fixed portion of the time, the outcome is reported incorrectly. If we still desire to know the true bias of coin, we have an analogy with quantum measurement (at least for the purpose of estimation).

There are many methods for the estimation of quantum states (also known as *quantum state tomography* [166]) which include, for example, linear inversion, maximum likelihood, and Bayesian mean. But none of them is clearly "the most accurate" for data of finite size N. Even the upper limits on accuracy are as yet unknown, which makes it difficult to say that a given method is "accurate enough". We address this problem here by (i) calculating the minimum achievable error for single-qubit tomography with N Pauli measurements, (ii) finding *minimax* estimators that achieve this bound, and (iii) comparing the performance of known estimators.

In our decision theoretic approach [18], estimators can be ranked by their *worst-case* risk – the maximum, over all ρ, of the expected error. The best-performing estimator by this metric is called the *minimax* estimator. Different error metrics (fidelity, trace-norm, and so on) yield different minimax estimators; here we focus on *relative entropy* [189, 213] error (the canonical choice in classical predictive estimation and machine learning [133]). The minimax estimators for quantum tomography are impractical to calculate but they serve as a critical *benchmark*: a tomography algorithm is "good enough" inasmuch as its risk is close to that of the minimax estimator.

In Chapter 2, we construct minimax estimators for reconstructing single-qubit states from N measurements of the Pauli operators (σ_x, σ_y, σ_z); use them to get absolute lower bounds on achievable risk; and find that risk scales as $N^{-1/2}$ (for classical probabilities, risk scales as N^{-1}). We reproduced most features of quantum tomography, including $N^{-1/2}$ risk scaling, with our simpler "noisy coin" model.

Quantum mechanics gives the most accurate description of many physical systems of interest. In turn, the most accurate characterization of a quantum device is given by its quantum mechanical model. Thus, efficient methods for the honest estimation of the distribution of parameters in a quantum mechanical model are of utmost importance, not only for building robust quantum technologies, but to reach new regimes of physics. The quantum state is classical description of a preparation procedure producing physical systems. We can perform quantum mechanical measurements on these physical systems to ascertain a useful description of our preparation devices. However, it is likely the case that something, intentional or not, *will* happen to the physical system between its preparation and measurement and *almost certainly* the case that something intentional *must* happen it if we are to utilize the physical system for technological gain. Thus, we must also characterize how the state of the physical system *transforms* by our actions upon it.

Whereas we approached the estimation of quantum states in Chapter 2 from a global perspective (a more top-down approach), in Chapter 3 we tackle the estimation of parameters in quantum dynamical processes from the perspective of precise prior knowledge (a more bottom-up approach). This makes the problem somewhat less difficult. On the other hand, we add two additional challenges: (1) we consider *adaptively* designing experiments and (2) we consider experimentally restricted control and measurement paradigms. The former is not necessary of course – but we will find that it drastically improves the accuracy with which we can determine unknown parameters. The latter, however, is almost tautologically necessary; the parameters necessary for precise control are the very ones we are trying to estimate!

In Chapter 3 we apply Bayesian statistical principles to the problem of characterizing quantum devices. We do so for sample problems of varying complexity. For models that adhere to certain information theoretic assumptions, we can derive protocols which are provably optimal by also obtaining analytic expressions, and lower bounds, on the accuracy of generic protocols. This is the ideal situation because it is important to know how well current and prospective strategies do with respect to what is optimal. Moreover, we found that the absolute best strategies require the solutions of complicated optimization problems and will therefore be infeasible to implement in practice. However, we derive intuition from those solutions to motivate practically implementable strategies which come close to optimal but are vastly simpler to perform. That is, we derived heuristics which offer an extreme boost in efficiency at a small cost in accuracy. Indeed, we show that the asymptotic scaling of our protocol is equivalent to the optimal solution.

Although we illustrated the utility of our procedure for a specific estimation problem, the methodology applies more generally. However, it does not extend to fully generic estimation problems. For that we look to classical machine learning methods [62] and provide a proof of principle that they will produce robust protocols which will significantly improve the accuracy and efficiency with which we can obtain optimal, reliable estimates of parameters in dynamical models of quantum systems.

In the final chapter we explore a different facet of quantum information theory which overlaps with the foundations of quantum mechanics. Since the advent of quantum theory, identifying the quintessential feature of the theory has remained an open problem. That is, we have yet to answer the long standing question of specifying the conditions under which a given physical process is "truly quantum". In Chapter 4, we explore this question in the context of one of the oldest notions of "quantumness" – namely, quasi-probability.

There exists an enormous variety of ways to map quantum theory to a classical probabilistic framework [67]. However, building on previous work, we show it is impossible for the resulting theory to be a proper classical theory. In other words, "negative probability" must emerge somewhere. This conclusion was suspected for a very long time and many attempts have been made in the past, but have failed since additional assumptions were needed or tacitly applied. This results has implications primarily in the field of quantum foundations. It should also find use wherever quasi-probability representations are employed (notably: quantum optics [188], quantum chaos [207], and decoherence theory [122]) as it clarifies the connection between the various definitions of quasi-probability representations. We also comment on recent progress identifying *subtheories* or operationally restricted sets of quantum operations which allow a classical representation [215, 219].

Chapter 2

Quantum State Estimation

We will call *state estimation* the general situation surrounding the following question: how should one estimate the parameters (generally, probabilitic) in an assumed (generally, probabilistic) model based on obtained data? Classical state estimation is sub-discipline of decision theory [18] and, more broadly, information theory [56]. The phrase "quantum state estimation" may evoke two connotations. The first is the now usual sense in which all the elements of a classical theory are replaced with their "quantum analogue". For example, in some quantum computing models, we replace bits with quantum bits (qubits) and gates with quantum gates and so on. This is not the interpretation we are interested in. We are interested in obtaining a classical description of a quantum state [166]. This is a means to an end; for at some point in the future this procedure will not be necessary. However, since scientists presently communicate with classical information, this task is a daily requirement. And thus we quickly come to the urgent question: given experimental data, what is the best procedure for estimating the classical description of the quantum state?

A quantum theoretical model of a physical system is easy enough to state and understand provided we have some knowledge of the following linear algebra terms which unfortunately must be assumed of the reader: inner product space, Hermitian transpose, positive semi-definite matrix, identity matrix and trace. The model is as follows:

1. To each physical system, we assign an inner product space $\mathcal{H}$ of dimension $d < \infty$.
2. Each preparation procedure is represented by a positive semi-definite matrix ρ with $\mathrm{Tr}(\rho) = 1$.

3. Each measurement procedure is represented by set of positive semi-definite matrices $\{E_k\}$ where each E_k represents an outcome of the measurement and the set satisfies $\sum_k E_k = \mathbb{1}$.

4. The probability of outcome k given preparation ρ is

$$\Pr(E_k|\rho) = \mathrm{Tr}(E_k\rho). \tag{2.1}$$

Quantum state estimation[1], or *tomography*, proceeds in two steps: (1) measuring identically prepared systems in different bases to collect data D; and (2) approximating the state ρ by plugging the data into an estimator $\hat{\rho}(D)$ of the preparation procedure. Clearly, tomography is of interest to experimentalists who need to know what state they are preparing. But, there is a recent growing interest in the theoretical aspects of the problem. What we have been doing so far has "worked" for the for small dimensional systems. But it is not clear if, or how well, it will work for larger, more complex, systems.

Before we define what "best" means, let us state the basic problem more formally. Suppose the same preparation procedure is performed N times, and each outcome k of the measurement $\{E_k\}$ occurs n_k times. The problem can now be concisely stated as follows: infer the state ρ from the experimental data $\{n_k\}$. Note that even if we assume, which we do, there is some "true" state ρ, this is an impossible task for finite N. Thus, we will need to find some (generalized) metric $d(\rho, \hat{\rho})$ which measures the performance of our estimation procedure $\hat{\rho}$.

For one, the estimator should (must!) be accurate; technically, we demand it have low expected error (or *risk*) for all true states. Some popular estimators (e.g. linear inversion, or maximum likelihood) have no provable accuracy properties for finite number of measurements N. Others (Bayesian mean estimation) are provably optimal only on *average* over a particular ensemble of input states – which isn't particularly helpful, since device states in the laboratory are not selected at random.

Instead, estimators can be ranked by their worst-case risk: the maximum value, over all true states ρ, of the expected error. The best-performing estimator by this metric is called the minimax estimator. Different error metrics (fidelity, trace-norm and so on) yield different minimax estimators. Here we focus on relative entropy error (the canonical choice in classical predictive estimation and machine learning). As in many cases, the minimax estimators for quantum tomography are strange, unwieldy, and impractical for laboratory

[1] A collection of papers on quantum state estimation, each with some historical references is in [166]. There is also a historical review in reference [22].

use. But they serve a crucial purpose as a benchmark: an estimator is “good enough” inasmuch as its risk is close to that of the minimax estimator.

This chapter is structured as follows. The natural, perhaps obvious, first step is to look to the classical theory of estimation. Unfortunately, this does not quite work as we will see in section 2.1. We find a better classical analog of the problem in section 2.2, which we call the *noisy coin*. We show how the formalism of the noisy coin can give insight into the solution of the fully quantum mechanical problem in section 2.3. We discuss the future directions of this research in section 2.4.

Portions of the this chapter is based on the reference [68].

2.1 Why Classical Estimation Fails

One might expect that since a quantum state can be interpreted classically as a calculation tool to predict the probability distribution of the outcomes of future measurements, it might behave the same way when we try to estimate it from past data. Wrong! And, here we will see why.

2.1.1 Classical Estimation

Few things are as ubiquitous in information theory as the Shannon entropy [195] and two people: Alice and Bob. In classical statistics the entropy quantifies the average uncertainty in a random variable. More intuitively, it measures the expected *surprise* an agent, say Bob, should feel upon the outcome of an experiment the possible outcomes for which he has assigned probabilities. For example, suppose Bob tosses a die and he deems the probability of each face to be p_k. The Shannon entropy of a toss of a d-sided die is

$$H(p) = -\sum_{k=1}^{d} p_k \log p_k.$$

The toss has the most capacity to surprise when the die is unbiased (H is maximized at $p_k = 1/d$). If Bob is certain that the die is loaded ($p_1 = 1$ and $p_{k>1} = 0$, say) then Bob expects to be unsurprised by the outcome of the toss and hence the entropy is zero.

The Shannon entropy has operational significance as the amount of bits required to describe the expected outcome of the toss [56]. Suppose Bob tosses the die N times and he is interested only in the frequency with which each outcomes occurred. Then he can store, *code* or *compress*, the outcome of the N tosses with a bit string of expected length approaching $NH(p)$ as $N \to \infty$. Bob achieves this theoretical lower bound, called the *Shannon bound*, by assigning short codes to outcomes he deems more probable. But so far Bob has known or had some confidence in assigning the probability p. But what happens if he is unsure?

Now suppose Bob is in the possession of a die which he wants to know the bias of. Bob is a public relations professional employed by a celebrity or a government administration or an oil company and is hence very busy. Bob decides to hire an "expert", Alice, to tell him what the bias is. Call Alice's opinion of the bias p and her reported bias $\hat{p}$. Bob would clearly like to know p. To ensure an honest report Bob arranges to pay Alice $R(\hat{p}_k)$ dollars if he tosses the coin and face "k" occurs. Alice's honesty is ensured provided the payments are made according to

$$\sum_{k=1}^{d} p_k R(\hat{p}_k) \leq \sum_{k=1}^{d} p_k R(p_k). \tag{2.2}$$

In other words, Alice's *expected* pay is maximized when reporting her honest opinion. It is a theorem of Aczel and Pfanzagl[2] that a function R satisfying equation (2.2) must be of the form $R(x) = C \log x + B$ for some constants $C \geq 0, B$. Equivalently, Alice's expected loss when reporting $\hat{p}$ is

$$C \sum_{k=1}^{d} p_k \log \frac{p_k}{\hat{p}_k},$$

which is, up to the constant C, the definition of a quantity introduced by Kullback and Leibler in 1951 [134] which now bears their names. It is given the symbol $D(p\|\hat{p})$ and also called *relative entropy.*

Now suppose Alice lied on her resume and she is no "dice-bias-determining-expert". Her payoff constrains her to be honest, but how should she generate an opinion to be honest about? The possibilities are endless. On a whim, she decides to toss the die a fixed number N times and if the outcomes are $n := \{n_k\}$, where each n_k denotes the number of times each face occurred, she will estimate the bias as $\hat{p}(n)$. That is, $\hat{p}$ is a rule (a function, or map) taking every possible measurement outcome to an estimate of the bias.

If Alice assumes there is some "true" bias p which she should be reporting, the difference between the amount of money she *could* make and the amount she expects is the weighted

[2] The original reference is [6]. For a proof which does not assume any regularity conditions on R, see [76]. For a generalization, see [3]. Also, [5, 4].

average of $D(p\|\hat{p}(n))$ over possible outcomes of her tosses. This is called the risk and is explicitly given as

$$R(p,\hat{p}) := \sum_n \Pr(n|p) D(p\|\hat{p}(n)),$$

where

$$\Pr(n|p) = N! \prod_k \frac{p_k^{n_k}}{n_k!}$$

is the distribution of outcomes. When considered as a function of p, it is called the *likelihood function.*

Suppose Bob uses Alice's estimate to store sequences of coin tosses. In order to do so efficiently he attempts to achieve the Shannon bound using an appropriate code based on Alice's reported $\hat{p}$. Then the risk gives the *redundancy* of his code: the expected difference of the actual code length from the theoretical minimum.

The risk can written equivalently in the more compact notation

$$R(p,\hat{p}) = \mathbb{E}_n[D(p\|\hat{p}(n))].$$

with the understanding that $\mathbb{E}_n[\cdot]$ is the average with respect to the probability distribution of n. Supposing Alice has a prior probability distribution $\pi(p)dp$ in mind at the time of reporting $\hat{p}$, her expected loss is

$$r(\pi,\hat{p}) := \int R(p,\hat{p})\pi(p)dp = \mathbb{E}_p[R(p,\hat{p})],$$

which is called the *Bayes risk* of $\hat{p}$.

Amongst the infinitely many possible estimators are a few special ones each having its own justification. Here we define the minimax and Bayes estimators. Loosely speaking, the minimax estimator is the one which does best in the worst case; the Bayes estimator is the one which does best on average with respect to a single prior.

A minimax estimator $\hat{p}_{\text{minimax}}$ is

$$\hat{p}_{\text{minimax}} := \underset{\hat{p}}{\operatorname{argmin}} \max_p R(p,\hat{p}).$$

This is the estimator which has the best worst case behaviour regardless of the probability of such an event. A Bayes estimator is

$$\hat{p}_\pi := \underset{\hat{p}}{\operatorname{argmin}}\, r(\pi,\hat{p}).$$

This is the estimator which has the best expected performance with respect to a given prior π. The Bayes risk *of the prior* π is the expected risk under π when using its Bayes estimator: $r(\pi) := r(\pi, \hat{p}_\pi)$. That is, it is the expected risk when choosing the best estimate with respect to the given prior π. A prior satisfying

$$\pi_{\text{LFP}} := \operatorname*{argmax}_{\pi} r(\pi),$$

is called a *least favorable prior.* If

$$r(\pi) = \max_{p} R(p, \hat{p}_\pi),$$

then $\hat{p}_\pi$ is minimax and π is least favorable. For the case of relative entropy, the Bayes estimators $\hat{p}_\pi$ are unique and hence the unique minimax estimators as well. We call this result *Bayes-minimax duality.*

For example, the "add-β" estimators are the simplest non-trivial (linear, but not constant) estimators given by

$$\hat{p}_{k,\beta}(n_k) := \frac{n_k + \beta}{N + d\beta}.$$

They are not minimax. However, it turns out that they are "not much worse" than the minimax estimator. First we note that the minimax risk is lower bounded by the Bayes risk for any prior:

$$\inf_{\hat{p}} \max_{p} R(p, \hat{p}) \geq r(\pi).$$

Breass and company [28] calculated a lower bound on the minimax risk using the uniform prior which agrees with the asymptotic results of Rukhin [181] and Krichevskiy [133]. They found, as $N \to \infty$,

$$\inf_{\hat{p}} \max_{p} R(p, \hat{p}) \geq 0.5\frac{1}{N} + O\left(\frac{1}{N}\right).$$

Whereas, the best "add-β" estimator has asymptotic risk

$$\inf_{\hat{p}_\beta} \max_{p} R(p, \hat{p}) = \beta_0\frac{1}{N} + O\left(\frac{1}{N}\right),$$

where $\beta_0 \approx 0.509$ and is achieved by the "add-β_0" estimator [181, 133]. Thus, the "add-β_0" estimator is trivial to implement and incurs only a constant fraction of excess risk. The risk of some "add-β" estimators for a die with $d = 2$ sides (also known as a coin) are depicted in figure 2.1 and figure 2.2.

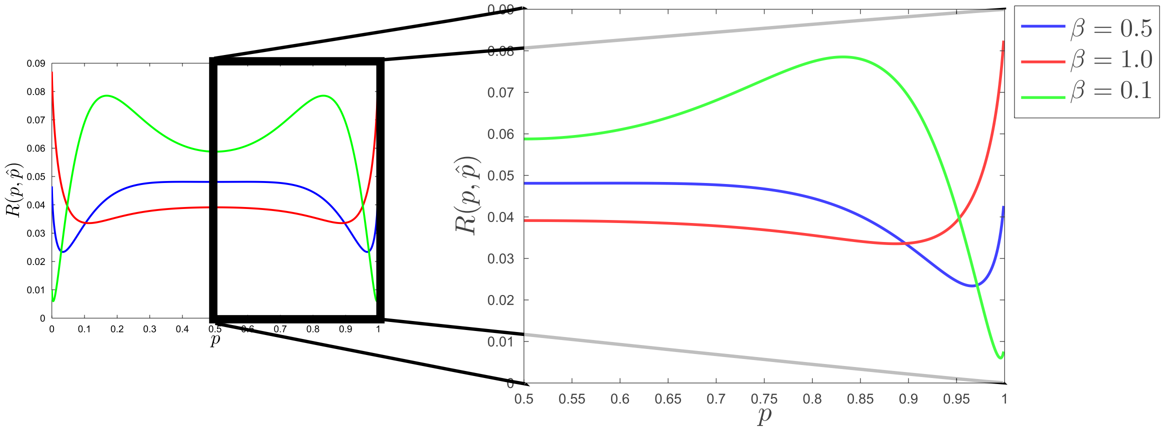

Figure 2.1: The risk of the hedged maximum likelihood estimator for $N = 10$ coin flips. On the left is the risk for all coin biases. Note that if the estimator possess the symmetry $\hat{p}(N-n) = \hat{p}(n)$, then it also posses the symmetry $R(1-p,\hat{p}) = R(p,\hat{p})$. Thus we will only display the non-redundant information, as shown on the right. The risk of the same strategies for $N = 100$ is shown in figure 2.2.

2.1.2 Quantum Relative Entropy

A qubit quantum state is equivalent to the bias of the coin in the sense that once it is known, we can probabilistically predict the outcome of future "flips". But estimating the quantum state from current data is somehow "harder" than classical probabilistic estimation. To see this, we must return to our story of Alice and Bob.

Alice does well with the coin, but this was only a test. Bob has a black-box, with the label "V-Wade", which he has been promised prepares a qubit which he would like to know the quantum state of. He asks Alice, who happens also to be an experimental physicist, to determine the state of his qubit. Alice reports $\hat{\rho}$ but Bob would like to know her honest opinion ρ for the state of the qubit. To ensure her honesty, Bob performs a measurement $\{E_i\}$ and will pay Alice $R(q_i)$ if he obtains outcome E_i, where $q_i = \operatorname{Tr}(\hat{\rho}E_i)$. Denote the honest probabilities of Alice by $p_i = \operatorname{Tr}(\rho E_i)$. Then her honesty is ensured if

$$\sum_i p_i R(p_i) \leq \sum_i p_i R(q_i).$$

The Aczel-Pfanzagl theorem holds and thus $R(p) = C \log p + B$. Thus, Alice's expected

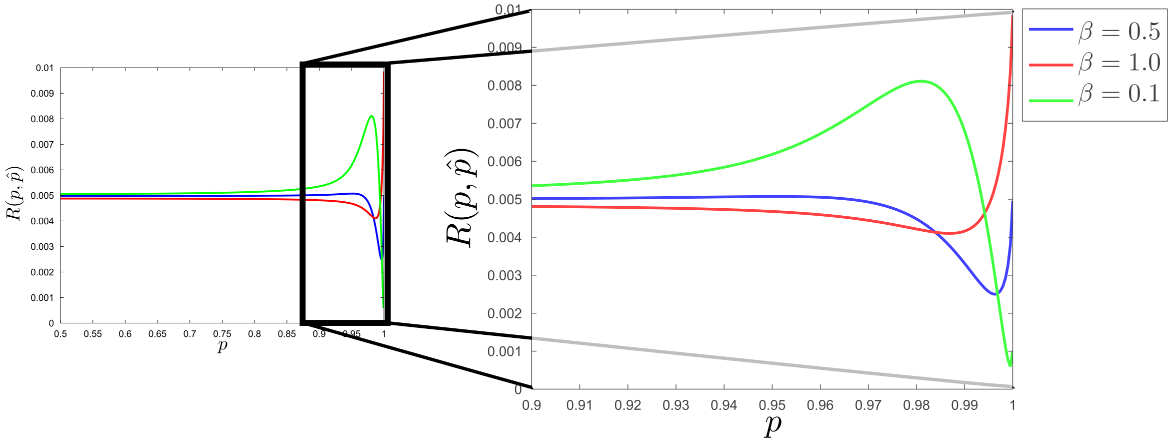

Figure 2.2: The risk of the hedged maximum likelihood estimator for $N = 100$ coin flips. As explained in figure 2.1, we need only see half the state space due to symmetry. However, we can also see that the "interesting region" becomes squashed near the boundary of the state space. This behaviour is generic for the problems we consider in the thesis so pay close attention to the region that is being plotted in the remainder. In this plot and 2.1, note that we can start to see the $O(1/N)$ behaviour of the (near) optimal "add-1/2" strategy.

loss is (up to a constant C)

$$\sum_i p_i(\log p_i - \log q_i) = \sum_i \mathrm{Tr}(\rho E_i) \log \left[\frac{\mathrm{Tr}(\rho E_i)}{\mathrm{Tr}(\hat{\rho} E_i)}\right].$$

For example [24], if Bob performs a measurement in the diagonal basis of Alice's reported state, then her expected loss is the *quantum relative entropy* [3]

$$D(\rho\|\hat{\rho}) = \mathrm{Tr}(\rho \log \rho) - \mathrm{Tr}(\rho \log \hat{\rho}).$$

Since the quantum relative entropy is strictly convex in its second argument, in this example, Alice is constrained to be honest since the minimum of $D(\rho\|\hat{\rho})$ is uniquely obtained at $\hat{\rho}$. This is not true for any measurement Bob can make (take the trivial measurement for example). So, we naturally must ask which measurements can Bob make to ensure Alice's honesty? That is, which measurements are characterized by

$$\sum_i \mathrm{Tr}(\rho E_i) \log \left[\frac{\mathrm{Tr}(\rho E_i)}{\mathrm{Tr}(\hat{\rho} E_i)}\right] = 0 \Leftrightarrow \hat{\rho} = \rho?$$

[3] See a history of this quantity in reference [79].

In some sense it does not matter since Alice does not know, in general, the measurement Bob is to perform. For Alice, she cares only about minimizing her expected worst case loss, which is the *quantum Kullback information* [61]

$$K(\rho\|\hat{\rho}) = \max_{\{E_i\}} \sum_i \mathrm{Tr}(\rho E_i) \log \left[\frac{\mathrm{Tr}(\rho E_i)}{\mathrm{Tr}(\hat{\rho} E_i)} \right].$$

Moreover, she cannot assume Bob will only measure one copy of the qubit. Supposing Bob can measure N copies the qubit with a possibly joint measurement, Alice's worst case expected loss is

$$\begin{aligned} K_N(\rho\|\hat{\rho}) &= \max_{\{E_i\}} \sum_i \mathrm{Tr}(\rho^{\otimes N} E_i) \log \left[\frac{\mathrm{Tr}(\rho^{\otimes N} E_i)}{\mathrm{Tr}(\hat{\rho}^{\otimes N} E_i)} \right] \\ &= D(\rho\|\hat{\rho}) \text{ as } N \to \infty \end{aligned}$$

and, non-asymptotically, $K_N(\rho\|\hat{\rho}) \leq D(\rho\|\hat{\rho})$ [111].

In exactly the same way as the classical scenario, the quantum relative entropy gives the redundancy of the codes used in data compression when the wrong quantum state is used [189]. That is, if Bob uses Alice's state $\hat{\rho}$ to perform *quantum data compression* and the true state is ρ, his codewords will have, on average, $D(\rho\|\hat{\rho})$ more qubits than necessary.

2.1.3 Quantum State Estimation

We are a little ahead of ourselves however. How does Alice come to the decision $\hat{\rho}$ in the first place? The natural thing, for an experimental physicist such as herself, to do is to perform a measurement of her own! Amongst the many possible schemes she could come up with, let us assume she is to perform a fixed measurement $\{E_k\}$ a total of N times. The set of possible outcomes is $\mathcal{O} := \{\{n_k\} : \sum_k n_k = N\}$, where each n_k denotes the number of times E_k occurred. Her reported state is a map $\hat{\rho} : \mathcal{O} \to \mathbb{D}(\mathcal{H})$ which takes experimental data $\{n_k\}$ to a density matrix. If Alice assumes there is a "true" state ρ, her expected loss is (bounded above but also asymptotically given by) the average of the quantum relative entropy over the possible experimental outcomes:

$$R(\rho, \hat{\rho}) := \sum_{\{n_k\}} \Pr(\{n_k\}|\rho) D(\rho\|\hat{\rho}(\{n_k\})),$$

where the likelihood function is

$$\Pr(\{n_k\}|\rho) = N! \prod_k \frac{\mathrm{Tr}(\rho E_k)^{n_k}}{n_k!}$$

a multinomial distribution.

The beauty of the general decision theoretic approach is that the methodology is independent of the model and dictates that we should minimize the risk R. Thus, all the definitions of the various estimation strategies are essentially the same. A minimax estimator $\hat{\rho}_{\text{minimax}}$ is

$$\hat{\rho}_{\text{minimax}} := \operatorname*{argmin}_{\hat{\rho}} \max_{\rho} R(\rho, \hat{\rho}).$$

This is the estimate which has the best worst case behaviour regardless of the probability of such an event. Supposing Alice has a prior distribution $\pi(\rho)d\rho$ in mind when it comes time to report $\hat{\rho}$, her expected loss (with respect to π) is

$$r(\pi, \hat{\rho}) := \int R(\rho, \hat{\rho})\pi(\rho)d\rho,$$

which is called the *Bayes risk* of $\hat{\rho}$ (with respect to π). A Bayes estimator is

$$\hat{\rho}_{\pi} := \operatorname*{argmin}_{\hat{\rho}} r(\pi, \hat{\rho}).$$

This is the estimator which has the best expected performance with respect to a given prior π. The Bayes risk *of the prior* π is the expected risk under π when using its Bayes estimator: $r(\pi) := r(\pi, \hat{\rho}_{\pi})$. That is, it is the expected risk when choosing the best estimate with respect to the given prior π. The prior

$$\pi_{\text{LFP}} := \operatorname*{argmax}_{\pi} r(\pi),$$

is called a least favorable prior. The *Bayes-minimax* duality theorem remains unchanged: if

$$r(\pi) = \max_{\rho} R(\rho, \hat{\rho}_{\pi}),$$

then $\hat{\rho}_{\pi}$ is minimax and π is least favorable. We also know exactly what the Bayes estimators are. Given any prior π, the Bayes estimator is the mean of the Bayesian updated posterior [208] (this is more generally true for any *strictly proper scoring rule*, not just the quantum relative entropy [24]).

2.1.4 Hedged Maximum Likelihood

Given that "add-β" estimators are near-optimal for classical probability estimation, Blume-Kohout generalized these strategies to the estimation of quantum states [23] – but what he found was surprising!

First we note that classical "add-β" estimators achieve the maximum of a *hedged* likelihood function $h_\beta(p) \times \Pr(n|p)$ where the hedging function is

$$h_\beta(p) := p^\beta (1-p)^\beta.$$

In order to generalize this to the estimation of quantum states, the hedging function ought to be "add-β" if the measurement outcomes are mutually exclusive (only one outcome of a set can occur). That is, if the measurement is a standard projective measurement (if the set of measurement operators $\{P_k\}$ are orthogonal projectors) the estimated quantum state should be the density operator

$$\hat{\rho} = \sum_k \frac{n_k + \beta}{N + 2\beta} P_k$$

when the measurement yields outcomes $\{n_k\}$. It was shown that the only measurement-independent hedging function which achieves this is

$$h_\beta(\rho) = \det(\rho)^\beta.$$

For an arbitrary measurement, then, the *hedged maximum likelihood* (HML) estimators are defined as

$$\hat{\rho}_\beta := \operatorname*{argmax}_\rho h_\beta(\rho) \prod_k \frac{\mathrm{Tr}(\rho E_k)^{n_k}}{n_k!}.$$

The intuition behind this choice of hedging function stems from the fact that $\det(\rho)$ is the product of the eigenvalues of ρ and is unitarily invariant. So these should be nearly optimal having only a constant offset from the asymptotic $O(1/N)$ risk scaling of classical probability estimation, right? Wrong! As depicted in figure 2.4, the risk of the HML estimators is significantly worse near the boundary of the state space. Next, we will investigate this curiosity from a less quantitative point of view.

2.1.5 Conceptual Lessons Learned

Why is quantum state estimation different from classical probability estimation? The key feature is measurement. In classical probability, measuring reveals the state of system (Alice and Bob get to see if the die toss produced "1" or "2" or...). In quantum theory, however, the true state is not revealed and the relationship between the true state and the outcome of the experiment depends strongly on what measurement is made. Perhaps, then, we should look for a better classical analogue of quantum state estimation. The intuition for the solution presented in the next section is better described pictorially, as is done in figure 2.5.

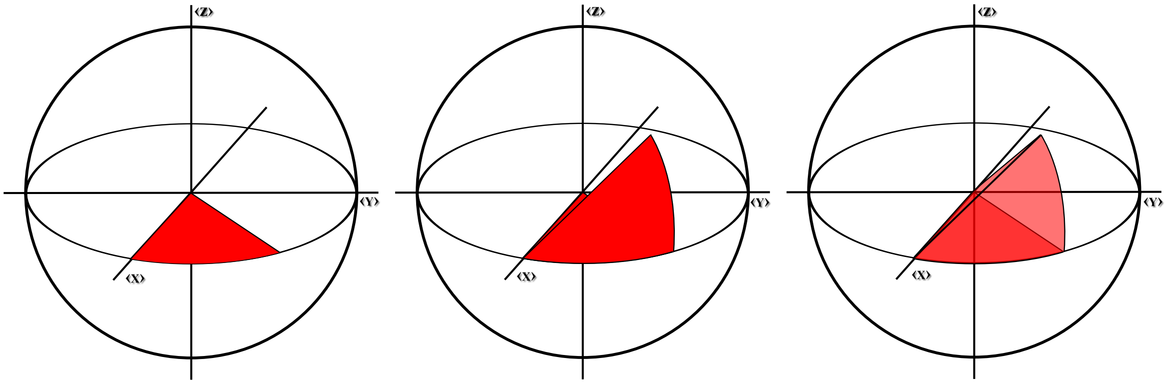

Figure 2.3: As in figure 2.1 for the classical problem, the quantum state estimation problem possesses convenient symmetries depending on the measurements that are made. Here we have assumed an equal number N of Pauli X, Y and Z measurements are made and the estimator satisfies $\hat{\rho}(N - n_x, N - n_y, N - n_z) = \hat{\rho}(n_x, n_y, n_z)$. The corresponding non-redundant region in the Bloch sphere is constructed following the left to right flow in the figure. First create a pie slice in the XY-plane which makes an angle of 45^o from the X-axis. From this point off the axis, move along a great circle toward the north pole approximately 54.7^o (the so-called "magic angle" – $\arctan \sqrt{2}$). Finally, enclosed the convex region which connects this point to the origin. This is the region is the quantum state space which contains no redundant information on the risk of a symmetric estimator. Of course, it will be difficult to plot the risk in the region since we are missing a fourth spacial dimension. Thus, the risk will be plotted along the 3 rays from the origin on the edges of this regions and it will become clear how this contains the majority of the useful information on the performance of the estimators we will consider.

2.2 Noisy Coins: a Better Classical Analogue

So we have seen how classical probably estimation fails for quantum state estimation. However, it seems (see figure 2.5 again) that we have been comparing to a poor classical analogy. What we need is a classical problem where the sampling distribution is different from the one we want to estimate. For this we invent the "noisy coin". Adding noise – random bit flips with probability α – to the coin flip data separates the effective probability of "heads",

$$q = \alpha + p(1 - 2\alpha), \tag{2.3}$$

from the true probability p. Quite a lot of the complications that ensue can be understood as stemming from a single underlying schizophrenia in the problem: there are now *two*

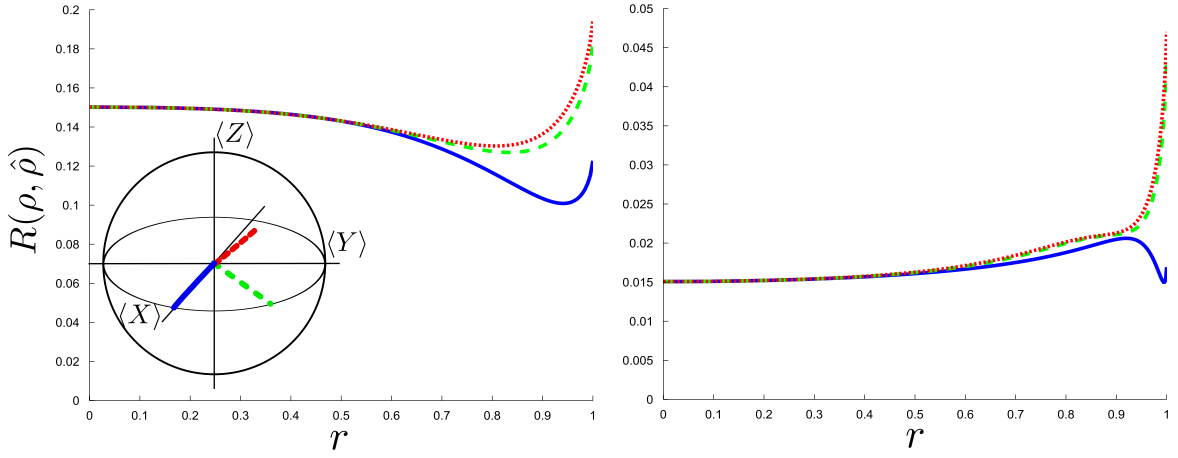

Figure 2.4: The risk of the hedge maximum likelihood estimator for $\beta = 0.5$ for $N = 10$ Pauli X, Y and Z measurements on the left and $N = 100$ on the right. The horizontal axis is label by r, the radius of the Bloch sphere. Note that three rays are plotted. As noted in figure 2.3, these rays form the edges of the non-redundant region of the Bloch sphere. Comparing them to figures 2.1 and 2.2, which display the risk of the hedged maximum likelihood estimators for a coin, we can see that quantum behaviour of these strategies is both qualitatively and quantitatively different. Indeed, it is clear that the risk for a qubit is not going to be $O(1/N)$. More measurements would show that the risk is $O(1/\sqrt{N})$ near the boundary and remains $O(1/N)$ in the "bulk" of the state space.

relevant probability simplices, one for q and one for p. One part of the problem (the data, and therefore the likelihood function) essentially live on the q-simplex. The other parts (the parameter to be estimated, and therefore the risk function) live on the p-simplex. The core problem here is sampling mismatch. We sample from q, but the risk is determined by p.

As we attempted to do for the quantum mechanical problem, we will try to replicate the classical "add-β" estimator via hedging the likelihood function. The "hedging function" for the noisy coin is $h(p) = \prod_k p_k^\beta = p^\beta(1-p)^\beta$ (which is analogous to the quantum mechanical version: $h(\rho) = \det(\rho)^\beta$). We define the hedged maximum likelihood (HML) estimator for the noisy coin as

$$\hat{p}_\beta(n) = \operatorname*{argmax}_p h(p) \Pr(n|p).$$

See figure 2.6 for a graphical depiction of the effect of hedging. For a noiseless coin, this is identical to adding β fictitious observations of each possible event – but for $\alpha > 0$, they

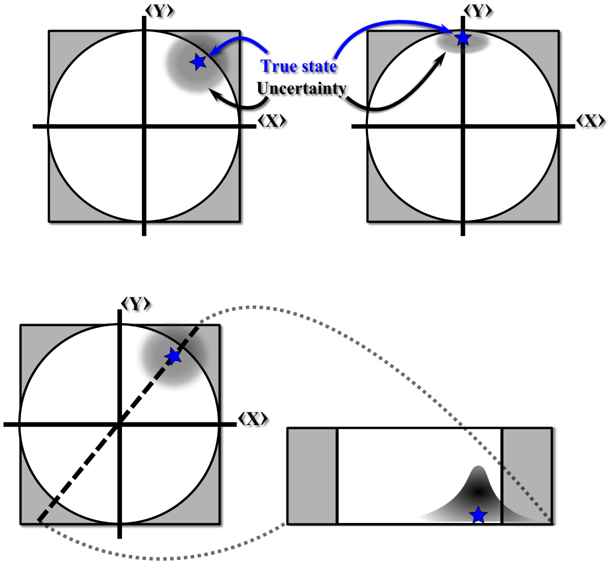

Figure 2.5: Plotted is a slice of the Bloch sphere. The outer shaded box is a slice of the "Bloch cube" which is the boundary of the set of possible empirical frequencies. The measurement axes are X and Y. Notice that when the state lies in a measurement axis, the measurements act as (and are formally equivalent to) tosses of two independent coins. However, when the state lies outside the measurement axes, it is possible to obtain a frequency of outcomes which lies outside the allowed states. If we imagine that the state lies along a hypothetical measurement axis, the measurement of the quantum system is equivalent to what we call a *noisy coin*.

are not equivalent. The hedging function modification is sensitive to the $p_k = 0$ boundary of the simplex, and inexorably forces the maximum of $h(p)\Pr(n|p)$ away from it (since $h(p)\Pr(n|p)$ remains log-convex, but equals zero at the boundary).

For the noisy coin we can solve for the HML estimator explicitly. It is given by

$$\hat{p}_\beta = \frac{\hat{q}_\beta - \alpha}{1 - 2\alpha},$$

where $\hat{q}_\beta$ is the zero of the cubic polynomial

$$(N + 2\beta)\hat{q}^3 - (N + n + 3\beta)\hat{q}^2 + (n + \beta + N\alpha - N\alpha^2)\hat{q} + n\alpha^2 - n\alpha.$$

that lies in $[\alpha, 1 - \alpha]$.

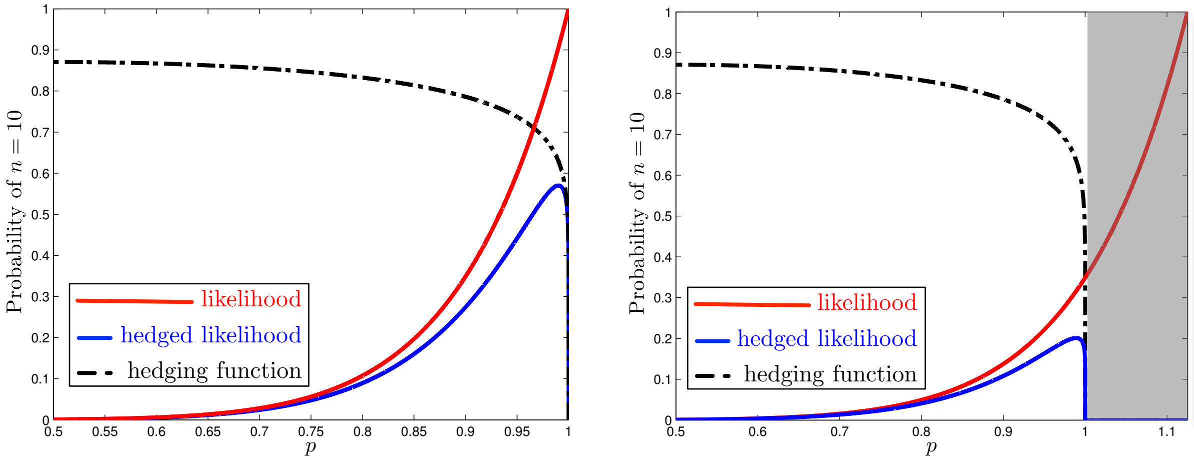

Figure 2.6: Illustration of a hedging function ($\beta = 0.1$), and its effect on the likelihood function for noiseless (left) and noisy (right) coins. In the left plot, we show the hedging function over the 2-simplex (dotted black line), the likelihood function for an extreme data set comprosing 10 heads and 0 tails (red line), and the corresponding *hedged* likelihood (blue line). The right plot shows the same functions for a noisy coin with $\alpha = 0.1$. The shaded regions are outside the p-simplex (and therefore forbidden), but correspond to valid q values. Note that the *unconstrained* maximum of $\Pr(n|p)$ lies in the forbidden region where $p > 1$, and therefore the maximum of the constrained likelihood is on the boundary ($p = 1$) – a pathology that hedging remedies.

Figure 2.7 illustrates how $\hat{p}_\beta$ depends on $\frac{n}{N}$. In the "bulk", far from the simplex boundary (0 and 1), hedging has relatively little effect – it behaves essentially the same as the maximum likelihood estimator ($\beta = 0$). In fact, hedging yields an approximately linear estimator akin to "add-β". But as we approach 0 or 1, the effect of hedging increases. When a linear estimator intersects 0, at $n = \alpha N$, $\hat{p}_\beta = O(1/\sqrt{N})$. This fairly dramatic shift occurs because the likelihood function is approximately Gaussian, with a maximum at $p = 0$ and a width of $O(1/\sqrt{N})$. $\Pr(n|p)$ declines rather slowly from $p = 0$, and $p = O(1/\sqrt{N})$ is not *substantially* less likely than $p = 0$, so the hedging imperative to avoid $\hat{p} = 0$ pushes the maximum of $h(p)\Pr(n|p)$ far inside the simplex.

How accurate are these hedged estimators? Figure 2.8 shows the *pointwise* risk as a function of the true p, for different amounts of hedging (β). For the noiseless ($\alpha = 0$) coin, $\beta = 1/2$ yields a nearly flat risk profile given by $R(p) \approx 1/2N$. In contrast, hedged estimators for the noiseless coin yield similar profiles that rise from $O(1/N)$ in the interior to

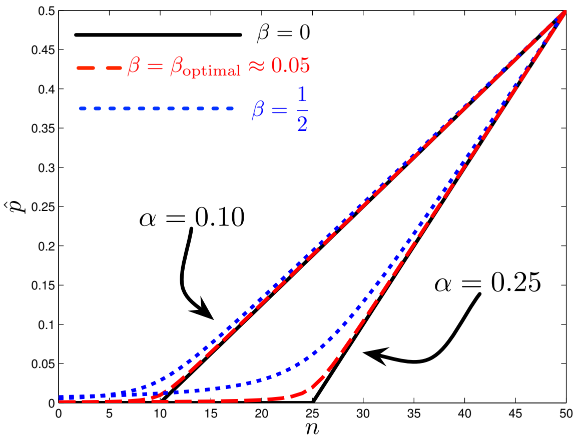

Figure 2.7: HML (hedged maximum likelihood) estimators $\hat{p}_\beta(n)$ are shown for several values of β, and compared with the maximum likelihood (ML) estimator $\hat{p}_{\text{ML}}(n)$. $N = 100$ in all cases. Whereas the ML estimator is linear in n until it encounters $p = 0$, the hedged estimator smoothly approaches $p = 0$ as the data become more extreme. Increasing β pushes $\hat{p}$ away from $\hat{p} = 0$.

a peak of $O(1/\sqrt{N})$ around $p = O(1/\sqrt{N})$. This is exactly the same behaviour we witness for the quantum problem in figure 2.4. At first glance, this behavior suggests a serious flaw in the hedged estimators. The peak around $p \approx O(1/\sqrt{N})$ is of particular concern, since in all cases the risk is $O(1/\sqrt{N})$ there. But in fact, this behavior is generic for the noisy coin. Minimax estimators have similar $O(1/\sqrt{N})$ errors, and hedged estimators turn out to perform quite well. However, they are *not* minimax, or even close to it! As we shall soon see, the noisy coin's "intrinsic risk" profile is far from flat. The minimax estimator attempts to flatten it – at substantial cost.

Some simple estimators (such as "add-1/2") are nearly minimax for the noiseless coin. This is not true for the noisy coin in general, because (as we shall see) the minimax estimators are somewhat pathological. So we used numerics to find good approximations to minimax estimators for noisy coins [68]. An example is plotted in 2.9, which illustrates minimax estimators for several N and α, while Figure 2.10 shows the resulting risk profiles. The minimax risk is $O(1/\sqrt{N})$ – *not* $O(1/N)$ as for the noiseless coin. The minimax estimators are highly biased toward $p \approx 1/\sqrt{N}$ – not just when p is close to the boundary (when bias is inevitable) but also when p is in the interior! This effect is truly pathological, although it can easily be explained. Low risk, of order $1/N$, can easily be achieved in the interior. However, the minimax estimator seeks at all costs to reduce the *maximum* risk,

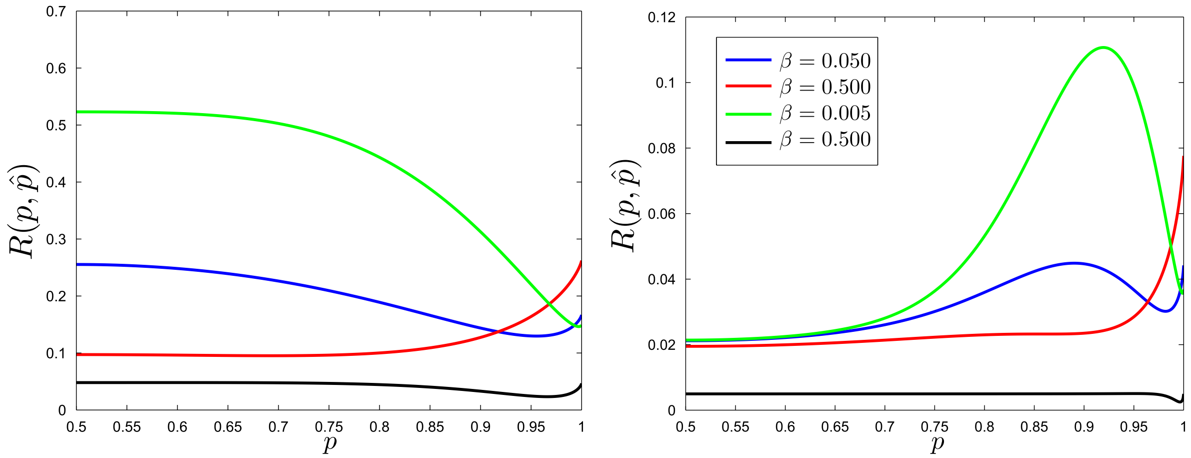

Figure 2.8: The risk profile $R(p, \hat{p}_\beta)$ is shown for several hedge maximum likelihood estimators for $N = 10$ on the left and $N = 100$ on the right. A noisy coin with $\alpha = 1/4$ has been estimated using three different HML estimators. The optimal $\beta \approx 0.05$ balances boundary risk against interior risk. Increasing β increases boundary risk, while decreasing it increases interior risk. Risk approaches $O(1/\sqrt{N})$ for noisy coins, vs. $O(1/N)$ for noiseless coins which is reproduced in the black curve for comparison.

which is achieved near $p \approx 1/\sqrt{N}$. By biasing heavily toward $p \approx 1/\sqrt{N}$, the estimator achieves slightly lower maximum risk–at the cost of dramatically increasing its interior risk from $O(1/N)$ to $O(1/\sqrt{N})$.

2.2.1 Bimodal Risk and the Comprimising Optimality of Hedging

The preceding analysis made use of an intuitive notion of pointwise "intrinsic risk" – i.e., a lower bound $R_{\min}(p)$ on the expected risk for any given p. Formally, no such lower bound exists. We can achieve $R(p') = 0$ for *any* p', simply by using the estimator $\hat{p} = p'$. But we can rigorously define something very similar, which we call *bimodal risk*.

The reason that it's not practical to achieve $R(p') = 0$ at any given p' is, of course, that p' is unknown. We must take into account the possibility that p takes some other value. Least favorable priors are intended to quantify the risk that ensues, but a LFP is a property of the entire problem, not of any particular p'. In order to quantify "how hard is

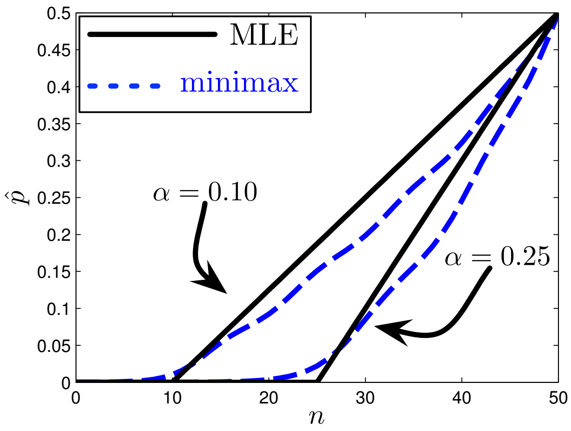

Figure 2.9: Minimax and maximum likelihood estimators are shown for $N = 100$ and $\alpha = 1/10, 1/4$. Note that the minimax estimator is grossly biased in the interior – a pathological result of the mandate to minimize *maximum* risk at all costs.

a particular p' to estimate," we consider the set of bimodal priors,

$$\pi_{w,p',p''}(p) = w\delta(p - p') + (1 - w)\delta(p - p''),$$

and maximize Bayes risk over them. We define the bimodal risk of p' as

$$R_{\text{bimodal}}(p') = \max_{w,p''} r(\pi_{w,p',p''}).$$

The bimodal risk quantifies the difficulty of distinguishing p' from *just one* other state p''. As such, it is always a lower bound on the minimax risk.

Figure 2.11 compares the bimodal risk to the pointwise risk achieved by the minimax and (optimal) HML estimators. Note that the bimodal risk function is a strict lower bound (at every point) for the minimax risk, but not for the pointwise risk of any estimator (including the minimax estimator itself). However, every estimator exceeds the bimodal risk at at least one point, and almost certainly at *many* points. Figure 2.11 confirms that the noisy coin's risk is dominated by the difficulty of distinguishing $p \approx 1/\sqrt{N}$ from $p = 0$. States deep inside the simplex are far easier to estimate, with an expected risk of $O(1/N)$.

Minimax is an elegant concept, but for the noisy coin it does not yield "good" estimators. In a single-minded quest to minimize the maximum risk, it yields wildly biased estimates in the interior of the simplex. This is reasonable only in the case where p is truly

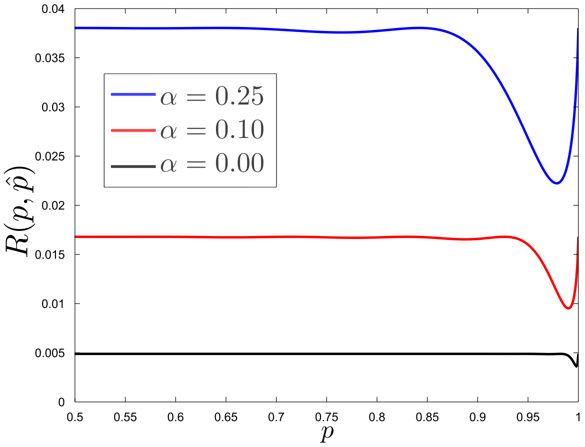

Figure 2.10: The risk profile $R(p,\hat{p})$ is shown for the minimax estimators of figure 2.9 ($N = 100$; $\alpha = 1/10, 1/4$) and for a noiseless coin (also $N = 100$). No estimator can achieve lower risk across the board – but the minimax estimator's risk is very high in the interior ($p > 1/\sqrt{N}$) compared with HML estimators. The $O(1/\sqrt{N})$ risk of HML estimators is intrinsic to noisy coins.

selected by an adversary. In the real world, robustness against adversarial selection of p is good, but should not be taken to absurd limits.

This leaves us in need of a quantitative criterion for "good" estimators. Ideally, we would like an estimator that achieves (or comes close to achieving) the "intrinsic" risk for every p. The bimodal risk $R_{\text{bimodal}}(p)$ provides a reasonably good proxy – or, more precisely, a lower bound – for intrinsic risk (in the absence of a rigorous definition). This is not a precise quantitative framework, but it does provide a reasonably straightforward criterion: we are looking for an estimator that closely approaches the bimodal risk profile.

Hedged estimators are a natural ansatz, but we need to specify β. Whereas the noiseless coin is fairly accurately estimated by $\beta = 1/2$ for all N, the optimal value of β varies with N for noisy coins. Local maxima of the risk are located at $p = 0$ and at $p \approx 1/\sqrt{N}$, one or both of which is always the global maximum. So, to choose β, we minimize maximum risk by setting them equal to each other. Elsewhere [68], we showed that this optimum approaches $\beta_{\text{optimal}} \approx 0.0389$ for large N. This value is obtained for a large range of α's, as shown in Figure 2.12. We conclude that while optimal hedging estimators probably do not offer strictly optimal performance, they are (i) easy to specify and calculate, (ii) far better than minimax estimators for almost all values of p, and (iii) relatively close to the

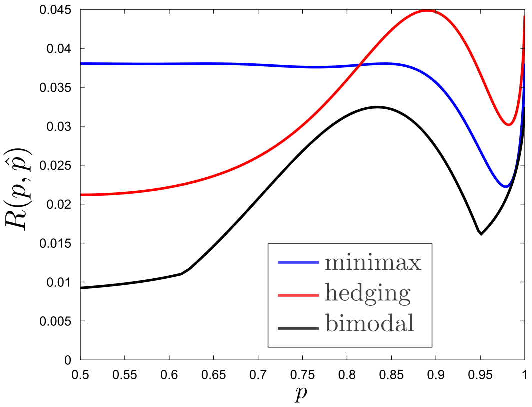

Figure 2.11: The risk profiles of the optimal HML estimator (red) and the minimax estimator (blue) are compared with the bimodal lower bound $R_{\text{bimodal}}(p)$. Note that while the HML maximum risk exceeds the minimax risk (as it must!), it is competitive – and HML is *far* more accurate in the interior. The bimodal lower bound supports the conjecture that HML is a good compromise, since the HML risk exceeds the bimodal bound by a nearly constant factor.

lower bound defined by bimodal risk.

On the theoretical side, bimodal prior also provide us with a relatively simple method to obtain asymptotic lower bound on the minimax risk. We can obtain this asymptotic lower bound using the normal approximation to the binomial distribution: $n \sim \mathcal{N}[Nq, Nq(1-q)]$. Note that for $\alpha > 0$ the normal approximation is valid for all p whereas it is invalid for $p \to 0$ when $\alpha = 0$. The Bayes risk is asymptotically given by

$$r(\pi) = \frac{1}{2}[R(p_0, \hat{p}) + R(p_1, \hat{p})].$$

Now let us choose $p_0 = 0$ and $p_1 = 1/\sqrt{N}$ and the very crude bound $R(p_1, \hat{p}) \geq 0$. We focus then on the first term, for which

$$D(0\|\hat{p}(n)) = -\log(1 - \hat{p}(n)) \sim \hat{p}(n)$$

is a monotonically increasing function of n. We can make use of the measure theoretic Chebyshev inequality:

$$R(0, \hat{p}) \geq \frac{1}{2} q(t) \Pr(n \geq t).$$

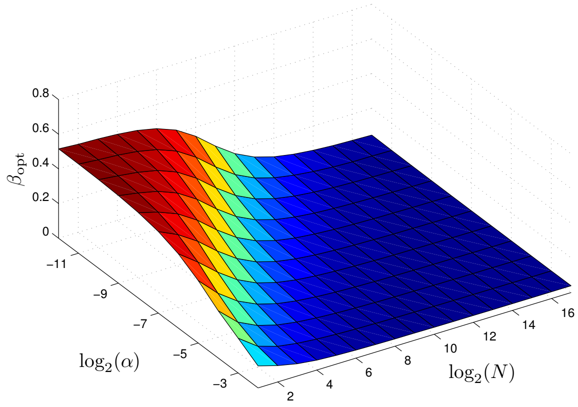

Figure 2.12: The optimal value of β for $N = 2 \dots 2^{17}$ and $\alpha = 2^{-12} \dots 2^{-2}$. It approaches the optimal noiseless value $\approx 1/2$ when $N \ll \alpha^{-1}$. It rapidly declines and at roughly $N \approx \alpha^{-1}$ it is well within 10^{-2} of what appears to be its asymptotic value $\beta_{\rm optimal} \approx 0.0389$.

The choice $t = \alpha N$, the mean of the distribution, is convenient since it can be shown that

$$\hat{p}(\alpha N) = \frac{1}{\left(1 + e^{-2+\frac{1}{2\alpha(1-\alpha)}}\right)} \frac{1}{\sqrt{N}} =: 2B(\alpha)\frac{1}{\sqrt{N}}, \tag{2.4}$$

where the implicitly defined scaling constant B is independent of N. Then the risk is lower bounded by

$$\max_p R(p, \hat{p}) \geq B(\alpha)\frac{1}{\sqrt{N}} + O\left(\frac{1}{N}\right).$$

2.3 Optimal Qubit State Estimation

Now, we apply the intuition afforded to us by the noisy coin to quantum state estimation. Consider the following question: what is the quantum state of a qubit when is it known that identical preparations of it have been measured in each of the bases X, Y and Z and the outcomes are $(n_x, N - n_x)$, $(n_y, N - n_y)$ and $(n_z, N - n_z)$, respectively? As depicted in figue 2.5, this is conceptually similar to estimating a noisy coin bias.

When the data set is $(n_x, n_y, n_z) =: n \in \{0, 1, \dots, N\}^3$, the estimate is a map $n \mapsto \hat{\rho}(n)$ and the likelihood function, the probability of data given the true state, simplifies to the product of three binomial distributions:

$$\Pr(n|\rho) = \prod_{k=x,y,z} \binom{N}{n_k} p_k^{n_k}(1-p_k)^{N-n_k},$$

where $p_x = \frac{1}{2}(1 + \mathrm{Tr}(\rho X))$ and similarly for p_y and p_z. These probabilities are the coordinates of the Bloch sphere representation of ρ. Given the results for the risk of the noisy coin, we can intuit that the risk of estimating qubits is at least $O(1/\sqrt{N})$. In fact, we can show this formally. Let

$$\rho = \sum_i p_i |p_i\rangle\langle p_i|, \quad \hat{\rho} = \sum_i \hat{p}_i |\hat{p}_i\rangle\langle \hat{p}_i|,$$

be the spectral decompositions of ρ and $\hat{\rho}$. Then we have

$$D(\rho\|\hat{\rho}) = \sum_i p_i \log p_i - \sum_{ij} p_i \log \hat{p}_j \left|\langle p_i, \hat{p}_j\rangle\right|^2.$$

Application of Jensen's inequality gives $D(\rho\|\hat{\rho}) \geq D(p\|m)$, the classical relative entropy between the probability distribution p of eigenvalues of ρ and the probability distribution m defined as

$$m_i = \sum_j \hat{p}_j \left|\langle p_i, \hat{p}_j\rangle\right|^2.$$

The probabilities can be written

$$p_x = p_0(1-\alpha_x) + (1-p_0)\alpha_x,$$

with $\alpha_x = \frac{1}{2}(1 + \langle p_0|X|p_0\rangle)$ and similarly for p_y and p_z. Since the measurements are independent, a second application of Jensen's inequality gives

$$R(\rho,\hat{\rho}) \geq \sum_{k=x,y,z} \sum_{n_k} \binom{N}{n_k} p_k^{n_k}(1-p_k)^{N-n_k} D(p_k\|m_k).$$

This is equivalent, in the worst case, to the risk of estimating three independent biases of noisy coins with noise levels α_x, α_y and α_z. Thus

$$\max_{\rho\in\mathbb{D}(\mathcal{H})} R(\rho,\hat{\rho}) \geq (B(\alpha_x) + B(\alpha_y) + B(\alpha_z))\frac{1}{\sqrt{N}} + O\left(\frac{1}{N}\right),$$

where B is implicitly defined in Eq. 2.4.

Applying our intuition gained from the noisy coin, we see that each axis of the Bloch sphere acts a noisy coin whose effective "noise" is determined by the angle from the nearest measurement axis. We expect, then, that the minimax risk should attempt to be flat, at $O(1/\sqrt{N})$, as possible over all states. For smallish N, we can numerically construct minimax estimators. We do so with two algorithms, both of which rely on Bayes-minimax duality. That is, we construct Bayes estimator from approximately least favorable priors. Each prior $\pi(\rho)d\rho$ defines a Bayesian mean estimator $\hat{\rho}_\pi(n)$, which is Bayes for $\pi(\rho)$. Its risk profile $R(\rho, \hat{\rho}_\pi)$ provides both upper and lower bounds on the minimax risk – the maximum value is an upper bound and the average is a lower bound.

As is often the case for discretely distributed data, the minimax priors appear to always be discrete [137]. We searched for least favorable priors (holding N fixed) using the algorithm of Kempthorne [125]. We defined a prior with a few support points, and let the location and weight of the support points vary in order to maximize the Bayes risk. Once the optimization equilibrated, we added new support points at local maxima of the risk, and repeated this process until the algorithm found priors for which the maximum and Bayes risk coincided to within 10^{-3} relative error.

To get a sense of the how challenging this is, note that minimax estimators for a simple coin had only been numerically found for up to tens of flips. Therefore, it shouldn't be surprising that this algorithm was only able to find minimax estimators for up to $N = 10$ Pauli measurements on a qubit. Utilizing everything we know about the symmetry of the problem, we can reduce the dimension of the search space by almost 100-fold. However the difficulty still lies in the fact that maximizing the Bayes risk in this way consists of multiple global optimizations problem over a space whose dimension is increasing cubically in N. For example, to find the least favorable for $N = 10$ required iteratively maximizing the Bayes risk over spaces of dimension 8 up to 28, which took about 4 days on a modern laptop.

All is not lost, however! We can find approximately least favorable priors via Monte Carlo methods. The key idea is that, by fixing states and optimizing only over the weights, the problem reduces to a convex optimization problem, which can be solved much more efficiently. The solution of this problem will not produce a qualitatively accurate least favorable prior, but the estimators derived from them will arbitrarily close to minimax provided enough support points are randomly chosen. We demonstrate the least favorable priors found by both of our algorithms in figure 2.13 for a *rebit*. A rebit is a qubit restricted to the XY plane, which plane allows us to more easily visualize the risk and least favorable priors. The minimax risk of the qubit and rebit is plotted in figure 2.14, which shows the scaling is equivalent in either case.

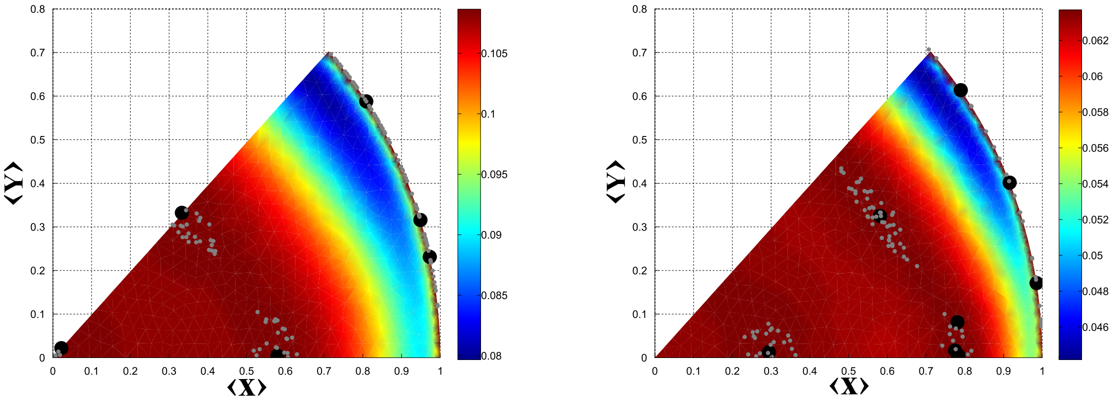

Figure 2.13: The risk profile $R(p, \hat{p})$ is shown for the Bayes estimators of the numerically found least favorable prior. As noted in figure 2.3, it is difficult to visualize a function over the interior of the Bloch sphere so we first illustrate the idea with a *rebit*: a qubit restricted to the XY plane. The left is for $N = 10$ Pauli X and Y measurements while the right plot has $N = 20$. The larger black dots are the support points of the least favorable prior as found by our first algorithm. The smaller grey dots are the support points retained by the Monte Carlo algorithm.

In figure 2.15, the risk of the numerically computed minimax estimators is plotted along with the bimodal risk. As we expected from the noisy coin analogy, the "intrinsic risk" of states furthest away from a measurement axis is higher. Those near the boundary on these "bad" axes are the only ones which contribute to the $O(1/\sqrt{N})$ scaling while all other states have $O(1/N)$ intrinsic risk. The minimax estimator ignores this structure, as it must, and has the less favorable $O(1/\sqrt{N})$ everywhere. We must remedy this pathology. Fortunately, we now know to look to the hedged maximum likelihood estimators for the desired fix. Also plotted in figure 2.15 is the risk of the HML estimators, which we see are a good compromise – mimicking the profile of the bimodal risk.

Therefore, we propose that the HML estimator is the optimal choice for quantum state estimation. Although it is not universal in the strict sense of minimax, it is nearly minimax and has quadratically improved performance for most states. Moreover, it is much easier to calculate that the minimax estimator. The analysis of hedging for the noisy coin given us the optimal hedging parameter $\beta_{\text{optimal}} \approx 0.0389$.

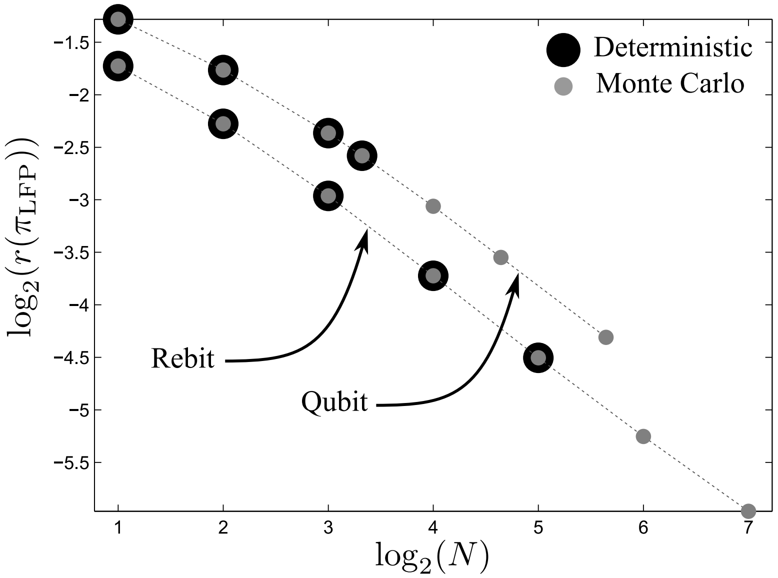

Figure 2.14: The Bayes risk of the numerically computed least favorable priors for a qubit and rebit as a function of the number of measurements. The larger data markers are for the deterministic algorithm while the smaller grey dots represent the values found for a single run of the Monte Carlo algorithm which used 500 initial Monte Carlo samples from a "flat" (or Hilbert-Schmidt or uniform) prior.

2.3.1 Adaptive State Estimation

In this chapter, we have considered the worst-case performance, with respect to relative entropy loss, of estimators for quantum states and its classical analogue: the noisy coin. We have shown that both have a worst-case risk that scales as $O(1/\sqrt{N})$ and is overwhelmingly dominated by nearly-pure states. Hedged maximum likelihood estimators achieve asymptotic scaling with a constant penalty near the boundary. More importantly, for all states in the bulk region (not within $O(1/\sqrt{N})$ of the boundary), HML estimators achieve $O(1/N)$ risk.

We have seen that each axis of the Bloch sphere acts a noisy coin whose effective "noise" is determined by the angle from the nearest measurement axis. What this suggests is that states whose axes lie in a measurement axis should behave as a noiseless coin - and hence have the more favorable risk scaling $O(1/N)$. To see this, suppose the true state lies on one the axes defined by the measurements. Without loss of generality, let $\rho = pX^+ + (1-p)X^-$.

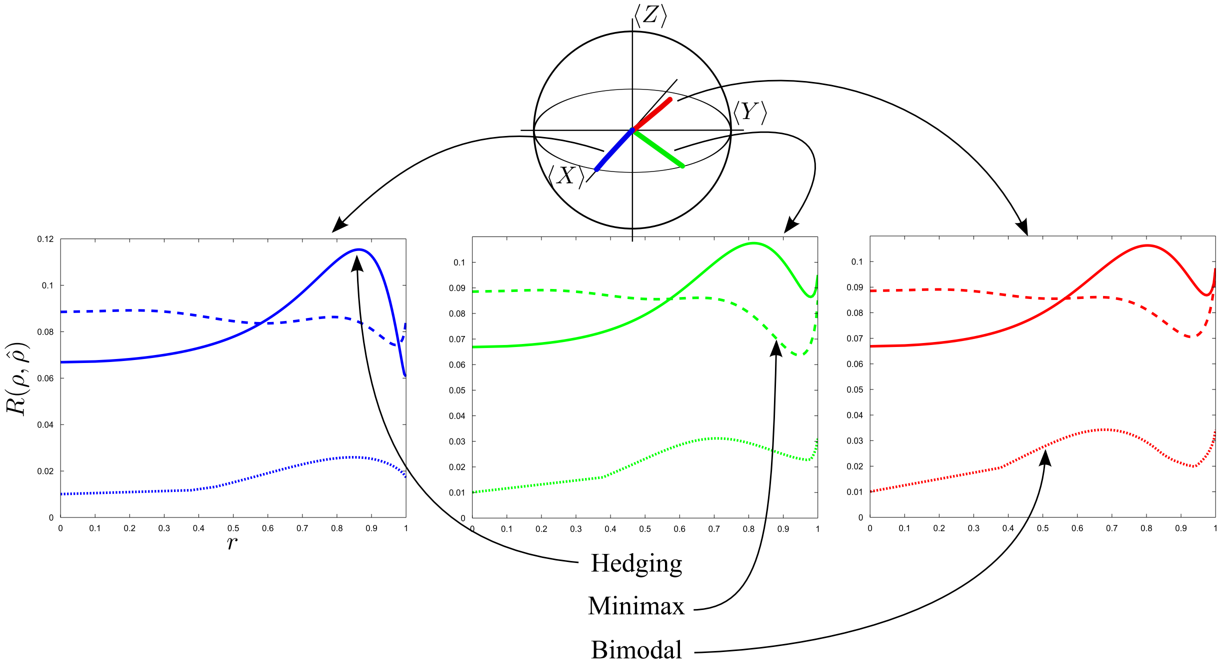

Figure 2.15: The risk profile $R(p, \hat{p})$ is shown for the HML estimator, minimax estimator and bimodal risk lower bound. The risk is plotted for $N = 25$ Pauli X, Y and Z measurements. Note that the HML risk profile matches that of the bimodal risk.

Then,

$$\Pr(n|\rho) = \binom{N}{n_x}\binom{N}{n_y}\binom{N}{n_z} p^{n_x}(1-p)^{N-n_x}\left(\frac{1}{2}\right)^{2N},$$

and

$$\begin{aligned} R(\rho, \hat{\rho}) &\geq \left(1 - \frac{1}{2^N}\right)^2 \sum_{n_x} \binom{N}{n_x} p^{n_x}(1-p)^{N-n_x} D(p\|m) \\ &\geq \sum_{n=0}^{N} \binom{N}{n} p^n (1-p)^{N-n} D(p\|m), \end{aligned}$$

which, in the worst case, is identical to the risk of estimating the bias of a noiseless coin. Thus, when ρ has an eigenbasis equivalent to a measurement basis,

$$\max_{\rho \in \mathbb{D}(\mathcal{H})} R(\rho, \hat{\rho}) \geq 0.5\frac{1}{N} + O\left(\frac{1}{N^2}\right).$$

This results suggests for the estimating a general qubit state, one should attempt to adapt the measurements to align with the eigenbasis of the state. If such an adaptive scheme were possible, our result proves it would achieve the favorable risk scaling of noiseless coin bias estimation.

Adaptive tomography has been considered before. For example, it has been shown for an alternative risk function, namely *fidelity*, that the Bayes risk with respect to the class of *Bures* priors can be minimized, albeit asymptotically, by only a single adaptive step [12]. The number of measurements to spend on determining the adaptation depends on the prior. Therefore, if one is concerned about worst-case risk, as we are here, it is not clear from an experimental perspective, where finite data counts may be very low, how many measurements should be performed before adapting to achieve the minimax bound.

Moreover, for non-asymptotic data counts, it may be the case that more than a single adaptation is necessary to achieve good performance. However, our numerical results suggest that the limit of adapting after every step [118] is far from necessary. Indeed, the returns of multiple adaptation diminish quite rapidly when using optimal HML estimators. This is illustrated in figure 2.16.

2.4 Discussion and Future Directions

The noisy coin analogy is a convenient and intuitive proxy for the problem of quantum state estimation – but only when the measurements are binary. It is likely the case that an analogy to noisy *dice* is appropriate for a more general measurement. However, it is not clear that, for example, hedging will produce the same favorable properties nor is it obvious that the numerically found optimal parameters for hedging will be the same, although we conjecture this to be the case.

We are also one step removed from the truly *universal* estimators for classical prediction since our strategies still depend on N. That is, the number of measurements N must be known. This is not a problem for given data, but it is certainly inconvenient, and probably ultimately infeasible, to have to compute an estimator for every N. In classical theory, estimators are usually designed to "deal with" any N. The "add-β" estimators, although they depend on N, are trivial to implement. To explicitly remove the dependence on N, *cumulative risk* is usually replaces the standard risk function:

$$R_N(\rho,\hat{\rho}) \mapsto \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} R_n(\rho,\hat{\rho}).$$

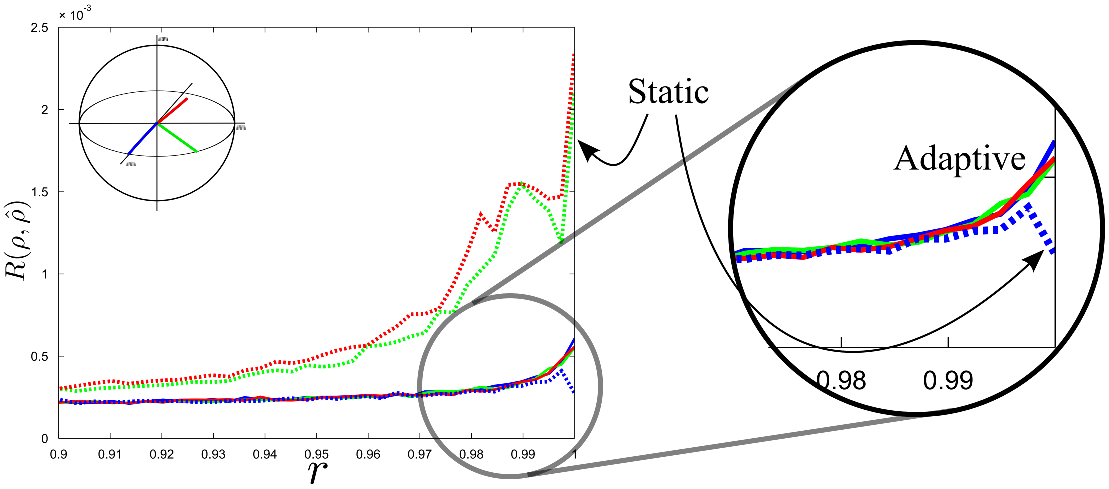

Figure 2.16: A comparison of static versus adaptive tomography for $N = 10^4$ measurements in each basis (plotted is an average of $N = 3 \times 10^3$ randomly selected data sets). The worst-case performance of any static protocol will be along the "noisiest" axis (the one directed toward the vertices of the Bloch cube in red) and have risk $O(1/\sqrt{N})$. Asymptotically equal is the risk of states on the rays 45^o from the measurement axes (green). The measurement axes behave as "noiseless" coins with $O(1/N)$ risk (blue). For our adaptive protocol, all states have risk scaling $O(1/N)$.

In the classical context [50], the least favorable priors are no longer discrete and actually turn out to be uniquely Jeffrey's prior. This is nice because, for one, discrete priors are conceptually pathological – no sane "agent" is going to assign such a prior. Moreover, there are many contexts in which Jeffrey's prior has been shown to be the unique choice [121, 95]. It would be interesting to know if the same phenomenon occurs in the quantum mechanical setting.

Chapter 3

Quantum Process Estimation

In Chapter 2 we used the following quantum mechanical model:

1. To each physical system, we assign an inner product space $\mathcal{H}$ with dimension $d < \infty$.
2. Each preparation procedure is represented by a positive semi-definite matrix ρ with $\mathrm{Tr}(\rho) = 1$.
3. Each measurement procedure is represented by set of positive semi-definite matrices $\{E_k\}$ where each E_k represents an outcome of the measurement and the set satisfies $\sum_k E_k = \mathbb{1}$.
4. The probability of outcome k given preparation ρ is

$$\Pr(E_k|\rho) = \mathrm{Tr}(E_k\rho). \tag{3.1}$$

The operational axioms above, although sufficient to obtain predictions provided one knows the state immediately before the measurement, leaves out the possibility of *transformations*. The idea is that one knows some initial preparation and final measurement but something happens to the system in the time between the two. In quantum theory this something takes the form of a *completely positive and trace preserving* (CPTP) map. A positive trace preserving map is a linear transformation taking density matrices to density matrices. Such a map is *completely* positive if the map induced when a second arbitrary system is added remains positive.

A CPTP map is vector in the space of linear operators acting on the vector space of linear operators acting on the Hilbert space $\mathcal{H}$ (call these *superoperators*). There are many ways to representation such an object. The usual Krauss decomposition and Stinespring dilation [163] are not unique and hence not very useful for characterizing quantum processes. There are two useful (invertible) representations which revolve around the *standard basis* of d^2 linear operators $E_{ij} = |i\rangle\langle j|$, where the $|i\rangle$ are the standard orthonormal basis for $\mathcal{H}$, and the *Choi matrix* $\Phi_{ik,jl} = \langle i|\Phi(|k\rangle\langle l|)|j\rangle$, for some superoperator Φ.

The *natural, linear* or *Liouville representation* is

$$K(\Phi) = \sum_{ijkl} \Phi_{ik,jl} E_{ik} \otimes E_{jl},$$

and is defined such that $K(\Phi)\text{vec}(\rho) = \text{vec}(\Phi(\rho))$. Note that $K(\Phi)$ is an operator acting on $\mathcal{H} \otimes \mathcal{H}$. The action of the superoperator is recovered via

$$\Phi(\rho) = \sum_{ijkl} \Phi_{ik,jl} E_{ik} \rho E_{jl}^{\dagger}.$$

Had we chosen a different basis $\{B_\alpha\}$, instead of $\{E_{ij}\}$, such that

$$\Phi(\rho) = \sum_{\alpha\beta} \chi_{\alpha\beta} B_\alpha \rho B_\beta^{\dagger},$$

then

$$K(\Phi) = \sum_{\alpha\beta} \chi_{\beta\alpha} B_\alpha \otimes \overline{B_\beta},$$

and the coefficients in this basis are together called the χ*-matrix*. References [49, 175] give an algorithm to determine the χ-matrix which was first experimentally determined in NMR [46] [1]. The Choi matrix was experimentally determined in references [27, 222].

The other useful representation is the *Choi-Jamiolkowski representation* and is given by

$$J(\Phi) = \sum_{ij} \Phi(E_{ij}) \otimes E_{ij}.$$

[1] This article reports the determination of the χ-matrix for NMR quantum computer running a controlled-NOT pulse sequence. The χ-matrix is used to compute the minimum and average gate fidelity. In order to characterize the decoherence process, the authors chose to fit χ to a model χ_m which has far fewer parameters and is motivated by the large body of NMR knowledge. The results show the model is in good agreement with the obtained data. Note that reference [162] is the first to report an experiment performing quantum process estimation, which was done for a qubit. See reference [27] for an example of dynamical process estimation in NMR. For a linear optical implementation, see [158].

The inversion is provided by the formula

$$\Phi(\rho) = \mathrm{Tr}_2[J(\Phi)(\mathbb{1} \otimes \rho^{\mathrm{T}})].$$

This representation was found to be useful in the characterizing the full set of CPTP maps for maximum likelihood estimation [77, 183, 166].

In references [140, 57] a simple overview of Choi's proof is given which shows how to construct the Kraus operators using what is now known as the Choi-matrix or Choi-Jamiolkowski isomorphism. The matrix

$$\mathcal{T}(\Phi) = \sum_{ij} E_{ij} \otimes \Phi(E_{ij}),$$

where $E_{ij} = |i\rangle\langle j|$ are elements of the standard basis. Note that this is a density matrix when Φ acts on one-half of the maximally entangled state

$$|\Phi\rangle = \frac{1}{\sqrt{d}} \sum_i |i\rangle \otimes |i\rangle \, .$$

This suggests the following recipe for quantum process estimation:

1. Prepare the maximally entangled state $|\Phi\rangle$.
2. Send one half of that state through Φ.
3. Determine by state estimation the joint density matrix $\mathcal{T}(\Phi)$.

The Kraus operators can then be determined by linear algebra. Note that determining the joint density matrix requires d^4 parameters (d^2 are constrained by trace preservation) which is in agreement with estimation schemes which determine the effect of Φ on d^2 basis states.

All these techniques rely on the ability to perform reliable quantum *state* estimation. But it is interesting to note that, while necessary for these schemes, optimal state estimation is not sufficient for optimal process estimation. Since state estimation will never be perfect, statistical errors alone might suggest a non-CPTP map has been performed. However, there is an added level of complexity here at the fundamental level since there exist physical maps which are non-CPTP (as noted in textbook accounts [163] and by experimental evidence [222]). So we cannot simply slough off those estimates which are non-CPTP as we did with negative and zero eigenvalue estimates in state estimation. Thus there is much theoretical work to be done in order to determine what optimal means for process estimation.

In any case, what has been discussed is *full* quantum process estimation – there really is a "black box" we know *nothing* about. While conservative, this is highly unlikely – especially if *we* are the ones trying to build a quantum information processing device, rather than being given an unlabeled one. That is, we can assume we have some additional knowledge of the process. When this additional knowledge reduces the number of parameters of the process to specify it is sometimes called *partial* process estimation or partial tomography. If our goal is to build a quantum information processing device, we must consider also an additional complication; a characterization of a process at a "snap-shot" in time is not nearly as useful as a characterization of the *dynamics* a quantum system undergoes (the latter can considered as a snap-shot at *all* times). Since, for closed systems, evolution is given by a Hamiltonian operation, this is usually called *Hamiltonian estimation*.

One way to adapt the above schemes to Hamiltonian estimation is by stroboscopically estimating snap-shots of the process at fixed times and then use various algorithms to invert these to find the Hamiltonian (see [164] and references therein). At each time, an *ensemble* measurement is performed, whereas here we take a different, more general, approach. We allow the possibility to estimate the Hamiltonian after every single *projective* measurement. Moreover, we also consider adaptively changing the measurement parameters between run of the experiment. We do so in a fully parametric Bayesian framework.

The outline of this chapter is as follows. First, in section 3.1, we review the framework of Bayesian experimental design and apply it to a simple Hamiltonian model for a qubit. In section 3.2 we use statistical techniques to provide analytic and asymptotic lower bounds on the derived estimation protocols. In section 3.3 we apply sequential Monte Carlo techniques from machine learning to more difficult, higher dimensional parameter space, examples. Finally, in section 3.4, we discuss the viability of the derived algorithms for use in generic Hamiltonian estimation problems.

This chapter is based on work presented in references [72, 71] and unpublished results obtained in collaboration with Chris Granade and David Cory.

3.1 Bayesian Experimental Design

Bayesian experimental design (see, e.g. [146]) is a methodology to ascertain the utility of a proposed experiment. Bayesian experimental design has been successfully applied to problems in experimental physics, such as in the recent examples of [64] and [217].

In classical theories of physics and statistics, the measurement simply reveals the state of the system at that instant. By contrast, quantum theory presents with the following physical (and conceptual) barrier: no single measurement can reveal the state. Rather, each potential kind of experiment admits a probability distribution from which we draw our data. Thus, the methodology of experimental design seems tailor-made for quantum theory.

The structure of this section is as follows. We begin by reviewing the general outline of Bayesian experimental design. We then apply the technique to devise an algorithm for the estimation of quantum Hamiltonian parameters. We show that in a particular case, this strategy is nearly globally optimal and demonstrate its improvement over standard algorithms numerically. Finally we conclude with a discussion on the applicability of this technique to real experiments on more complex quantum systems.

3.1.1 Formalism

We assume some initial experiment E has been performed and data D has been obtained. The goal is to determine $\Pr(\Theta|D,E)$, the probability distribution of the model parameters Θ given the experimental data. To achieve this we use Bayes' rule

$$\Pr(\Theta|D,E) = \frac{\Pr(D|\Theta,E)\Pr(\Theta|E)}{\Pr(D|E)},$$

where $\Pr(D|\Theta,E)$ is the *likelihood function*, which is determined through the process of modeling the experiment, and $\Pr(\Theta|E)$ is the *prior*, which encodes any *a priori* knowledge of the model parameters. The final term $\Pr(D|E)$ can simply be thought as a normalization factor.

At this stage we can stop or obtain further data. Experimental design is well suited to quantum theory since an arbitrary fixed measurement procedure does not give maximal knowledge as is often assumed in the statistical modeling of classical system. We conceive, then, of possible future data D_1 obtained from a, possibly different, experiment E_1. The probability of obtaining this data can be computed from the distributions at hand via marginalizing over model parameters

$$\Pr(D_1|E_1,D,E) = \int \Pr(D_1|\Theta,E_1)\Pr(\Theta|D,E)d\Theta.$$

We can use this distribution to calculate the expected *utility* of an experiment

$$U(E_1) = \sum_{D_1} \Pr(D_1|E_1,D,E)U(D_1,E_1),$$

where $U(D_1, E_1)$ is the utility we would derive if experiment E_1 gave result D_1. This could in principle be any function tailored to the specific problem. However, for scientific inference, a generally well motivated measure of utility is *information gain* [141]. In information theory, information is measured by the entropy

$$U(D_1, E_1) = \int \Pr(\Theta|D_1, E_1, D, E) \log \Pr(\Theta|D_1, E_1, D, E) d\Theta.$$

Thus, we search for the experiment which maximizes the expected information in the final distribution. That is, an optimal experiment $\hat{E}$ is one which satisfies

$$U(\hat{E}) = \max_{E_1} \Big\{ \sum_{D_1} \Pr(D_1|E_1, D, E) \times \\ \int \Pr(\Theta|D_1, E_1, D, E) \log \Pr(\Theta|D_1, E_1, D, E) d\Theta \Big\}.$$

3.1.2 Application to Simple Example

As an example of how to apply the Bayesian experimental design formalism to problems in quantum information, we consider a simple situation with a single qubit. In particular, we suppose that the qubit evolves under an internal Hamiltonian

$$H = \frac{\omega}{2}\sigma_z.$$

Here ω is an unknown parameter whose value we want to estimate. An experiment consists of preparing a single known input state $\psi_{\text{in}} = |+\rangle$, the $+1$ eigenstate of σ_x, evolving under the Hamiltonian H for a controllable time t and performing a measurement in the σ_x basis. This is the simplest problem where adaptive Hamiltonian estimation can be used and is the problem studied in reference [193].

In the language of Bayesian inference, the data $D \in \{0, 1\}$ is the outcome of the measurement. An experiment E consists of a specification of time the t that the Hamiltonian is on, while the model parameter Θ is simply ω. The likelihood function is given by the Born rule

$$\Pr(D = 0|\Theta, E) = \left| \langle +| \, e^{i\frac{\omega}{2}\sigma_z t} \, |+\rangle \right|^2 = \cos^2\left(\frac{\omega}{2}t\right). \tag{3.2}$$

This distribution is plotted in figure 3.1.

Experimental design is a decision theoretic problem based on the utility function

$$U(t) = \sum_D \Pr(D|t) \int \Pr(\omega|D, t) \log \Pr(\omega|D, t) d\omega.$$

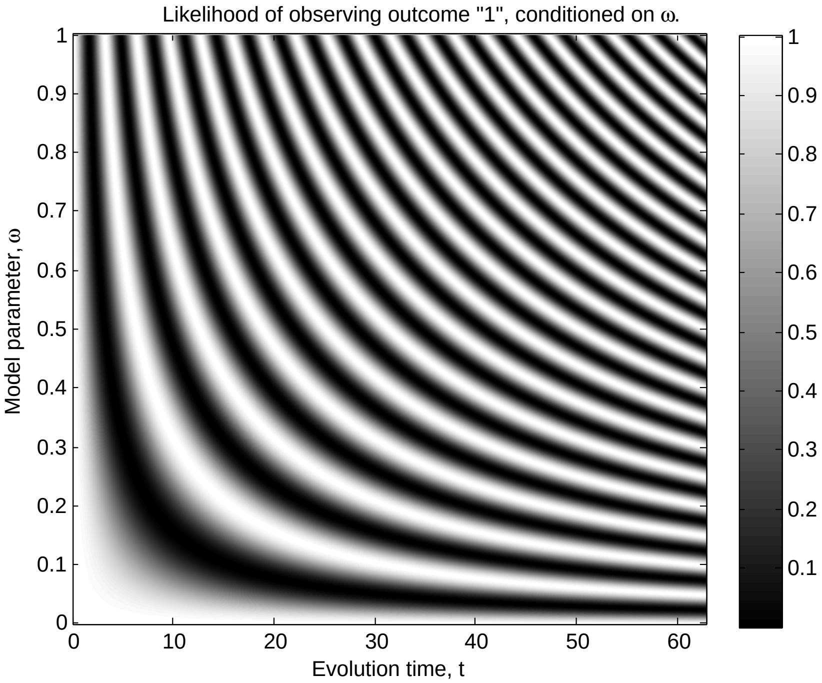

Figure 3.1: The likelihood function (3.2). Usually we think of a probability distribution as a function of the variable on the *left* of the conditional and for a fixed set of parameters on the *right*. However, the terminology "likelihood function" implies the opposite interpretation. Here we have fixed the datum $D = 1$ and plotted the probability distribution for variable ω and t.

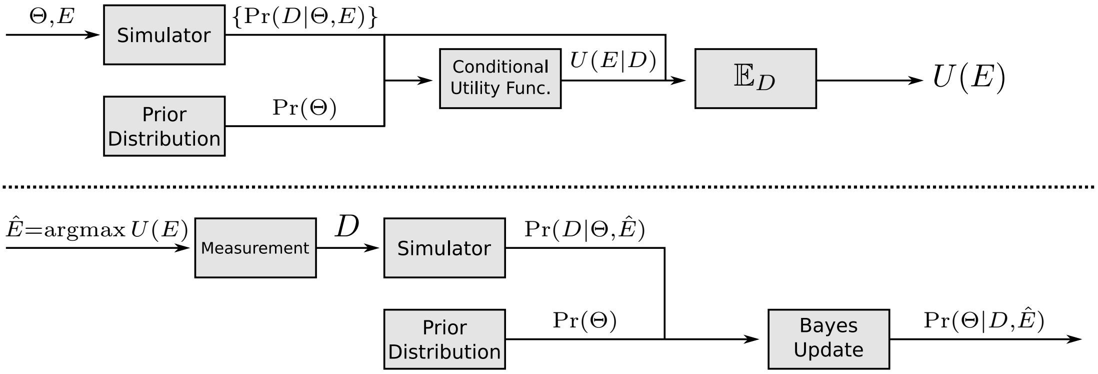

Figure 3.2: Overview of a step in the online adaptive algorithm for finding locally optimal experiments. Top: Method for calculating the utility function $U(E)$, given a simulator and a prior distribution $\Pr(\Theta)$ over model parameters Θ. Bottom: Method for updating prior distribution with results D from chosen actual experiment.

The optimal design is any value of t which maximizes this quantity. We proceed by performing the optimal experiment and obtaining data D_1. Using Bayesian inference we update our prior $\Pr(\omega)$ via Bayes' rule:

$$\Pr(\omega|D_1) = \frac{\Pr(D_1|\omega)\Pr(\omega)}{\Pr(D_1)}.$$

If we are not satisfied, we can repeat the process where this distribution becomes the prior for the new experimental design step. This algorithm is depicted in figure 3.2.

3.1.3 Estimators, Squared Error Loss and a Greedy Alternative to Information Gain

The preceding problem had a single unknown variable. If we desire an estimate $\hat{\Theta}$ of the true value Θ, the most often used figure of merit is the *squared error loss*:

$$L(\Theta, \hat{\Theta}) = \left|\Theta - \hat{\Theta}\right|^2.$$

The *risk* of an estimator $\hat{\Theta} : \{D, D_1, D_2, \ldots, D_N\} \mapsto \mathbb{R}$ is its expected performance with respect to the loss function:

$$R(\Theta, \hat{\Theta}) = \sum_{\{D, D_1, D_2, \ldots, D_N\}} \Pr(\{D, D_1, D_2, \ldots, D_N\}|\Theta) L(\Theta, \hat{\Theta}).$$

For squared error loss, the risk is also called the *mean squared error*. The average of this quantity with respect to some prior $\Pr(\Theta) =: \pi(\Theta)$ is the *Bayes risk* of π,

$$r(\pi, \hat{\Theta}) = \int R(\Theta, \hat{\Theta})\pi(\Theta)d\Theta,$$

and the estimator which minimizes this quantity is called a *Bayes estimator*. In this case the Bayes estimator is the mean of the posterior distribution[2]. Let us assume then that the estimators we choose are Bayes. Let us also choose a uniform prior for Θ. Then, the final figure of merit is the average mean squared error (AMSE):

$$r = \int R(\Theta, \hat{\Theta})d\Theta.$$

We would like a strategy which minimizes this quantity. Non-adaptive Fourier and Bayesian strategies were investigated and compared to an adaptive strategy in reference [193]. Their adaptive strategy fits into the Bayesian experimental design framework when the utility is measured by the *variance* of the posterior distribution:

$$V(D_1, E_1) = -\int \Pr(\Theta|D_1, E_1, D, E)(\Theta^2 - \mu(D_1, E_1))^2 d\Theta,$$

where

$$\mu(D_1, E_1) = \int \Pr(\Theta|D_1, E_1, D, E)\Theta d\Theta$$

is the mean of the posterior. Recall that the mean is a Bayes estimator of AMSE, so $\mu = \hat{\Theta}$. For a single measurement this utility function satisfies $V = -r$. That is, maximizing the utility *locally* at each step of the algorithm is equivalent to minimizing the AMSE at each step. Hence, when using the negative variance as our utility function, the adaptive strategy summarized in figure 3.2 is an example not only of a local optimization, but also a *greedy algorithm* with respect to the AMSE risk.

As experiments are designed and measurements are made, a *decision tree* is built up (a cartoon of this is shown in figure 3.3). We can also write the risk of this strategy recursively as follows. Suppose at the N'th, and final, measurement we have the updated distribution π_{N-1}. Then, the risk of the local strategy is

$$l_N(\pi_{N-1}, \Theta) = \sum_{D_N} \Pr(D_N|\Theta, \hat{E}_N)L(\Theta, \mu(D_N, \hat{E}_N)),$$

[2] Note that in any case where the loss function is *strictly proper*, i.e. is equal to zero if and only if the estimate is equal to the true state, the Bayes estimator is the posterior mean [24].

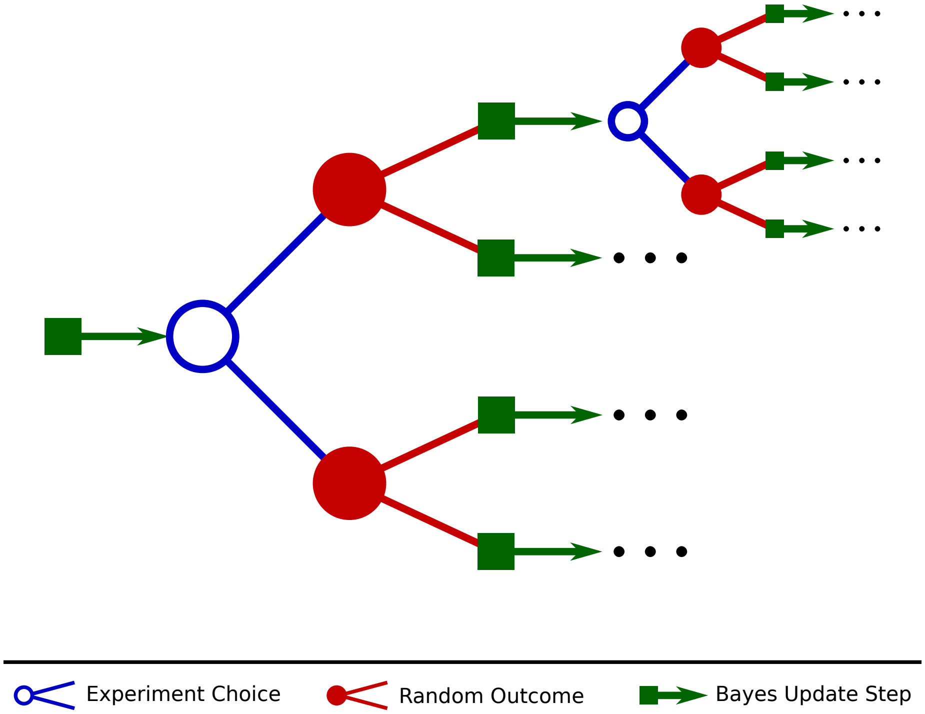

Figure 3.3: Example decision tree for experiment design choice.

where $\hat{E}_N$ is the locally optimal design satisfying

$$\hat{E}_N = \operatorname*{argmin}_{E_N} \int \sum_{D_N} \Pr(D_N|\Theta, E_N) L(\Theta, \mu(D_N, E_N))) \pi_{N-1}(\Theta) d\Theta.$$

The expected risk at any other stage is

$$l_n(\pi_{n-1}, \Theta) = \sum_{D_n} \Pr(D_n|\hat{E}_n) l_{n+1} \left(\frac{\Pr(D_n|\Theta, \hat{E}_n)\pi_{n-1}(\Theta)}{\int \Pr(D_n|\Theta, \hat{E}_n)\pi_{n-1}(\Theta) d\Theta} \right),$$

where $\hat{E}_n$ is, again, the locally optimal design satisfying

$$\hat{E}_n = \operatorname*{argmin}_{E_n} \int \sum_{D_n} \Pr(D_n|\Theta, E_n) L(\Theta, \mu(D_n, E_n))) \pi_{n-1}(\Theta) d\Theta.$$

Then, the Bayes risk of the greedy strategy is

$$\int l_1(\pi_0, \Theta) \pi_0(\Theta) d\Theta.$$

Again, it is clear that the greedy algorithm is globally optimal on the final decision, as there is no further hypothetical data to consider. That is, the optimal solution at the N'th measurement is

$$g_N(\pi_{N-1}, \Theta) = \sum_{D_N} \Pr(D_N|\Theta, \hat{E}_N) L(\Theta, \mu(D_N, \hat{E}_N)),$$

where $\hat{E}_N$ is the locally optimal design satisfying

$$\hat{E}_N = \operatorname*{argmin}_{E_N} \int \sum_{D_N} \Pr(D_N|\Theta, E_N) L(\Theta, \mu(D_N, E_N))) \pi_{N-1}(\Theta) d\Theta.$$

However, the globally optimal risk at any other stage

$$g_n(\pi_{n-1}, \Theta) = \sum_{D_n} \Pr(D_n|\tilde{E}_n) g_{n+1} \left(\frac{\Pr(D_n|\Theta, \tilde{E}_n)\pi_{n-1}(\Theta)}{\int \Pr(D_n|\Theta, \tilde{E}_n)\pi_{n-1}(\Theta) d\Theta} \right),$$

where now $\tilde{E}_n$ is the globally optimal design satisfying

$$\tilde{E}_n = \operatorname*{argmin}_{E_n} \int \sum_{D_n} \Pr(D_n|\Theta, E_n) g_{n+1} \left(\frac{\Pr(D_n|\Theta, E_n)\pi_{n-1}(\Theta)}{\int \Pr(D_n|\Theta, E_n)\pi_{n-1}(\Theta) d\Theta} \right) \pi_{n-1}(\Theta) d\Theta.$$

Then, the Bayes risk of the greedy strategy is

$$\int g_1(\pi_0, \Theta)\pi_0(\Theta)d\Theta.$$

In general, $l_1(\pi_0, \Theta) \neq g_1(\pi_0, \Theta)$. Nor is it the case that

$$\int l_1(\pi_0, \Theta)\pi_0(\Theta)d\Theta = \int g_1(\pi_0, \Theta)\pi_0(\Theta)d\Theta$$

for an arbitrary prior. However, for the special case of the uniform prior, we have found numerically that the Bayes risk of the greedy strategy and the Bayes risk of the global strategy are similar enough that the greedy strategy is useful.

3.1.4 Performance Comparisons

In reference [193], it was shown via simulation that the posterior variance of the greedy strategy is best fit by an exponentially decreasing function of N, the total number of measurements. In contrast, all off-line strategies decrease at best as a power of N.

In Figure 3.4, we show that the local information gain optimizing algorithm also enjoys an exponential improvement in accuracy over naive off-line methods. Moreover, we show Nyquist rate sampling is unnecessary and, indeed, sub-optimal. All results stated are obtained using a uniform prior on $[0, 1]$ and are computed numerically by exploring every branch of the decision tree, in contrast to simulation.

In order to be "fair" to the off-line methods, we restricted the adaptive methods to explore the same experimental design specifications. That is, for this particular problem, the adaptive algorithm was allowed to select measurement times from $[0, N_{\max}\pi]$, where $N_{\max}$ is the total number of measurements. In principle, these methods could only do better with a larger design specification.

3.1.5 Nyquist Considered Harmful

We have shown for the problem of estimating the parameter in a simple Hamiltonian model of qubit dynamics an adaptive measurement strategy can exponentially improve the accuracy over offline estimation strategies. Moreover, we have shown that sampling at the Nyquist rate is not optimal in the case of strong measurement. We have derived

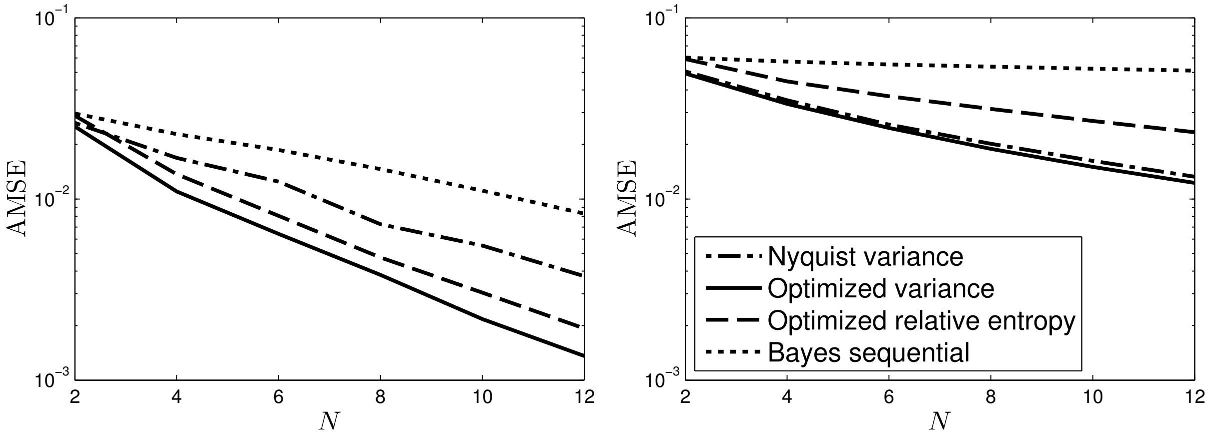

Figure 3.4: Performance of the estimation strategies. The Bayesian sequential and the strategy labeled "Nyquist" are sampled at the Nyquist rate. The "optimized" strategies find the global maximum utility (using Matlab's "fmincon" starting with the optimal Nyquist time). In each case, $N_{\max} = 12$ measurements are considered. Left: the ideal model with no noise. Right: a more realistic model with 25% noise and an addition relaxation process (known as T_2) which exponentially decays the signal (to half its value at $t = 10\pi$).

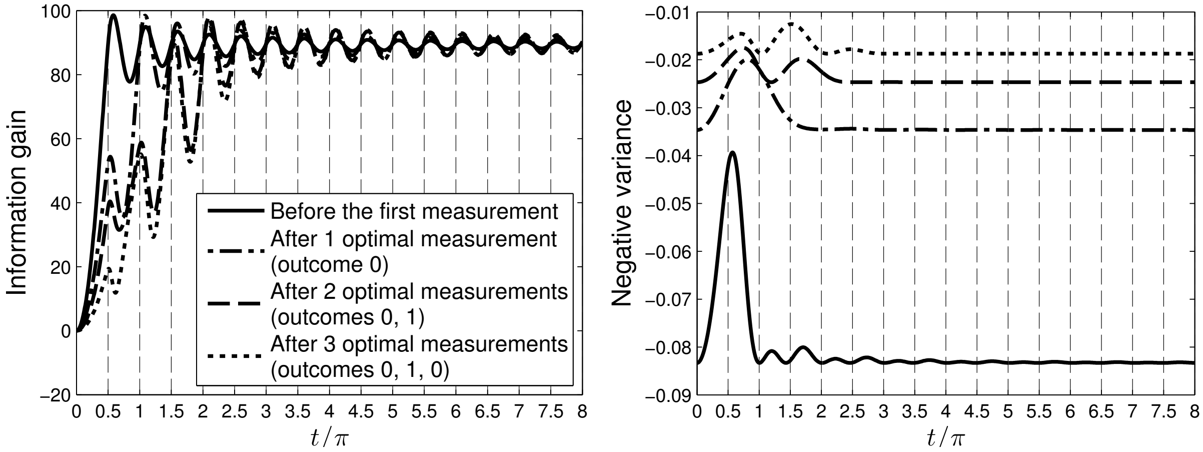

Figure 3.5: The information gain (left) and variance (right) utilities for the prior followed by three simulated measurements. The vertical grid lines indicate the Nyquist times. Note that the times at which the utilities are maximized do not necessarily increase with the number of measurements.

a recursive solution to the risks for both the local and global optimal strategies. Using this solution, we numerically found that the local strategy is nearly optimal in the special case of a uniform prior. That the greedy algorithm is nearly optimal in a case relevant to experiment demonstrates that an adaptive Bayesian method may be computationally feasible, in that an implementation need not consider all possible future data when choosing each experiment.

Together, these results demonstrate the usefulness of an adaptive Bayesian algorithm for parameter estimation in quantum mechanical systems, especially in comparison with other algorithms in common use. In the presence of noise, this improvement becomes still more stark, as demonstrated by the results shown in Figure 3.4.

Why is it the case that the Nyquist times are not optimal? First, why should we expect them to be optimal? The Nyquist theorem states that a signal which contains no frequencies higher than $\omega_{\max}$ is completely and unambiguously characterized by a discrete set of samples taken at a rate greater than or equal to $\pi/2\omega_{\max}$. However, the classical notion of sampling fails for the strong-measurement case that we consider here. What we have is a periodic *probability* distribution which can be sampled, not a periodic function whose *values* can be ascertained. That is, there is no *signal*, in the classical sense of the word, which can be reconstructed. The failure of the Nyquist rate sampling is exemplified in Figure 3.5.

3.2 A Lower Bound on Local Strategies

In the last section and in reference [193], measurement adaptive tomography was suggested as an efficient means of performing partial quantum process tomography. Little is known about optimal protocols when realistic experimental restrictions are imposed — as opposed to the case where one is allowed arbitrary quantum resources[3]. Indeed, even in the simplest example we have looked at above, not even bounds have been given on the proposed protocols. Here, we give analytic bounds on both non-adaptive and adaptive estimation protocols for the simple Hamiltonian parameter estimation problem. Moreover, we derive estimation protocols which asymptotically achieve these bounds. Adaptive protocols are typically difficult to implement because a complex optimization problem must be solved after each measurement. We instead derive a heuristic that is easy to implement *and* achieves the exponentially improved asymptotic risk scaling of the optimal solution.

3.2.1 Mean Squared Error Lower Bound

Recall the problem specified in section 3.1.2. If we desire an estimate $\hat{\omega}$ of the true value ω, a commonly used figure of merit is the *squared error loss*:

$$L(\omega, \hat{\omega}) = |\omega - \hat{\omega}|^2 .$$

The *risk* of an estimator, which is a function that takes data sets $(D, T) := (\{d_k\}, \{t_k\})$ to estimates $\hat{\omega}(D, T)$, is its expected performance with respect to the loss function:

$$R(\omega, \hat{\omega}) = \sum_D \Pr(D|\omega, T) L(\omega, \hat{\omega}(D, T)).$$

For squared error loss, the risk is also called the *mean squared error* (MSE).

The difficulty here is that the random outcomes of the measurements are *not* identically distributed. In fact, since they depend on the measurement time, each one could be different. Although, asymptotic results exist for non-identically distributed random variables[4], these results are derived for insufficient statistics, such as the sample mean. Moreover, we desire to provide computationally tractable heuristics that permit useful estimates with a finite number of samples.

[3] As in the standard phase estimation protocol. See e.g. [47].

[4] The frequentist reference is [113], while a useful Bayesian reference is [223].

Although it is quite difficult to obtain exact expressions for the risk for arbitrary measurement times, in some cases we have obtained an asymptotically tight lower bound. For *unbiased* estimators, we can appeal to the Cramer-Rao bound [56]

$$R(\omega, \hat{\omega}) \geq \frac{1}{\mathcal{I}(\omega)}, \tag{3.3}$$

where

$$\mathcal{I}(\omega) = -\sum_D \Pr(D|\omega, T) \frac{\partial^2 \log(\Pr(D|\omega, T))}{\partial \omega^2} \tag{3.4}$$

is called the Fisher information. In our particular case, the Fisher information reduces to quite a simple form in

$$\mathcal{I}(\omega) = \sum_{k=1}^{N} t_k^2, \tag{3.5}$$

which is conveniently independent of ω. Let us derive this here. We show that for the simple model represented by the likelihood function presented in equation (3.2), the Fisher information reduces to the form claimed in (3.5). To show this, we first note that the likelihood for a vector $D = (d_1, d_2, \ldots, d_k)$ of observations at times $T = (t_1, t_2, \ldots, t_k)$ is given by a product of the likelihoods for each individual measurement,

$$\Pr(D|\omega, T) = \prod_k \Pr(d_k|\omega, t_k).$$

Thus, the log-likelihood function is simply a sum over the individual log-likelihoods. Since the derivative operator commutes with summation, we obtain that

$$\frac{\partial^2}{\partial \omega^2} \log \Pr(D|\omega, T) = \sum_k \frac{\partial^2}{\partial \omega^2} \log \Pr(d_k|\omega, t_k).$$

This in turn implies that the Fisher information for a vector of measurements is given by the sum for each measurement of that measurement's Fisher information.

To calculate the single-measurement Fisher information, we find the second derivative of the log-likelihood for a single measurement is given by

$$\frac{\partial^2}{\partial \omega^2} \log \Pr(d_k|\omega, t_k) = t_k^2 \frac{(2d_k - 1)(1 - 2d_k + \cos(\omega t_k))}{((2d_k - 1)\cos(\omega t_k) - 1)^2}.$$

Thus, we find that the single-measurement Fisher information is given by

$$\begin{aligned}
\mathcal{I}(\omega|t_k) &= -\sum_{d_k\in\{0,1\}} \Pr(d_k|\omega,t_k)\frac{\partial^2}{\partial\omega^2}\log\Pr(d_k|\omega,t_k) \\
&= t_k^2 \sum_{d_k\in\{0,1\}} \frac{(2d_k-1)\,(1-2d_k+\cos{(\omega t_k)})}{2\,(2d_k-1)\cos{(\omega t_k)}-2} \\
&= t_k^2.
\end{aligned}$$

We conclude that $\mathcal{I}(\omega|T)=\sum_k t_k^2$, as claimed. Later we show that this bound becomes exponentially suppressed when we include noise in our model. In general, this quantity is dependent on the true parameter ω.

The Bayesian solution considers the average of the risk, called the *Bayes risk*, with respect to some prior $\pi(\omega)$:

$$r(\pi,\hat{\omega}) = \int R(\omega,\hat{\omega})\pi(\omega)d\omega.$$

As in references [193], we choose a uniform prior for $\omega\in(0,1)$. Then, the final figure of merit is the *average* mean squared error:

$$r(\hat{\omega}) = \int R(\omega,\hat{\omega})d\omega.$$

The goal is to find a strategy which minimizes this quantity. Although there exist Bayesian generalizations of the Cramer-Rao bound [87], ours is independent of ω and thus remains unchanged by integration over the parameter space:

$$r(\hat{\omega}) \geq \frac{1}{\sum_{k=1}^N t_k^2}. \tag{3.6}$$

Note also that, in general, Bayesian Cramer-Rao bounds require fewer assumptions to derive than the standard (frequentist) bound. Although they are the same for this model, they differ for a more general model considered later. In broad strokes, the difference in practice between Bayesian and frequentist methods is averaging versus optimization. Below we demonstrate a heuristic strategy which draws from both methods to achieve the goal of determining the measurement times which give the lowest possible achievable bound on the Bayes risk (3.6).

As useful as the Bayesian Cramer-Rao lower bound (3.6) is, it is simple to see that it is not always achievable. We can obtain a lower bound by considering the best protocol we could possibly hope for in any two-outcome experiment. In such a protocol, one bit of experimental data provides exactly one bit of certainty about the parameter ω. If we learn the bits of ω in sequence, at each step k, our risk is upper bounded by the worst-case where all the remaining bits of ω are either all 0 or all 1. In either case, the error incurred by estimating a point between the two extremes is given by $\sum_{n=k+2}^{\infty} 2^{-n} = 2^{-(k+1)}$, leading to the best possible MSE after N measurements being $2^{-2(N+1)}$, even though we can make a smaller Cramer-Rao bound by choosing times that grow faster than this exponential function. Note that we can achieve risk scaling as 2^{-N} via the standard phase estimation protocol [47], but that this protocol requires quantum resources which are *not* part of our model.

3.2.2 Lower Bounds for Some Sampling Schemes

Let us consider a couple of examples for which the lower bound can be further simplified. First, consider the case when all the measurement times are the same. This is by far the simplest case, since the outcomes become identically distributed. Recall $\omega \in (0,1)$. Then, the measurement time should be less then the first Nyquist time, $t \leq \pi$, or the data will be consistent with more than one ω. That is, for $t > \pi$ (but less than 2π, say), the likelihood function will have two equally likely maxima. We minimize the risk, then, by choosing $t = \pi$. Then, the maximum likelihood estimator (MLE), for example, will be asymptotically efficient [137] achieving the Cramer-Rao lower bound

$$r(\hat{\omega}_{\mathrm{MLE}}) = \frac{2}{\pi^2 N} + O\left(\frac{1}{N^2}\right).$$

Now consider a uniform grid of times. Since $\omega \in (0,1)$, we should choose the Nyquist sampling rate: $t_k = k\pi$. Then, for any estimator $\hat{\omega}$ using data collected at these measurement times, the Cramer-Rao bound gives

$$r(\hat{\omega}) \geq \frac{6}{\pi^2 N(1+N)(1+2N)} = \frac{3}{\pi^2 N^3} + O\left(\frac{1}{N^4}\right).$$

Again, the maximum likelihood estimator will be asymptotically efficient. However, since the likelihood function will have many local maxima, the maximum likelihood estimator is non-trivial to find as gradient methods are not guaranteed to work. Bayesian estimators were derived in [193], where simulations yielded $\sim 1/N^3$ risk scaling which is asymptotically efficient.

Note that since we are considering a uniform spacing of times, we can apply a Fourier estimation technique without worrying about spectral aliasing introduced by non-uniformity [148]. That is, we apply the discrete Fourier transform and estimate the peak of the power spectrum. Since the resolution in the frequency domain is $1/N\triangle t$, we expect the Bayes risk to be

$$r(\hat{\omega}_{\text{Fourier}}) = \frac{1}{\pi^2 N^2}.$$

The sampling theorem requires that we sample from a *deterministic* function, not a probability distribution. In practice, this condition is often approximately satisfied by sampling some stable statistic such as the mean value of the distribution at each time. This can be achieved by measuring at the same time until a sufficiently accurate estimate of the mean at that time is obtained, then repeating this for many other times. But as we have shown, this method can be quadratically improved by performing every *single* measurement at a different time.

3.2.3 Lower Bound on Local Adaptive Strategies

It has been shown that Bayesian adaptive solutions lead to risk decreasing exponentially with the number of measurements [193]. However, these results are given by fits to numerical data. Here, we give an analytic lower bound on the risk of these protocols.

Our local (in time) Bayesian adaptive protocol can be described as follows: (1) begin with a uniform prior $\Pr(\omega)$ and determine the first measurement time $t_1 \approx 1.136\pi$ which minimizes the average (over the two possible outcomes) variance of the posterior distribution; (2) perform a measurement at t_1, record the outcome d_1, and update the distribution $\Pr(\omega) \mapsto \Pr(\omega|d_1, t_1)$ via Bayes' rule; (3) repeat step (1) replacing the current prior with the current posterior. Note that the expected variance in the posterior is the Bayes risk. Thus, the protocol attempts to minimize the risk assuming the next measurement is the last. Strategies that are local in this sense are called a *greedy* strategies, as opposed to strategies which attempt to minimize the risk over all future experiments.

For some choices of measurement times, including those given by the protocol above, the posterior will be approximately normally distributed[5]. This is guaranteed in the asymptotic limit, but the posterior distribution near its peak is also remarkably well approximated by a Gaussian after as few as 15 reasonably chosen measurements (we found a uniform grid $t_k = k\pi$ to be sufficient for "warming up" to the Gaussian approximation). Thus, we

[5] This is true asymptotically and higher order corrections can be used if required [223].

approximate the current distribution (at given some sufficiently long measurement record D) as

$$\Pr(\omega|D) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\omega-\mu)^2}{2\sigma^2}},$$

with some arbitrary mean μ and variance σ^2 implied by D. The expected posterior variance (which is equal to the Bayes risk) of the probability distribution of the next measurement is

$$r(t) = \sigma^2 \left(1 + \frac{t^2\sigma^2 \sin(\mu t)^2}{-e^{t^2\sigma^2} + \cos(\mu t)^2}\right), \tag{3.7}$$

which oscillates with frequency 2μ within an envelope $\sigma^2 \left(1 - t^2\sigma^2 e^{-t^2\sigma}\right)$. Next, we derive expressions for posterior distributions under the assumption of a normally-distributed prior, and then apply these expressions to show the asymptotic scaling of the Bayes risk. We also derive update rules that allow for expedient implementation of the greedy algorithm.

Under the assumption of a normally-distributed prior, all prior information about the parameter ω can be characterized by the mean μ and variance σ^2 of the prior distribution. Thus, we shall write our priors as $\Pr(\omega|\mu, \sigma^2)$ to reflect the assumption of normality. Then, the probability of obtaining a datum d at time t given such prior information is then given by

$$\Pr(d|t; \mu, \sigma^2) = \int_{-\infty}^{\infty} \Pr(d|t, \omega) \Pr(\omega|\mu, \sigma^2) d\omega = \frac{1}{4}\left(2 - (2d-1)\left(1 + e^{2i\mu t}\right) e^{-\frac{1}{2}t\left(\sigma^2 t + 2i\mu\right)}\right).$$

Applying Bayes' rule then produces the posterior distribution

$$\begin{aligned}
\Pr(\omega|d, t; \mu, \sigma^2) &= \frac{\Pr(\omega|\mu, \sigma^2) \Pr(d|t, \omega)}{\Pr(d|t; \mu, \sigma^2)} \\
&= \frac{\sqrt{\frac{2}{\pi}} e^{-\frac{(\mu-\omega)^2}{2\sigma^2}} ((1-2d)\cos(t\omega) + 1)}{\sigma\left(2 - (2d-1)\left(1 + e^{2i\mu t}\right) e^{-\frac{1}{2}t(\sigma^2 t + 2i\mu)}\right)}.
\end{aligned}$$

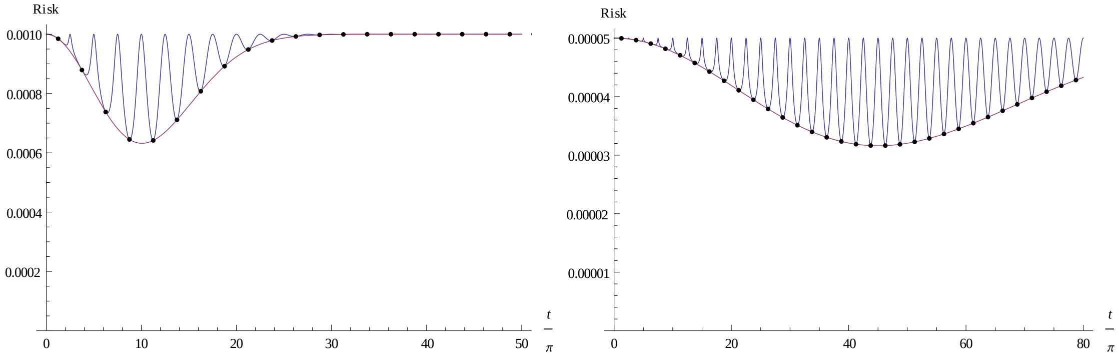

Figure 3.6: The risk envelope $E(t, \sigma^2)$, and the risk $r(t; \mu, \sigma^2) \geq E(t, \sigma^2)$ for the examples where $\mu = 0.4$ and $\sigma^2 = 10^{-3}$ (left) and $\sigma^2 = 5 \times 10^{-5}$ (right). Note that as σ^2 shrinks, the intersections between E and r (marked by dots) become more tightly packed.

The mean and variance of this distribution are given by:

$$\begin{aligned}
\mathbb{E}[\omega|d, t; \mu, \sigma^2] &= \frac{2\left((2d-1)e^{-\frac{1}{2}\sigma^2 t^2}\left(\sigma^2 t \sin(\mu t) - \mu\cos(\mu t)\right) + \mu\right)}{2 - (2d-1)\left(1 + e^{2i\mu t}\right)e^{-\frac{1}{2}t(\sigma^2 t + 2i\mu)}} \\
\mathbb{V}[\omega|d, t; \mu, \sigma^2] &= \mu^2 + \sigma^2 - \frac{2\left((2d-1)e^{-\frac{1}{2}\sigma^2 t^2}\left(\sigma^2 t \sin(\mu t) - \mu\cos(\mu t)\right) + \mu\right)}{2 - (2d-1)\left(1 + e^{2i\mu t}\right)e^{-\frac{1}{2}t(\sigma^2 t + 2i\mu)}} \\
&\quad - \frac{2(2d-1)\sigma^2 t e^{i\mu t}\left(\sigma^2 t\cos(\mu t) + 2\mu\sin(\mu t)\right)}{(2d-1)\left(1 + e^{2i\mu t}\right) - 2e^{\frac{1}{2}t(\sigma^2 t + 2i\mu)}}
\end{aligned}$$

To chose optimal times, we wish to pick t so as to minimize the expected value over of the variance, where this expectation is taken over possible data. Based on the previous expressions, we find that

$$\mathbb{E}_d[\mathbb{V}_\omega[\omega|d, t; \mu, \sigma^2]] = \sigma^2\left(1 + \frac{t^2\sigma^2\sin(\mu t)^2}{-e^{t^2\sigma^2} + \cos(\mu t)^2}\right),$$

in agreement with Equation (3.7).

This expected variance, which describes our risk incurred by measuring at a given t, is bounded below by an envelope $E(t, \sigma^2) = \sigma^2\left(1 - t^2\sigma^2 e^{-t^2\sigma^2}\right)$. A pair of examples of the envelope $E(t, \sigma^2)$ and achievable risk $r(t; \mu, \sigma^2)$ is illustrated in figure 3.6.

Note that the envelope is minimized by $\hat{t} = \underset{t}{\operatorname{argmin}}\, E(t, \sigma^2) = 1/\sigma$. Moreover, the expected variance saturates the lower bound at intervals in t of $1/\mu$, but the width of the envelope's minimum grows as $1/\sigma^2$, so that as more measurements are performed, the bound becomes a good approximation for the minimum achievable risk. Thus, in the asymptotic limit of large numbers of experiments, we have that the risk at scales with each step as the minimum of the envelope,

$$\frac{E(\hat{t}, \sigma^2)}{\sigma^2} = 1 - e^{-1} \approx 0.632.$$

We conclude that in the asymptotic limit, the risk decays as $e^{-N \ln 0.632} \approx e^{-0.458N}$, where N is the number of measurements performed, and this is achieved at measurement times which scale as

$$t_k \sim \frac{1}{\sigma(1 - e^{-1})^{k/2}} \approx \frac{1.26^k}{\sigma}.$$

These times are guaranteed to be optimal only in the asymptotic limit. For finite numbers of samples, we suggest two simple heuristics. First, we suggest the use of exponentially increasing times, where the base of the exponent is optimized offline, followed by the use of the maximum likelihood estimator for these times. Second, we suggest a simpler adaptive scheme based on the assumption that the distribution remains Gaussian after each measurement. Making use of this normality assumption, we only need update equations for the mean and variance of the distribution over ω. In deriving the update equations, we also take into account the oscillations of the expected Bayes risk by finding the nearest achievable minima to the one given by the lower bound. We state without derivation the update rules for μ and σ^2 after obtaining a measurement result d from an experiment performed at time t, under the assumption of an normal prior. For the simple model described by equation (3.2),

$$\mathbb{E}\left[\omega|d\right] = \mu - \frac{\pi(2d-1)\sigma^2(-1)^k\left(2k-1\right)\exp\left(-\frac{\pi^2\sigma^2(1-2k)^2}{8\mu^2}\right)}{2\mu} \tag{3.8}$$

$$\mathbb{V}\left[\omega|d\right] = \sigma^2 - \frac{\pi^2(1-2d)^2\sigma^4\left(1-2k\right)^2\exp\left(-\frac{\pi^2\sigma^2(1-2k)^2}{4\mu^2}\right)}{4\mu^2}, \tag{3.9}$$

where $k = \text{round}\left[\frac{\mu}{\pi\sigma} + \frac{1}{2}\right]$ is used to pick the intersection of $E(t, \sigma^2)$ and $r(t; \mu, \sigma^2)$ to the minimum of E.

3.2.4 Generalizations to Noise Models

In practice, we will have to consider not only experimental restrictions but also noise and relaxation processes. Processes which do not affect the quantum state can be effectively

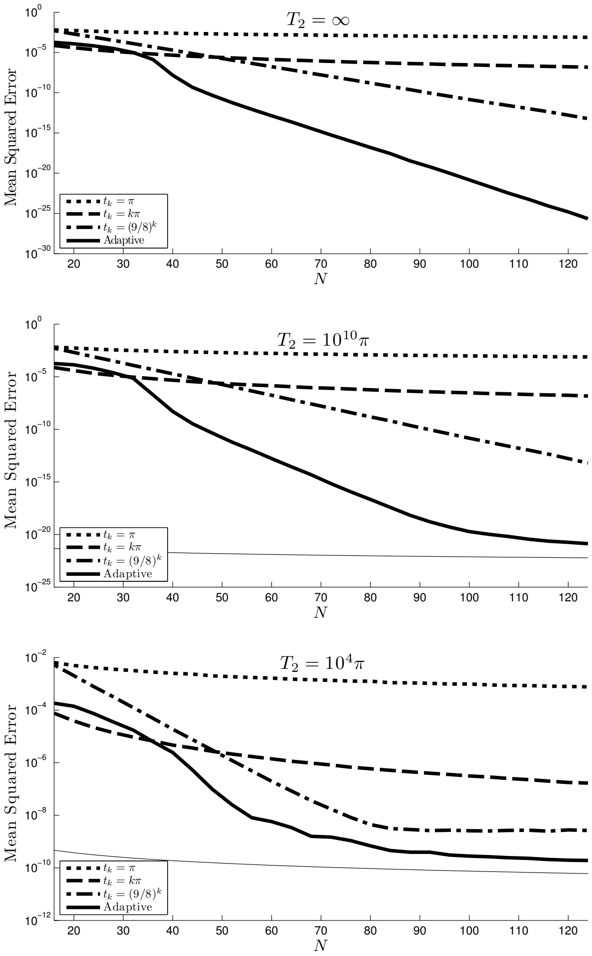

Figure 3.7: The Bayes risk – the average (over a uniform prior) mean (over data) squared error – of the strategies discussed in the thesis. Data points are at evenly spaced measurement numbers $N \in \{16, 20, 24, \ldots, 124\}$ and the lines are linear interpolants to guide the eye. Each data point is the average of 10^4 simulations. In each figure, the noise parameter $\eta = 1$ since its inclusion only gives a constant offset. From top to bottom, the relaxation characteristic time is $T_2 = \infty, 10^{10}\pi, 10^4\pi$. The thin solid lines indicate the lower bound given by Equation (3.12).

modeled by random bit-flip errors occurring with probability $1-\eta$. Processes which do affect the quantum state (decoherence) are modeled by an exponential decay of phase coherence[6] with characteristic time T_2. Since the state being measured lies in the xy-plane of the Bloch sphere, this loss of phase coherence manifests as an exponential decaying envelope being applied to the original likelihood (3.2). The model is thus fully specified by the likelihood function

$$\Pr(0|\omega, t, \eta, T_2) = \eta \left(e^{-\frac{t}{T_2}} \cos^2\left(\frac{\omega}{2} t\right) + \frac{1 - e^{-\frac{t}{T_2}}}{2} \right) + \frac{1-\eta}{2}. \tag{3.10}$$

For the model with finite T_2 and limited visibility, given by the likelihood function (3.10), we can follow the same logic. We find the second derivative of (3.10) with respect to ω gives us

$$\frac{\partial^2}{\partial \omega^2} \log \Pr(d_k|\omega, t_k) = \eta t_k^2 \cdot \frac{(2d_k - 1)\left(\eta\left(1 - 2d_k\right) + e^{\frac{t_k}{T_2}} \cos\left(\omega t_k\right)\right)}{\left(\eta\left(1 - 2d_k\right)\cos\left(\omega t_k\right) + e^{\frac{t_k}{T_2}}\right)^2}.$$

The expected value of this derivative then gives us the Fisher information for a single measurement in the finite-T_2 model,

$$\mathcal{I}(\omega|t_k) = \frac{\eta^2 t_k^2 \sin^2\left(\omega t_k\right)}{e^{\frac{2t_k}{T_2}} - \eta^2 \cos^2\left(\omega t_k\right)}.$$

Taking the sum of this information then produces the Cramer-Rao bound

$$R(\omega, \hat{\omega}) \geq \left(\sum_{k=1}^{N} \frac{t_k^2 \eta^2 \sin^2(\omega t_k)}{e^{\frac{2t_k}{T_2}} - \eta^2 \cos^2(\omega t_k)} \right)^{-1}. \tag{3.11}$$

Note that unlike the Cramer-Rao bound for the noiseless case, the above bound is not independent of ω and thus we must appeal to the Bayesian Cramer-Rao bound so that the measurement times can be chosen independently of the true parameter. However, the Bayesian bound turns out to be very loose. A sharper bound is given by first upper bounding each term in the denominator to give

$$r(\hat{\omega}) \geq \frac{1}{\eta^2 \sum_{k=1}^{N} t_k^2 e^{-\frac{2t_k}{T_2}}}.$$

[6] We do not include amplitude damping in our model since our populations remain equal throughout evolution and thus T_1 only manifests as a contribution to T_2.

The noise term (or visibility) η simply gives a constant reduction in the achievable accuracy. The relaxation process provides a more interesting dynamic as we see that the gains from longer times are exponentially suppressed. In other words, strategies are restricted to explore $t_k \leq T_2$. We can thus do no better than

$$r(\hat{\omega}) \geq \frac{e^2}{N\eta^2 T_2^2}. \tag{3.12}$$

The adaptive strategy discussed above can be generalized to include noise and relaxation. For the finite-T_2 model, the updated mean and variance are given by

$$\mathbb{E}\left[\omega|d\right] = \mu + \frac{\pi(2d-1)(-1)^k(2k-1)\sigma^2 \exp\left(-\frac{(\pi-2\pi k)\left(-2\pi k\sigma^2 T_2+4\mu+\pi\sigma^2 T_2\right)}{8\mu^2 T_2}\right)}{2\mu} \tag{3.13}$$

$$\mathbb{V}\left[\omega|d\right] = \sigma^2 - \frac{\pi^2(2d-1)^2(2k-1)^2\sigma^4 \exp\left(-\frac{(\pi-2\pi k)\left(-2\pi k\sigma^2 T_2+4\mu+\pi\sigma^2 T_2\right)}{4\mu^2 T_2}\right)}{4\mu^2}, \tag{3.14}$$

where in this case,

$$k = \operatorname{round}\left[\frac{\mu - \mu\sqrt{4\sigma^2 T_2^2+1}+\pi\sigma^2 T_2}{2\pi\sigma^2 T_2}\right].$$

To illustrate the performance of our adaptive strategy, we simulate the adaptive strategy along with offline strategies using identical times ($t_k = \pi$), linearly spaced times ($t_k = k\pi$) and exponentially sparse times ($t_k = (9/8)^k$). For each strategy, we perform simulations for experiments consisting of different numbers of samples N, up to $N = 124$, and repeat each such simulation 10^4 to obtain an estimate of the Bayes risk for that strategy and experiment size. In Fig. 3.7, we present the results of these simulations for the noiseless case, and for the cases $T_2 = 10^{10}\pi$ and $T_2 = 10^4\pi$.

Note that in all cases, the adaptive strategy achieves exponential scaling until the times selected reach $t = T_2$. At that point, the risk will then scale linearly if the remaining measurement times are $t = T_2$. However, if the protocol continues to select larger measurement times, the information gained from those measurements will tend to zero and the risk will remain constant.

3.3 Sequential Monte Carlo Algorithms

In this section we are going to consider the noise model from the previous section:

$$\begin{aligned}\Pr(0|\omega, T_2; t) &= e^{-\frac{t}{T_2}} \cos^2\left(\frac{\omega}{2} t\right) + \frac{1 - e^{-\frac{t}{T_2}}}{2}, \\ \Pr(1|\omega, T_2; t) &= 1 - \Pr(0|\omega, T_2; t),\end{aligned} \tag{3.15}$$

where now both ω and T_2 are unknown. The time t remains the only experimentally controllable parameter. We will find that, even for such a simple generalization as this, the methods discussed above are not adequate for this more general problem. We instead propose a Sequential Monte Carlo [62] algorithm that performs well with an experimentally viable amount of resources.

3.3.1 Cramer-Rao Bound

For one parameter models a commonly used figure of merit is the *squared error loss*: $L(\omega, \hat{\omega}) = |\omega - \hat{\omega}|^2$. Here we use a generalization called the *quadratic loss* which is suitable for a vector of parameters $\boldsymbol{x} := (\omega, T_2)$. It is defined as

$$L_{\boldsymbol{Q}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = (\boldsymbol{x} - \hat{\boldsymbol{x}})^{\mathrm{T}} \boldsymbol{Q} (\boldsymbol{x} - \hat{\boldsymbol{x}}),$$

where $\boldsymbol{Q}$ is a positive definite matrix on the space of unknown parameters. An *estimator* is a function $\hat{\boldsymbol{x}}$ that takes a set of observed data $D := \{d_k\}$ collected from a set of experiments with controls $C := \{t_k\}$ and produces a set of estimates for the unknown parameters $\boldsymbol{x}$. The *risk* of an estimator given a set of experiment designs C is its expected performance over all possible outcomes D with respect to the loss function:

$$R(\boldsymbol{x}, \hat{\boldsymbol{x}};\ C) = \mathbb{E}_{D|\boldsymbol{x};C}[L(\boldsymbol{x}, \hat{\boldsymbol{x}}(D, C))].$$

The Bayes risk is the average of this quantity with respect to a prior distribution on $\boldsymbol{x}$ and is given explicitly by

$$r(\pi; C) = \mathbb{E}_{\boldsymbol{x}}[R(\boldsymbol{x}, \hat{\boldsymbol{x}};\ C)]$$

where $\hat{\boldsymbol{x}}$ is assumed to be a Bayes estimator.

If we are to proceed by analogy with the single-parameter case, we want to prove asymptotic lower bounds on the risk $R(\boldsymbol{x}, \hat{\boldsymbol{x}};\ C)$ by finding the Fisher information. In the case of multiple parameters, the Fisher information is no longer a scalar, but a matrix defined by

$$\boldsymbol{I}(\boldsymbol{x}; C) = \mathbb{E}_{D|\boldsymbol{x};C}\left[\boldsymbol{\nabla}_{\boldsymbol{x}} \log\left(\Pr(D|\boldsymbol{x}; C)\right) \cdot \boldsymbol{\nabla}_{\boldsymbol{x}}^{\mathrm{T}} \log\left(\Pr(D|\boldsymbol{x}; C)\right)\right].$$

Importantly, the Fisher information does not depend at all on the prior distribution, and thus is calculated in the same way regardless of how many experiments have already been performed. The Cramer-Rao bound is then given by $\mathrm{Cov}(\hat{\boldsymbol{x}}) \geq \boldsymbol{I}(\boldsymbol{x}; C)^{-1}$, where $\boldsymbol{X} \geq \boldsymbol{Y}$ means that $\boldsymbol{X} - \boldsymbol{Y}$ is positive semi-definite. If we choose $\boldsymbol{Q} = \mathbb{1}$, then $R(\boldsymbol{x}, \hat{\boldsymbol{x}}; C) = \mathrm{Tr}(\mathrm{Cov}(\hat{\boldsymbol{x}})) \geq \mathrm{Tr}(\boldsymbol{I}(\boldsymbol{x}; C)^{-1})$. Clearly, this statement of the multivariate Cramer-Rao bound assumes that $\boldsymbol{I}$ is non-singular. Unfortunately, that assumption is not met for the given model in the case that we consider a single measurement.

Fortunately, we avoid this problem by considering the Bayesian information matrix $\boldsymbol{J}(\pi; C) = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{I}(\boldsymbol{x}; C)]$. However, note integration of these types are often intractable against analytic solutions. By numerically integrating, we can find lower bounds for specific values of t, T_2 and ω. Alternatively, we can apply an iterative algorithm[7] as follows. We will subscript the various quantities of interest by N, which means "at the N'th measurement". The Bayesian Cramer-Rao bound is given by

$$r(\pi; C_{N+1}) \geq \boldsymbol{J}_{N+1}(C_{N+1})^{-1}, \tag{3.16}$$

where the iteration is given by

$$\boldsymbol{J}_{N+1}(C_{N+1}) = \boldsymbol{J}(\pi; C_{N+1}) + \boldsymbol{J}_N(C_N)$$

and the initial condition is $\boldsymbol{J}_0 = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{\nabla}_{\boldsymbol{x}} \log\left(\pi(\boldsymbol{x})\right) \cdot \boldsymbol{\nabla}_{\boldsymbol{x}}^{\mathrm{T}} \log\left(\pi(\boldsymbol{x})\right)]$.

An additional complication is that the maximum likelihood estimator is not well-behaved in cases where the Fisher information matrix is singular. In particular, the likelihood function will *not* in general be locally normally distributed about its maximum. This also precludes making the assumption that our posterior approaches appropriate normality in the limit of many measurements, frustrating efforts at extending our update-rule heuristic to this more complicated case. However, next we will see that via a Sequential Monte Carlo algorithm we can efficiently compute Bayesian mean estimators and its respective lower bound.

[7] This type of algorithm as be given some attention recently for "state space models" and classical signals with additive noise [209].

3.3.2 Idea of Sequential Monte Carlo

Suppose we want to compute the integral of a function $f : \mathbb{R} \to \mathbb{R}$ with respect to some probability density p:

$$I = \int_a^b f(x)p(x)dx.$$

Then we could use various deterministic numerical integration techniques that we learned in our first year calculus courses. All of them break the interval $[a, b]$ in to smaller intervals: $[a, b] = [a, x_1) \cup [x_1, x_2) \cup \cdots \cup [x_{n-1}, b]$. Suppose for brevity that each of the n intervals is the same size: $\triangle x = (b - a)/n$. Then we approximate I, for example, by

$$\hat{I} = \sum_k f(x_k)p(x_k)\triangle x.$$

Using Taylor's theorem we can show that the leading order error term in such an approximation is $O(\triangle x^2)$ for each term. Since there are n terms, the error is bounded by n^{-1}. Using the same technique for a function $g : \mathbb{R}^2 \to \mathbb{R}$ will yield error scaling as $n^{-1/2}$, and so on, so that for a parameter space of dimension d, the error of deterministic numerical integration scales exponentially as $n^{-1/d}$. This is bad. And, for a long time it hindered progress of Bayesian techniques as the formulas above some times even require three nested integral expectations!

Monte Carlo methods attempt to remedy this situation by drastically reducing this error to a constant $n^{-1/2}$, in many cases. Now, general Monte Carlo methods have a rich and complicated history. However, we will be using a more recent variant which dates back to 1993 [91] but has been rediscover many times in a wide variety of scientific applications under the following names[8]: sequential Monte Carlo, particle filters, sequential importance sampling, Bayes filters, and so on. Fortunately, this field is yet to be plagued by the choice of name being decided by the charm of the acronym it can generate. For no good reason other than my personal proclivity, we will use the term sequential Monte Carlo. We will now sketch the idea behind the algorithm.

Recall the Bayes update rule for one datum D_1

$$\Pr(\boldsymbol{x}|D_1) \propto \Pr(D_1|\boldsymbol{x}) \Pr(\boldsymbol{x}).$$

[8] See, for example, [62] for a recent tutorial on these methods.

The distribution can be processed sequentially as more data arrive:

$$\begin{aligned}\Pr(\boldsymbol{x}|D_2, D_1) &\propto \Pr(D_2|\boldsymbol{x})\Pr(\boldsymbol{x}|D_1),\\ \Pr(\boldsymbol{x}|D_3, D_2, D_1) &\propto \Pr(D_3|\boldsymbol{x})\Pr(\boldsymbol{x}|D_2, D_1),\\ &\vdots\end{aligned}$$

Now it would be incredibly convenient, and possible more efficient, if the prior and posterior were both from a small family of probability densities. Such a class is a property of the likelihood function and is called its *conjugacy class* (see, for example, [19]). For example, a normally distributed prior is conjugate to a normal likelihood since its posterior will also be normal. Normality is often assumed for analytic and numeric tractability since we need only track a few parameters (mean and variance) which are sufficient to specify the normal distribution.

However, the likelihood functions that we will consider are not normal – not even asymptotically! But, it turns out that one simple class of prior is conjugate to *every* likelihood function: a Dirac delta-function, or more generally, a weighted sum of delta-functions. Using such a prior guarantees that at all times in the sequential Bayesian computation, we have the same delta-functions with possibly different weights. Not only does this greatly simply the analysis, it makes the required multidimensional integrals tractable.

Integration with respect to a sum of delta-functions becomes a simple summation over the finitely many support points. Moreover, and perhaps most importantly, this choice of prior can be made a good approximation to the correct distribution at every step in the computation. That is, we approximate an arbitrary distribution by

$$\Pr(\boldsymbol{x}|D) \approx \sum_{k=1}^{n} w_k(D)\delta(\boldsymbol{x} - \boldsymbol{x}_k),$$

where the weights at each step are iteratively calculated from the previous step via

$$w_k(D_{j+1}) = \sum_{k=1}^{n} \Pr(D_{j+1}|\boldsymbol{x}_k) w_k(D_j).$$

This will be a good approximation provided we feed in, at the initial stage, the appropriate weights $\{w_k\}$ and support points, often called *particles*, $\{\boldsymbol{x}_k\}$. Such a choice is to have equal weight, $w_k = 1/n$ for all k, and sampled particles according to the correct prior $\Pr(\boldsymbol{x})$.

Like all numerical techniques, this one requires its own bit of alchemy. The first problem any sequential Monte Carlo algorithm runs into is zero weights. This is doubly painful since we are effectively operating with fewer particles but using the same amount of computational resources. Since the support of our approximate distribution is a measure-zero set according to the correct distribution, all the weights will eventually be zero; we cannot avoid this. However, we can use various *resampling* techniques to postpone this disaster.

Generally, the idea behind resampling is to adaptively change the location of the particles to those which are most likely. The simplest of these types of algorithm chooses n particles (the original number), with replacement, according to the distribution of weights then reset the weights of all particles to $1/n$. Thus, zero weights particles are "moved" to higher weight locations.

Here we have used a more powerful general purpose resampling algorithm [9]. The idea is similar to above but that algorithm still potentially leaves large volumes of the parameter space unexplored. To remedy this, after the new particle is chosen, we apply a random perturbation to the particle location. Thus, when the same (probably high-weight) particle is chosen multiple times, the new particles are randomly spread around that location in the parameter space. Looking ahead at figure 3.8, we can see how the locations of the particles move along with the higher likelihood regions of the parameter space.

Using this method we can efficiently compute the require expectations in the experiment design protocols and Bayesian Cramer-Rao bounds since all integral expectations become summations. To illustrate how this algorithm works let us go back to the simple model with likelihood function given in equation (3.10). An example trial run is plotted in figure 3.8, for the case of known T_2, and in figure 3.9 for unknown T_2. The frequency of resampling is tunable but also depends on the random outcomes. However, in figures 3.8 and 3.9, we have forced the resampling before plotting so that each particle has the same weight.

3.3.3 Proof of Concept

Here we demonstrate the utility of the sequential Monte Carlo algorithm through its performance on our example problems. The model is that of equation (3.15). First, with known $T_2 = 100\pi$, the performance is plotted in figure 3.10 for a variable number of particles. As expected, the more Monte Carlo samples used, the more accurate the estimates of integrals and, hence, the more accurate the protocol. The same is true in figure 3.11, where the performance is plotted for unknown ω and unknown T_2.

[9] The details are well documented in reference [142].

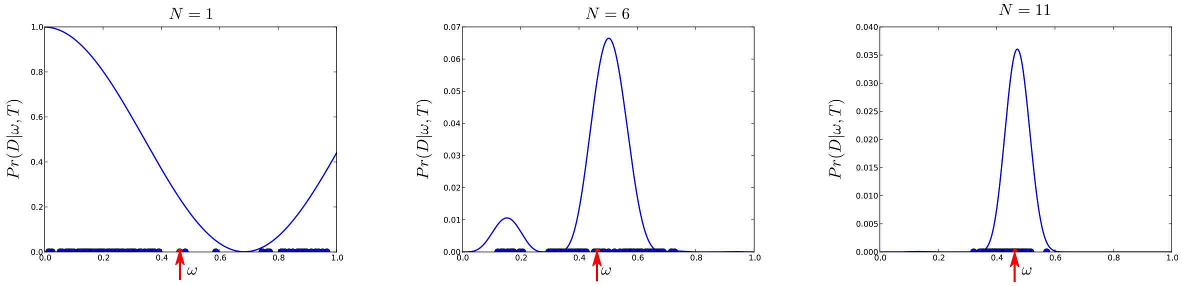

Figure 3.8: Left to right: the likelihood function for $N = 1, 6$, and 11 simulated measurements at random times in in $(0, 5\pi)$. The model is that given in equation (3.10) with $\eta = 1$ and $T_2 = 100\pi$. The red dot (and red arrow) is the randomly chosen true parameter ω. The blue dots are the $n = 100$ sequential Monte Carlo "particles".

3.4 Discussion and Future Directions

There are a few details to address under the rug in the discussion of the efficiency of the sequential Monte Carlo algorithm. The first thing to note is that, since the choice of resampling algorithm is usually tailored to the problem at hand, it is hard to say something in general about the algorithmic complexity of it. A more pressing issue for us, however, is that this algorithm cannot possibly be efficient in the number of physical systems we are trying to estimate the parameters of. This is simply because the calculation of the likelihood function requires a quantum simulation, which is an exponentially difficult problem in the number of physical systems.

Essentially, what we need to do is reduce the number of evaluations of the likelihood function. There are a number of options to consider. First, we can use a different utility function, such as information gain, rather than a variance reduction algorithm. Computing information gain requires no posterior updates to compute whereas computing the variance requires a hypothetical update of the posterior at each potential experiment[10]. We have been using the latter as we have been benchmarking the performance of our algorithm with a quadratic loss function. This is standard in statistical theory due to its various mathematical conveniences. However, it does not seem to posses any operational motivation. Thus, although maximizing information gain will be suboptimal with respect to quadratic loss, it might be optimal with respect to a more relevant, an operationally motivated, information theoretic loss function (such as relative entropy).

Using an alternative utility function can be seen as an approximation, or not. However,

[10] This preference was espoused also in reference [118] in the context of quantum state tomography.

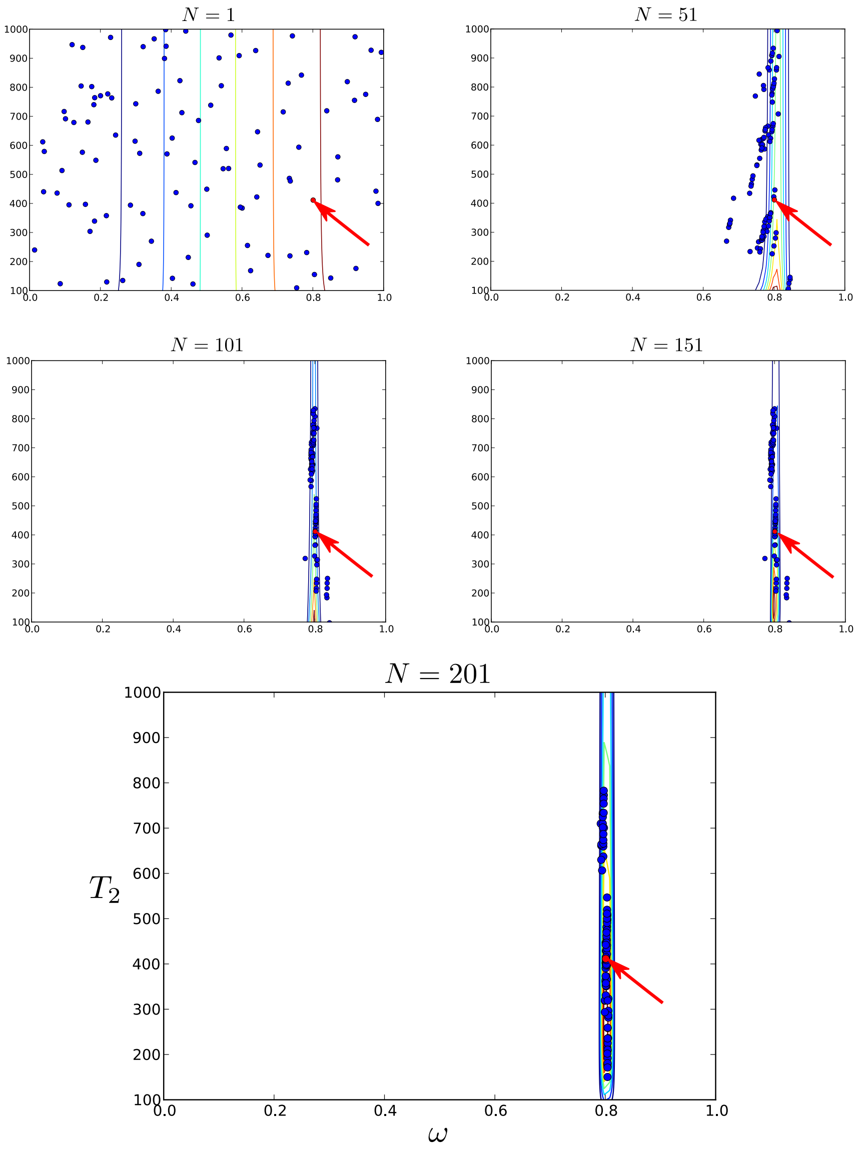

Figure 3.9: The likelihood function for $N = 1, 51, 101, 151$ and 201 simulated measurements at random times in in $(0, 20\pi)$. The model is that given in equation (3.10) with $\eta = 1$. The red dot (and red arrow) is the randomly chosen true parameter $\boldsymbol{x} = (\omega, T_2)$. The blue dots are the $n = 100$ sequential Monte Carlo “particles”.

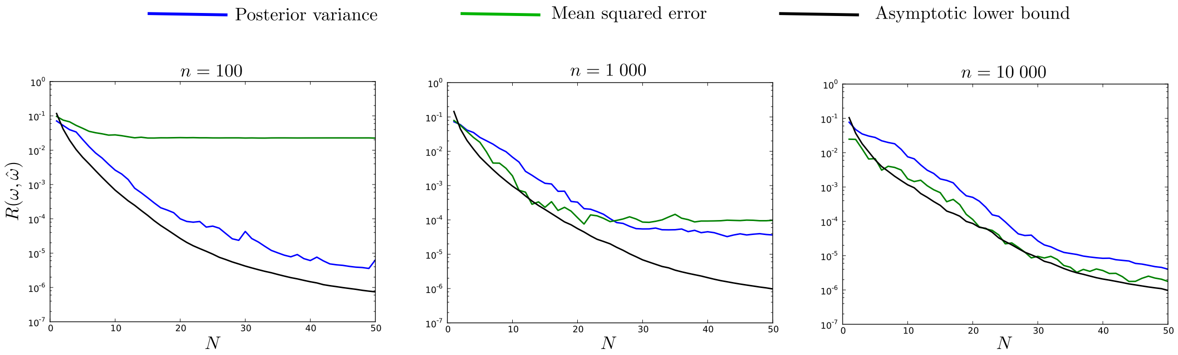

Figure 3.10: Left to right: the performance, as a function of the number of measurement N, of the sequential Monte Carlo algorithm for $n = 100, 1\,000$, and $10\,000$ particles. The model is that of equation (3.10) with $\eta = 1$ and $T_2 = 100\pi$ (see also figure 3.8). The blue curve is the posterior variance of the particles; green is the actual mean squared error and black is the asymptotic lower bound. The averages have been calculated from 100 trials.

we can make two explicit approximations to any utility function optimization. Firstly, we can use only the highest weighted particles to compute the expectation values appearing in the utility function[11]. Secondly, we can perform a *stochastic optimization* by Monte Carlo averaging the expectations over data sets as well [155]. Of course, the latter will be of little use for binary data produced by a measurement on a quibt since only one likelihood evaluation is necessary already. However, as the dimension of the underlying Hilbert space increases, so too will the number of outcomes of the measurements.

Note that sequential Monte Carlo algorithm presented above can also estimate credible regions and hyperparameters. Thus, we can also provide region and distribution estimates in addition to point estimates of parameters. The performance of the algorithm for these problems will be investigated in future research.

[11] This idea is due to Chris Granade and seems to perform well on our example problem in preliminary tests.

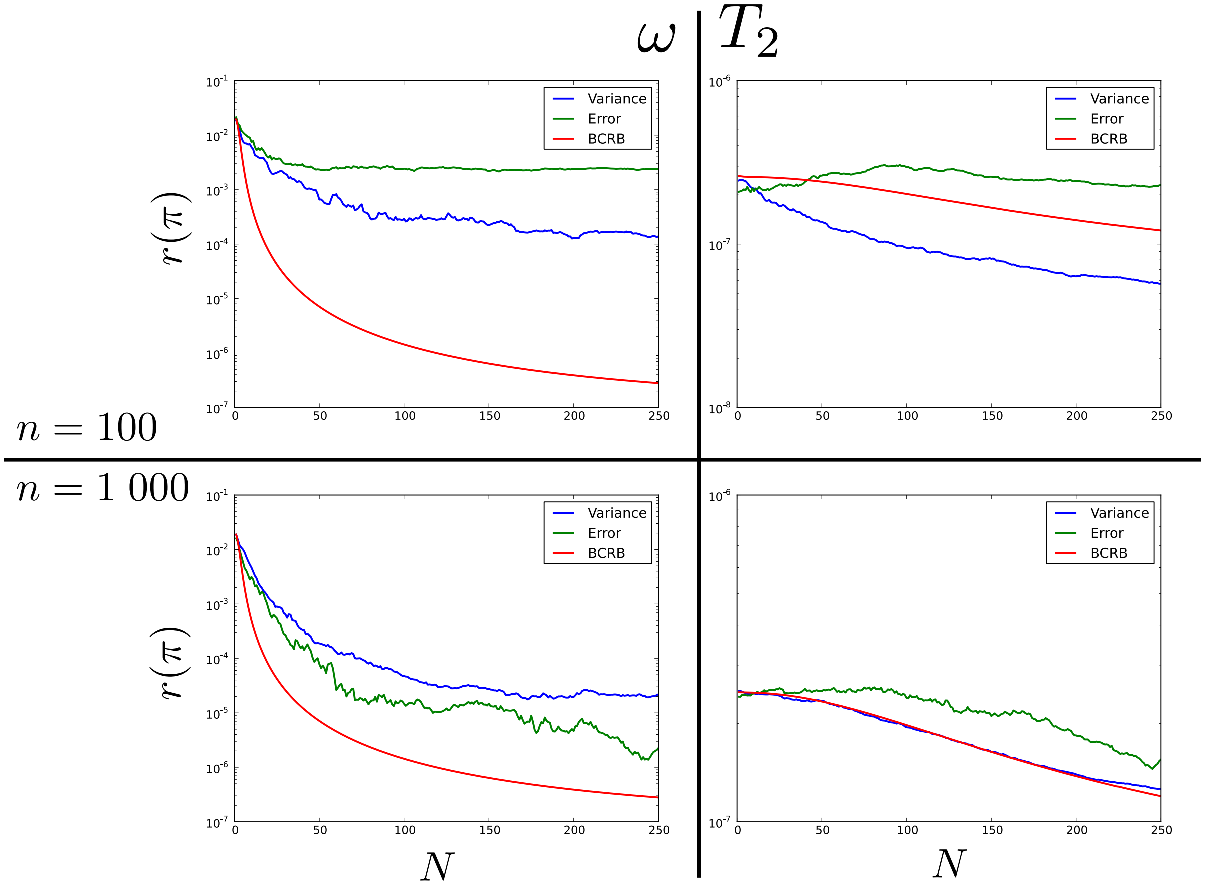

Figure 3.11: The performance, as a function of the number of measurement N, of the sequential Monte Carlo algorithm for $n = 100$ and 1 000, particles. The model is that of equation (3.10) with $\eta = 1$ and unknown T_2 (which is estimated as $\Gamma = 1/T_2$). The blue curve is the posterior variance of the particles; green is the actual mean squared error and red is numerically calculated Bayesian Cramer-Rao lower bound. The averages have been calculated from 100 trials.

Chapter 4

Necessity of Negativity in Quantum Theory

The quasi-probability representations of quantum theory evolved from the *phase space* representations. Phase space is a natural concept in classical theory since it is equivalent to the state space. The idea of formulating quantum theory in phase space dates back to the early days of quantum theory when the Wigner function was introduced [224]. The term quasi-probability refers to the fact that the function is not a true probability density as it takes on negative values for some quantum states.

The Wigner phase space formulation of quantum theory is equivalent to the usual abstract formalism of quantum theory in the same sense that Heisenberg's matrix mechanics and Schrodinger's wave mechanics are equivalent to the abstract formalism. The Wigner function representation is not the only quasi-probabilistic formulation of quantum theory. However, in most, if not all, of the familiar quasi-probability representations the kinematic or ontic space of the representation is presumed to be the usual canonical phase space of classical physics. In the broader context of attempting to represent quantum mechanics as a classical probability theory, the classical state space need not necessarily correspond to the phase space of some classical canonical variables. This is important since most quantum information protocols deal with finite systems rather than the continuous variables of classical physics.

There is a long standing problem of understanding the conditions under which a given physical process is "truly quantum" and this takes sharper focus in the context of quantum computing where a major open question is to determine the necessary and sufficient conditions for a quantum processor to exponentially outperform a classical one. While entanglement is the most notable criteria for quantumness, it can only be defined for composite systems. A non-negative quasi-probability function is a true probability distribution, prompting some authors to suggest that the presence of negativity in this function is a defining signature of non-classicality *for a single system*. Thus, a central concept in studies of the quantum-classical contrast in the quasi-probability formalisms of quantum theory is the appearance of *negativity*. However, it should be understood that any particular representation is non-unique and to some extent arbitrary since a state with negativity in one representation is positive in another.

In this chapter we complete the program set out in references [69, 70] by generalizing those results to quantum systems of arbitrary dimension and classical representations respecting the weakest – or, most general – set of restrictions. This chapter is structured as follows. First, in section 4.1, we review the ontological models framework. Then, we show how this leads to a no-go theorem and a natural generalization to quasi-probability in section 4.2. The formalism of references [69, 70] is briefly reviewed in section 4.3 and the fully general no-go theorem and generalized quasi-probability formalism is presented in section 4.4.

This chapter is based on references [73, 67].

4.1 Ontological Models and Non-contextuality

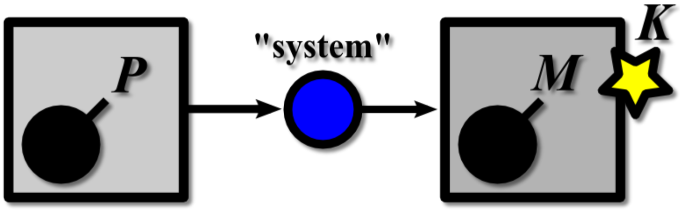

Figure 4.1: With appologies to my experimentalist friends: an abstraction of physical experiment. An *operational theory* is a mathematical model for the probability $\Pr(K|P, M)$ – the prediction that K will happen given settings P and M are chosen.

An *operational theory* [104, 199, 106] is an attempt to mathematically model a real physical experiment. The concepts in the theory are *preparations*, *systems*, *measurements* and *outcomes*. A preparation P is a proposition specifying how a real physical system has come to be the object of experimental investigation. We will reason about a set of mutually exclusive preparations $\mathcal{P}$. The system is assumed to then be measured according to some measurement procedure which produces an unambiguous answer called the outcome. The measurement procedure is specified by a proposition M belonging to a set of mutually exclusive possibilities $\mathcal{M}$. The outcome is specified by a proposition k which is assumed to belong to a set of mutually exclusive and exhaustive possibilities K. This means that, in any given run of the experiment, exactly one of K is true.

For example, P could be the setting of the knob on a "black box" which sends objects to another "black box" with knob setting M which outputs some sensory cue (e.g. an audible "click" or flash of light) labeled k as each system passes through. Note that P and M need not be statements about devices in a laboratory [184]; P could be a statement about a photon arriving from the sun, for example.

For a fixed preparation and measurement the outcome may not be deterministic. The role of the theory is then to describe the probability of each each outcome conditional on the the various combinations of preparations and measurements. That is, the theory should tabulate the numbers $\Pr(k|P \wedge M)$ for all k, P and M. For fixed P and M, the mutually exclusive and exhaustive property of K implies $\sum_k \Pr(k|P \wedge M) = 1$. An operational theory is then a specification $(\mathcal{P}, \mathcal{M}, K, \{\Pr(k|P \wedge M)\})$.

Quantum theory is an example of an operational theory where each preparation $P \in \mathcal{P}$ is associated with a density operator $\rho_P \in \mathbb{D}(\mathcal{H})$ via the mapping $\rho : \mathcal{P} \to \mathbb{D}(\mathcal{H})$. In general, this mapping is not required to be injective or surjective; different preparations may lead to the same density operator and there may not exist a preparation which leads to every density operator. Similarly, each measurement $M \in \mathcal{M}$ and outcome $k \in K$ is associated with an effect $E_{M,k} \in \mathbb{E}(\mathcal{H})$ via the mapping $E : \mathcal{M} \times K \to \mathbb{E}(\mathcal{H})$. Again this mapping need not be injective or surjective. Quantum theory prescribes the probabilities $\Pr(k|P \wedge M) = \mathrm{Tr}(\rho_P E_{M,k})$ which is called the Born rule. Since K is a set of mutually exclusive and exhaustive possibilities, for fixed M, $\{E_{M,k}\} \in \mathrm{POVM}(\mathcal{H})$.

Notice that an operational theory only specifies the probabilities given the preparations in the set $\mathcal{P}$. Suppose, for example, we are told a coin is tossed which determines which of two preparations procedures are implemented in a laboratory experiment. The operational theory does not offer the probability of a given outcome conditional on this information. However, the laws of probability provide us with the tools necessary to derive the desired probabilities. In general, the task is, given any disjunction $\bigvee_i P_i$ of propositions from the set of preparations $\mathcal{P}$, determine the probabilities $\Pr(k|\bigvee_i P_i \wedge M)$. Since the set of preparations is mutually exclusive, the laws of probability dictate

$$\Pr(k|\bigvee_i P_i \wedge M) = \sum_i \Pr(k|P_i \wedge M)\Pr(P_i). \tag{4.1}$$

Similarly, for any disjunction $\bigvee_j M_j$ of propositions from the set of measurements $\mathcal{M}$

$$\Pr(k|P \wedge \bigvee_j M_j) = \sum_j \Pr(k|P \wedge M_j)\Pr(M_j). \tag{4.2}$$

Putting Equations (4.1) and (4.2) together yields

$$\Pr(k|\bigvee_i P_i \wedge \bigvee_j M_j) = \sum_{ij} \Pr(k|P_i \wedge M_j)\Pr(P_i)\Pr(M_j). \tag{4.3}$$

Consider the example of quantum theory again. Let $\bigvee_i P_i$ be an arbitrary disjunction of preparations from $\mathcal{P}$. Then for fixed k and M

$$\begin{aligned}\Pr(k|\bigvee_i P_i \wedge M) &= \sum_i \Pr(k|P_i \wedge M)\Pr(P_i)\\ &= \sum_j \mathrm{Tr}(\rho_{P_i}E_{M,k})\Pr(P_i)\\ &= \mathrm{Tr}\left(\sum_i \Pr(P_i)\rho_{P_i}E_{M,k}\right)\end{aligned}$$

suggesting we define the function $\hat{\rho} : D(\mathcal{P}) \to \mathbb{D}(\mathcal{H})$

$$\hat{\rho} : \bigvee_i P_i \mapsto \textstyle\sum_i \Pr(P_i)\rho_{P_i}. \tag{4.4}$$

Similarly, let us define a function $\hat{E} : D(\mathcal{M}) \times K \to \mathbb{E}(\mathcal{H})$ as

$$\hat{E} : (\bigvee_j M_j, k) \mapsto \textstyle\sum_j \Pr(M_j)E_{M_j,k}. \tag{4.5}$$

Now, using $\hat{\rho}$ and $\hat{E}$ and defining $P_D := \bigvee_i P_i$ and $M_D = \bigvee_j M_j$ we have

$$\Pr(k|\bigvee_i P_i \wedge \bigvee_j M_j) = \mathrm{Tr}(\hat{\rho}_{P_D}\hat{E}_{M_D,k}). \tag{4.6}$$

This works for the probabilities of any operational theory and hence the set $(D(\mathcal{P}), D(\mathcal{M}), K, \mathcal{H}, \hat{\rho}, \hat{E})$ is an operational theory itself. Unless specified otherwise we will simply assume this analysis has been done when given an operational theory. This is equivalent to assuming the ρ and E of an operational theory $(\mathcal{P}, \mathcal{M}, K, \mathcal{H}, \rho, E)$ have ranges which are convex sets.

An *ontological model* is an attempt at interpreting an operational theory as an effectively epistemic theory of a deeper model describing the *real state of affairs* of the system. An ontological model posits a set Λ of mutually exclusive and exhaustive *ontic states* λ. Each preparation P is assumed to output the system in a particular ontic state λ. However, the experimental arrangement may not allow one to have knowledge of which state was prepared. This ignorance is quantified via a conditional probability $\Pr(\lambda|P)$. When this probability is viewed as a mapping $\mathcal{P} \to P(\Lambda)$ it is neither injective or surjective. That is, in general, different preparation may give the same probability distribution of Λ and, certainly, not all probability distributions are realized. A measurement M may not deterministically give an outcome k which reveals the state λ of the system. Each measurement can then only be associated with an conditional probability $\Pr(k|M \wedge \lambda) \in P(M)$. Again, when this probability is viewed as a mapping $\mathcal{M} \times K \to P(M)$ it is not assumed to be injective or surjective. An ontological model also requires that Λ be such that knowledge of λ renders knowledge of P irrelevant. In the language of probability,

CI k is conditionally independent of P given λ.

Summarizing, an ontological model is a set $(\mathcal{P}, \mathcal{M}, K, \Lambda, \{\Pr(\lambda|P)\}, \{\Pr(k|M \wedge \lambda)\})$ such that CI holds. As a consequence of CI, the law of total probability states

$$\Pr(k|P \wedge M) = \sum_{\lambda} \Pr(k|M \wedge \lambda) \Pr(\lambda|P). \tag{4.7}$$

Loosely speaking, *non-contextuality* encodes the property that operationally equivalent procedures are represented equivalently in the ontological model [199]. Two preparations are indistinguishable operationally if no measurement exists for which the probability of any outcome is different between the two. An ontological model is non-contextual (with respect to it preparations) if the probabilities over the ontic space for operationally equivalent preparations produce are equal. Similarly, measurements can be operationally equivalent and the ontological model can be non-contextual with respect to them. A mathematically concise definition of non-contextuality is as follows.

Definition 4.1.1. *Let* $(\mathcal{P}, \mathcal{M}, K, \Lambda, \{\Pr(\lambda|P)\}, \{\Pr(k|M \wedge \lambda)\})$ *define an ontological model.*

(a) Let $P, P' \in \mathcal{P}$*. The ontological model is* preparation non-contextual *if for all* $k \in K$*,* $M \in \mathcal{M}$ *and* $\lambda \in \Lambda$

$$\Pr(k|P \wedge M) = \Pr(k|P' \wedge M) \Rightarrow \Pr(\lambda|P) = \Pr(\lambda|P').$$

(b) Let $M, M' \in \mathcal{M}$. The ontological model is measurement non-contextual *if for all $k \in K$, $P \in \mathcal{P}$ and $\lambda \in \Lambda$*

$$\Pr(k|P \wedge M) = \Pr(k|P \wedge M') \Rightarrow \Pr(k|M \wedge \lambda) = \Pr(k|M' \wedge \lambda).$$

(c) The model is simply called a non-contextual ontological model *if it is both preparation and measurement non-contextual.*

4.2 Contextuality and Quasi-Probability

Recall that quantum theory as an operational model is defined via the mappings ρ and E. As an example, consider only those preparations associated with pure states. These pure states are determined by a mapping $\rho =: \psi : \mathcal{P} \to \mathbb{P}(\mathcal{H})$. The model of Beltrametti and Bugajski [17] posits the ontic space $\Lambda = \mathbb{P}(\mathcal{H})$ and a sharp probability distribution

$$\Pr(\lambda|P)d\lambda = \delta(\lambda - \psi_P)d\lambda \tag{4.8}$$

for each preparation P. This model suggests that the quantum state provides a complete specification of reality. Recall each measurement procedure is associated with a POVM via the mapping $E : \mathcal{M} \times K \to \mathbb{E}(\mathcal{H})$. Each measurement procedure M implies a conditional probability

$$\Pr(k|M \wedge \lambda) = \langle \lambda, E_{M,k}\lambda \rangle. \tag{4.9}$$

To show this is an ontological model, it remains only to verify that Equation (4.7) is satisfied. It follows that

$$\begin{aligned}\Pr(k|P \wedge M) &= \int_\Lambda \Pr(k|M \wedge \lambda) \Pr(\lambda|P)d\lambda \\ &= \int_\Lambda \langle \lambda, E_{M,k}\lambda \rangle \delta(\lambda - \psi_P)d\lambda \\ &= \langle \psi_P, E_{M,k}\psi_P \rangle,\end{aligned}$$

which is the Born rule for pure states. The Beltrametti-Bugajski model is an ontological model for pure state quantum theory which is preparation non-contextual. However, the range of the mapping ρ in this case is not convex. As we will now see, if we looked at all possible logical disjunctions of preparations in this model, so that range of the new mapping $\hat{\rho}$ is $\mathbb{D}(\mathcal{H})$, quantum theory admits no non-contextual model.

Lemma 4.2.1. *Suppose the convex range of E contains a basis (for $\mathbb{H}(\mathcal{H})$) and the ontological model $(\mathcal{P}, \mathcal{M}, K, \Lambda, \{\Pr(\lambda|P)\}, \{\Pr(k|M \wedge \lambda)\})$ is preparation non-contextual. Then, there exists a affine mapping $\mu : \mathrm{Ran}(\rho) \to P(\Lambda)$ satisfying $\mu(\lambda|\rho_P) = \Pr(\lambda|P)$.*

Proof. For equivalent disjunctions of preparations

$$\mathrm{Tr}(\rho_P E_{M,k}) = \Pr(k|P \wedge M) = \Pr(k|P' \wedge M) = \mathrm{Tr}(\rho_{P'} E_{M,k}). \tag{4.10}$$

Since the range of E is a basis it spans $\mathbb{H}(\mathcal{H})$ and therefore $\rho_P = \rho_{P'}$ if and only if P and P' are operationally equivalent. Then, from the definition of preparation non-contextuality, $\Pr(\lambda|P) \neq \Pr(\lambda|P')$ implies $\rho_P \neq \rho_{P'}$. Thus the mapping $\mu(\lambda|\rho_P) = \Pr(\lambda|P)$ is well-defined.

Now we show the mapping μ is affine. That is,

$$\mu(\lambda|p\rho_P + (1-p)\rho_{P'}) = p\mu(\lambda|\rho_P) + (1-p)\mu(\lambda|\rho_{P'})$$

for all $p \in [0, 1]$ and all $P, P' \in D(\mathcal{P})$. There is some $P'' \in D(\mathcal{P})$ such that $\rho_{P''} = p\rho_P + (1-p)\rho_{P'}$. That is to say, $P'' = P \vee P'$ while $\Pr(P) = p$ and $\Pr(P') = 1-p$. From the non-contextuality assumption this implies

$$\begin{aligned}\Pr(\lambda|P'') &= \Pr(\lambda|P \vee P') \\ &= \Pr(P)\Pr(\lambda|P) + \Pr(P')\Pr(\lambda|P').\end{aligned}$$

Applying the definition of μ to this yields

$$\mu(\lambda|\rho_{P''}) = p\mu(\lambda|\rho_P) + (1-p)\mu(\lambda|\rho_{P'}),$$

which proves μ is affine. □

Lemma 4.2.2. *Suppose the range of ρ contains a basis (for $\mathbb{H}(\mathcal{H})$) and the ontological model $(\mathcal{P}, \mathcal{M}, K, \Lambda, \{\Pr(\lambda|P)\}, \{\Pr(k|M \wedge \lambda)\})$ is measurement non-contextual. Then, there exists a unique convex-linear mapping $\xi : \mathbb{E}(\mathcal{H}) \to P(K)$ satisfying $\Pr(k|M \wedge \lambda) = \xi(E_{M,k})$.*

The proof of Lemma 4.2.2 is similar to that of Lemma 4.2.1. Together, the mappings μ and ξ are called a *classical representation of quantum theory.*

Lemma 4.2.3. *A classical representation satisfies, for all $\lambda \in \Lambda$, all $\rho \in \mathbb{D}(\mathcal{H})$ and all $E \in \mathbb{E}(\mathcal{H})$,*

(a) $\mu_\rho(\lambda) \in [0, 1]$ *and* $\sum_\lambda \mu_\rho(\lambda) = 1$*,*

(b) $\xi_E(\lambda) \in [0,1]$ *and* $\xi_{\mathbb{1}}(\lambda) = 1$,

(c) $\mathrm{Tr}(\rho E) = \sum_\lambda \mu_\rho(\lambda)\xi_E(\lambda)$.

In Reference [69] the name "classical representation" was defined to be a set of mappings satisfying (a)-(c). Through Lemmas 4.2.1 and 4.2.2 we have shown that the assumption of non-contexuality guarantees that these mappings are well defined, convex linear and satisfy (a)-(c). However, regardless of how we choose to arrive at a pair of convex linear mappings satisfying (a)-(c), one (or more) of the assumptions we make will be false as shown by the following theorem.

Theorem 4.2.4. *A classical representation of quantum theory does not exist.*

The theorem is implied by an earlier result on a related topic in reference [35]. It is also implied by the equivalence of a classical representation and a non-contextual ontological model of quantum theory [200] and the impossibility of the latter [199]. This theorem was proven in reference [69] using the notion of a frame and its dual, which we will see later. A direct and more intuitive proof was given in reference [70].

Now let us relax some assumptions in order to avoid a contradiction. First, we will relax the assumption of non-contextuality.

Definition 4.2.5. *A* contextual representation of quantum theory *is a pair of mappings* $\mu : \mathbb{D}(\mathcal{H}) \times C_{\mathcal{P}} \to P(\Lambda)$ *and* $\xi : \mathbb{E}(\mathcal{H}) \times C_{\mathcal{M}} \to P(\Lambda)$ *which satisfy, for all* $\lambda \in \Lambda$, *all* $\rho \in \mathbb{D}(\mathcal{H})$ *and all* $E \in \mathbb{E}(\mathcal{H})$, *all* $c_{\mathcal{P}} \in C_{\mathcal{P}}$, *and all* $c_{\mathcal{M}} \in C_{\mathcal{M}}$,

(a) $\mu_{\rho,c_{\mathcal{P}}}(\lambda) \in [0,1]$ *and* $\sum_\lambda \mu_{\rho,c_{\mathcal{P}}}(\lambda) = 1$,

(b) $\xi_{E,c_{\mathcal{M}}}(\lambda) \in [0,1]$ *and* $\xi_{\mathbb{1},c_{\mathcal{M}}}(\lambda) = 1$,

(c) $\mathrm{Tr}(\rho E) = \sum_\lambda \mu_{\rho,c_{\mathcal{P}}}(\lambda)\xi_{E,c_{\mathcal{M}}}(\lambda)$.

Here $C_{\mathcal{P}}$ *and* $C_{\mathcal{M}}$ *are the preparation and measurement* contexts.

Although it is clearly possible, the author is unaware if a contextual representation has been explicitly constructed. More common, however, is to relax the assumption of non-negative probability.

Definition 4.2.6. *A* quasi-probability representation of quantum theory *is a pair of affine mappings* $\mu : \mathbb{D}(\mathcal{H}) \to L(\Lambda)$ *and* $\xi : \mathbb{E}(\mathcal{H}) \to L(\Lambda)$ *which satisfy, for all* $\lambda \in \Lambda$, *all* $\rho \in \mathbb{D}(\mathcal{H})$ *and all* $E \in \mathbb{E}(\mathcal{H})$, *(a)* $\mu_\rho(\lambda) \in \mathbb{R}$ *and* $\sum_\lambda \mu_\rho(\lambda) = 1$,

(b) $\xi_E(\lambda) \in \mathbb{R}$ *and* $\xi_{\mathbb{1}}(\lambda) = 1$,

(c) $\mathrm{Tr}(\rho E) = \sum_\lambda \mu_\rho(\lambda)\xi_E(\lambda)$.

It is immediately clear that theorem 4.2.4 is equivalent to the following "negativity theorem":

Theorem 4.2.7. *A quasi-probability representation of quantum theory must have negativity in either its representation of states or measurements (or both).*

There are many instances of mappings μ satisfying the first requirement. In reference [69] it was shown by construction how to find a mapping ξ which, together with μ, satisfy all of them. Most of the mappings μ are called *phase space* functions as they conform to added mathematical structure not required in definition 4.2.6. A phase space representation is then a particular type of quasi-probability representation which we formally define as follows:

Definition 4.2.8. *If there exists a symmetry group on* Λ, G, *carrying a unitary representation* $U : G \to \mathbb{U}(\mathcal{H})$ *and a quasi-probability representation satisfying the covariance property* $U_g \rho U_g^\dagger \mapsto \{\mu_\rho(g(\lambda))\}_{\lambda\in\Lambda}$ *for all* $\rho \in \mathbb{D}(\mathcal{H})$ *and* $g \in G$, *then* $\rho \mapsto \mu_\rho(\lambda)$ *is a* phase space representation of quantum states*.*

4.3 Unification via Frames

A *frame* can be thought of as a generalization of an orthonormal basis [48]. However, the particular Hilbert space under consideration here is not $\mathcal{H}$. Considered here is a generalization of a basis for $\mathbb{H}(\mathcal{H})$, which is the set of Hermitian operators on an complex Hilbert space of dimension d. With the trace inner product $\langle A, B\rangle := \mathrm{Tr}(AB)$, $\mathbb{H}(\mathcal{H})$ forms a *real* Hilbert space itself of dimension d^2. Let Λ be some set of cardinality $d^2 \leq |\Lambda| < \infty$.

A frame[1] for $\mathbb{H}(\mathcal{H})$ is a set of operators $\mathcal{F} := \{F(\lambda)\} \subset \mathbb{H}(\mathcal{H})$ which satisfies

$$a\|A\|^2 \leq \sum_{\lambda\in\Lambda} \mathrm{Tr}[F(\lambda)A]^2 \leq b\|A\|^2, \tag{4.11}$$

[1] Frames have been considered in the context of quantum theory for other purposes in [190, 22].

for all $A \in \mathbb{H}(\mathcal{H})$ and some constants $a, b > 0$. This definition generalizes a defining condition for an orthogonal basis $\{B_k\}_{k=1}^{d^2}$

$$\sum_{k=1}^{d^2} \mathrm{Tr}[B_k A]^2 = \|A\|^2, \tag{4.12}$$

for all $A \in \mathbb{H}(\mathcal{H})$. The mapping $A \mapsto \mathrm{Tr}[F(\lambda)A]$ is called a *frame representation* of $\mathbb{H}(\mathcal{H})$.

A frame $\mathcal{D} := \{D(\lambda)\}$ which satisfies

$$A = \sum_{\lambda \in \Lambda} \mathrm{Tr}[F(\lambda)A]D(\lambda), \tag{4.13}$$

for all $A \in \mathbb{H}(\mathcal{H})$, is a *dual frame* (to $\mathcal{F}$). The *frame operator* associated with the frame $\mathcal{F}$ is defined as

$$S(A) := \sum_{\lambda \in \Lambda} \mathrm{Tr}[F(\lambda)A]F(\lambda).$$

If the frame operator is proportional to the identity *superoperator*, $S = a\tilde{\mathbb{1}}$, the frame is called *tight*. The frame operator is invertible and thus every operator has a representation

$$A = S^{-1}SA = \sum_{\lambda \in \Lambda} \mathrm{Tr}[F(\lambda)A]S^{-1}F(\lambda). \tag{4.14}$$

The frame $S^{-1}\mathcal{F}$ is called the *canonical dual frame*. When $|\Lambda| = d^2$, the canonical dual frame is the unique dual, otherwise there are infinitely many choices for a dual frame. A tight frame is ideal from the perspective that its canonical dual is proportional to the frame itself. Hence, the reconstruction is given by the convenient formula

$$A = S^{-1}SA = \frac{1}{a}\sum_{\lambda \in \Lambda} \mathrm{Tr}[F(\lambda)A]F(\lambda)$$

which is to be compared with

$$A = \sum_{k=1}^{d^2} \mathrm{Tr}[B_k A]B_k$$

which defines $\{B_k\}_{k=1}^{d^2}$ as an orthonormal basis.

Recalling the formal definition (4.2.6) of a quasi-probability representation, we have the following theorem

Theorem 4.3.1. *Two functions μ and ξ constitute a quasi-probability representation if and only if*

$$\mu_\rho(\lambda) = \mathrm{Tr}[\rho F(\lambda)]$$
$$\xi_E(\lambda) = \mathrm{Tr}[E D(\lambda)],$$

where $\{F(\lambda)\}$ is a frame and $\{D(\lambda)\}$ is one if its duals.

This was proven in reference [70][2]. This theorem allows us to make the following statement which is equivalent to the no-classical-representation theorem (4.2.4) and negativity theorem (4.2.7): there does not exists two frames of positive operators which are dual to each other.

With this results we can create quasi-probability representations of the whole operational formalism of quantum theory, not just states. First, we chose one of the discrete quasi-probability functions. Second, we identify the frame which gives rise to it. Lastly, we compute its dual frame to obtain the part of the quasi-probability representation mapping, ξ, which takes measurements to functions on the space Λ.

Suppose instead the functions μ and ξ are defined via

$$\mu_\rho(\lambda) = \mathrm{Tr}[\rho F(\lambda)]$$
$$\xi_E(\lambda) = \mathrm{Tr}[E F(\lambda)],$$

where $\{F(\lambda)\}$ is a frame. Then,

(a) $\mu_\rho(\lambda) \in [0,1]$ and $\sum_\lambda \mu_\rho(\lambda) = 1$,

(b) $\xi_E(\lambda) \in [0,1]$ and $\xi_{\mathbb{1}}(\lambda) = 1$,

(c) $\mathrm{Tr}(\rho E) = \sum_{\lambda,\lambda'} \mu_\rho(\lambda)\xi_E(\lambda')\mathrm{Tr}[D(\lambda)D(\lambda')]$.

In reference [69] this representation was called a *deformed* probability representation since states and measurements are represented as true probabilities but the law of total probability is deformed.

[2] Compare this to a similar result in a more operational setting in references [171, 172].

The frame formalism also provides a convenient transformation matrix to map between representations. We have

$$\begin{aligned}\mu_\rho(\lambda) &= \mathrm{Tr}[\rho F(\lambda)]\\ &= \sum_{\lambda'} \mathrm{Tr}[\rho F'(\lambda')]\mathrm{Tr}[D'(\lambda')F(\lambda)]\\ &= \sum_{\lambda'} T_{\lambda'\lambda}\mu'(\lambda'),\end{aligned}$$

where the matrix $T_{\lambda'\lambda}$ is the symmetric matrix which takes the μ representation to μ' representation.

Given a quasi-probability representation note that the frame satisfies

$$\sum_{\lambda\in\Lambda} F(\lambda) = \mathbb{1}.$$

Thus, if the quasi-probability representation satisfies $0 \leq \mu(\lambda) \leq 1$, the frame is an informationally complete positive operator valued measure (IC-POVM)[3]. Similarly, the dual frame satisfies

$$\mathrm{Tr}[D(\lambda)] = 1,$$

for all $\lambda \in \Lambda$. Thus, if the the quasi-probability representation satisfies $0 \leq \xi(\lambda) \leq 1$, the dual frame is a set of density operators. The definitions and results we have so far considered are tailored to the case $d < \infty$ – that is, finite dimensional quantum theory. Now we will extend them to infinite dimensions as done in reference [73].

4.4 Generalization to Arbitrary Quantum Systems

For the remainder suppose that the dimension of the Hilbert space $\mathcal{H}$ is arbitrary and let (Ω, Σ) be a measurable space. In this section we define the generalization of frame to the space of trace-class operators $\mathcal{T}_s(\mathcal{H})$. First, we need to introduce the notation which generalizes the familiar concepts from the finite dimensional analysis above. On the classical side, (Ω, Σ) denotes a measurable space, where Σ is a σ-algebra. Over this space, $\mathcal{M}_{\mathbb{R}}(\Omega, \Sigma)$ denotes the bounded signed measures while $\mathcal{F}_{\mathbb{R}}(\Omega, \Sigma)$ denotes the bounded measurable functions. A signed measure generalizes the usual notion of measure to allow for negative

[3] Frames have also been used in definition of informationally completeness in the context of tomography in reference [22].

values. The classical states are the probability measures, denoted $\mathcal{S}(\Omega,\Sigma) \subset \mathcal{M}_{\mathbb{R}}(\Omega,\Sigma)$. The classical effects are the measurable functions taking values in $[0,1]$. These are denoted $\mathcal{E}(\Omega,\Sigma) \subset \mathcal{F}_{\mathbb{R}}(\Omega,\Sigma)$. On the quantum side, the familiar symbol $\mathcal{H}$ denotes a Hilbert space. Over this space, $\mathcal{T}_s(\mathcal{H})$ denotes the trace-class self-adjoint operators while $\mathcal{B}_s(\mathcal{H})$ denotes the bounded self-adjoint operators. Quantum states are the density operators $\mathcal{S}(\mathcal{H}) \subset \mathcal{T}_s(\mathcal{H})$. Quantum effects are the positive operators $\mathcal{E}(\mathcal{H}) \subset \mathcal{B}_s(\mathcal{H})$ bounded above by $\mathbb{1}$. These are the elements of a positive operator valued measure (POVM).

Recall that a frame for Hermitian matrices is equivalent to an informationally complete *operator* valued measure (no positivity required). Essentially, it is a IC-POVM whose elements can have negative eigenvalues. So, we generalize the definition of informationally complete observable for infinite dimensions and define an *operator valued measure* as a map $F : \Sigma \to \mathcal{B}_s(\mathcal{H})$ satisfying $F(\emptyset) = 0$, $F(\Omega) = 1$ and

$$F\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} B_i,$$

where the sets $B_i \in \Xi$ are mutually disjoint and the sum converges in the weak sense. A *frame* for $\mathcal{T}_s(\mathcal{H})$ is an operator valued measure F for which the map $T : \mathcal{T}_s(\mathcal{H}) \to \mathcal{M}_{\mathbb{R}}(\Omega,\Sigma)$,

$$T(W) := \mathrm{Tr}(WF),$$

is injective. The map T is called a *frame representation* of $\mathcal{T}_s(\mathcal{H})$.

Similarly, generalizing the reconstruction formula for finite frames yields the generalized notion of a dual. That is, given a frame F, a *dual frame* to F is a map $D : \Omega \to \mathcal{B}_s(\mathcal{H})^*$ for which the function

$$SA(\cdot) := D_{(\cdot)}(A)$$

is measurable and satisfies

$$A = \int_{\Omega} SAdF, \tag{4.15}$$

for all $A \in \mathcal{B}_s(\mathcal{H})$.

Now we will generalize the definition of classical representation to the more general measurable space (Ω,Σ). A *classical representation of quantum mechanics* is a pair of mappings $T : \mathcal{S}(\mathcal{H}) \to \mathcal{S}(\Omega,\Sigma)$ and $S : \mathcal{E}(\mathcal{H}) \to \mathcal{E}(\Omega,\Sigma)$ such that

1. T and S are affine.
2. $S(0) = 0$.

3. For all $\rho \in \mathcal{S}(\mathcal{H})$ and $E \in \mathcal{E}(\mathcal{H})$,
$$\operatorname{Tr}(\rho E) = \int_{\Omega} (SE) d(T\rho). \tag{4.16}$$

As expected, we have

Theorem 4.4.1. *A classical representation of quantum mechanics does not exist.*

Proof. Suppose T and S form a classical representation of quantum mechanics. Conditions 1 and 2 imply that T and S can be extended to bounded linear functions $T : \mathcal{T}_s(\mathcal{H}) \to \mathcal{M}_{\mathbb{R}}(\Omega, \Sigma)$ and $S : \mathcal{B}_s(\mathcal{H}) \to \mathcal{F}_{\mathbb{R}}(\Omega, \Sigma)$. Thus for all $W \in \mathcal{T}_s(\mathcal{H})$ and $A \in \mathcal{B}_s(\mathcal{H})$, condition 3 gives
$$\operatorname{Tr}(WA) = \int_{\Omega} (SA) d(TW) = \operatorname{Tr}(W(T'SA)). \tag{4.17}$$
Hence for all $A \in \mathcal{B}_s(\mathcal{H})$, $A = T'SA$. S must be injective, and so S^{-1} exists, and if $R(S)$ is the range of S, then $T'|_{R(S)} = S^{-1}$. Therefore

$$T'\mathcal{E}(\Omega, \Sigma) \supset T'|_{R(S)}\mathcal{E}(\Omega, \Sigma) = S^{-1}\mathcal{E}(\Omega, \Sigma) \supset \mathcal{E}(\mathcal{H}).$$

Also, if $f \in \mathcal{E}(\Omega, \Sigma)$ and $W \in \mathcal{S}(\mathcal{H})$, then

$$0 \leq \int_{\Omega} f d(TW) = \operatorname{Tr}(W(T'f)) \leq \int_{\Omega} d(TW) = 1,$$
so $T'f \in \mathcal{E}(\mathcal{H})$ and $T'\mathcal{E}(\Omega, \Sigma) \subseteq \mathcal{E}(\mathcal{H})$. Therefore, $T'\mathcal{E}(\Omega, \Sigma) = \mathcal{E}(\mathcal{H})$.

However, this is a contradiction since $T'\mathcal{E}(\Omega, \Sigma) \subset \mathcal{E}(\mathcal{H})$ is a proper inclusion [205]. A short and intuitive proof that the inclusion is proper was briefly mentioned by Bugajski in Reference [33]. The proof relies on the notion of *coexistent effects* [110]. Intuitively, two (classical or quantum) effects are coexistent if they can be measured together. Any two classical effects are coexistent while two quantum effects are not necessarily coexistent. Since T' is a linear operator, it preserves the coexistence of effects. That is, the set $T'\mathcal{E}(\Omega, \Sigma)$ is one in which any two elements (quantum effects) are coexistent. But, again, not all quantum effects can be measured together (take two projectors in separate mutually unbiased bases, for example). Hence the inclusion is proper. □

In analogy with the finite dimensional case, we allow "negative probabilities" in a classical representation. A *quasi-probability representation of quantum mechanics* is a pair of mappings $T : \mathcal{S}(\mathcal{H}) \to \mathcal{M}_{\mathbb{R}}(\Omega, \Sigma)$ and $S : \mathcal{E}(\mathcal{H}) \to \mathcal{F}_{\mathbb{R}}(\Omega, \Sigma)$ such that

1. T and S are affine, T is bounded.

2. $T\rho(\Omega) = 1$.

3. $S(0) = 0$.

4. For all $\rho \in \mathcal{S}(\mathcal{H})$ and $E \in \mathcal{E}(\mathcal{H})$,

$$\mathrm{Tr}(\rho E) = \int_\Omega (SE)d(T\rho). \tag{4.18}$$

Since a quasi-probability representation in which $T\mathcal{S}(\mathcal{H}) \subset \mathcal{S}(\Omega, \Sigma)$ and $S\mathcal{E}(\mathcal{H}) \subset \mathcal{E}(\Omega, \Sigma)$ is a classical representation, we have immediately from Theorem 4.4.1,

Corollary 4.4.2. *A quasi-probability representation of quantum mechanics must have, for some $\rho \in \mathcal{S}(\mathcal{H})$, $E \in \mathcal{E}(\mathcal{H})$ either $(T\rho)(B) < 0$ for some $B \in \Xi$ or $(SE)(\omega) \notin [0,1]$ for some $\omega \in \Omega$.*

Note that, strictly speaking, we could have $SE(\omega) > 1$. However, we still refer to this as negativity since $SE(\omega)$ is meant to be interpreted as a probability and $1 - SE(\omega) < 0$ should stand on the same footing. Thus, the above corollary gives us the necessity of negativity in quasi-probability representations. Being more general, this theorem implies the previous three equivalent "negativity theorems" and says essentially the same; a classical representation does not exist and within a quasi-probability representation negativity must appear in the representation of the states or measurements or both.

It is obvious that every frame representation defines a quasi-probability representation. The converse is also true as established by the following theorem:

Theorem 4.4.3. *The pair (T, S) is a quasi-probability representation of quantum mechanics if and only if T is a frame representations of $\mathcal{T}_s(\mathcal{H})$ and S is a dual frame.*

Proof. Assume (T, S) is a quasi-probability representation. Define $F(B) := T'\chi_B$. By the definition of the dual map

$$\mathrm{Tr}[\rho F(B)] = \mathrm{Tr}[\rho T'\chi_B] = \int_\Omega \chi_B d(T\rho) = (T\rho)(B).$$

It is clear then that $F(\emptyset) = 0$ and $F(\Omega) = 1$ by normalization. The σ-additive can be verified by directly substituting the union of an arbitrary sequence of disjoint sets $\{B_i\} \subset \Sigma$. It is known that T and S can be uniquely extended to bounded linear mappings

$T : \mathcal{T}_s(\mathcal{H}) \to \mathcal{M}_{\mathbb{R}}(\Omega, \Sigma)$ and $S : \mathcal{B}_s(\mathcal{H}) \to \mathcal{F}_{\mathbb{R}}(\Omega, \Sigma)$. Now suppose TW is the trivial measure for some $W \in \mathcal{T}_s(\mathcal{H})$. From the last property of a quasi-probability representation, the Law of Total Probability, we have

$$\mathrm{Tr}(WA) = \int_\Omega (SA) d(TW) = 0.$$

Since this is true for all $A \in \mathcal{B}_s(\mathcal{H})$, $W = 0$. Therefore T is injective, and we have shown that T is a frame representation. Now define $D_\omega = S^* \delta_\omega$. So we have

$$D_\omega(A) = S^* \delta_\omega(A) = \int_\Omega (SA) d\delta_\omega = (SA)(\omega).$$

And, again from the Law of Total Probability,

$$\mathrm{Tr}[WA] = \int_\Omega (SA) d\mathrm{Tr}[WF] = \mathrm{Tr}\left[W \int_\Omega (SA) dF\right].$$

Hence

$$A = \int_\Omega (SA) dF,$$

and by definition D is dual to F. The proof of the converse should be clear. □

4.5 Discussion and Future Directions

The idea of using quasi-probability representations to identify non-classical features of quantum theory can be done in the following ways

1. Choose your favourite representation (the P function, for example) and define negatively represented states in this representation as the *quantum* ones.

2. Choose your favourite property of quantum states as your notion of quantumness (entanglement, for example) and find a representation in which negatively represented states have this property.

3. Within a quantum resource theory (magic state quantum computation, for example), identify a representation in which the "free" resources correspond exactly to classically represented states, measurements and transformation.

In the Appedix, we argue that only the last option is viable. Indeed, as shown in reference [215] and discussed in section A.2.1, the discrete Wigner function satisfies the constraint that the “free” resources in magic state distillation of qudits are represented classically.

It is not clear that such a procedure is always possible. Take a resource theory where the restricted set of quantum operations has, say, a classically efficient description. Such a resource theory is magic state distillation of qu*bits*. The Gottesman-Knill theorem [92] shows that the so-called stabilizer formalism (the restricted, or “free” resources in magic state distillation) has a classically efficient description. It is intuitive to suggest that one can find a quasi-probability representation (like the discrete Wigner function) in which these operations correspond to classically represented objects. However, this is not possible since, for qubits, one can violate a Bell-type inequality and such a classical representation is ruled out by Bell’s theorem. Now, it could be the case that locality and efficient simulatability are fundamentally distinct notions of classicality. If this is the case, it should be possible to find a representation in which the stabilizer formalism is classical yet also contextual in efficiently simulatable way. We leave this as an interesting question for future investigation.

Finally, we comment on the use of negativity and the quasi-probability formalism for measuring quantumness in an operationally restricted setting. Many aspects of quantum theory can be thought of as a resource theory [60, 94]. The most famous example of this is the restriction of local operations and classical communication (LOCC) which led to the theory of entanglement, the resource theory of LOCC. Entanglement is rich resource theory giving us *measures* of entanglement and specifies when one entangled state can be *converted* to another via LOCC. We conjecture the following: a proper resource theory should implicitly define a quasi-probability representation in which the restricted operations are represented classically (positive probability distributions, positive conditional distributions and stochastic transformations). If this conjecture is true, then negative quasi-probability is at least a necessary condition for any real, or suspected, quantum advantage. Evidence for this is provided in appendix A, where we discuss ongoing collaborative research which shows that this correspondence holds for the *magic state* model of qudit quantum computation and a generalization thereof to continuous variable systems.

APPENDICES

Appendix A

Negativity as a Resource for Non-classicality

Given this ubiquitous presence of “negative probability” in quantum theory, there is a strong tradition in physics of considering negativity of a particular quasi-probability functions as an indicator of non-classical features of quantum states. However, distinctions between quantum states are irrelevant outsides of the context of some operational setup. In other words, it depends on what we want do with a quantum system that defines which quantum states possess any non-classical features. This can be formalized via specific resource theories defined by operational restrictions on quantum theory. One important class of operational restrictions are resource theories for quantum computation. That is, there exist interesting models of computation in which some easily accessible operations are *not* able to perform an arbitrary quantum computation – but additional resource operations enable universal quantum control.

In section A.1 we review the standard interpretations of negativity in particular qausi-probability representations as non-classicality indicators. We also give arguments for the failure of this interpretation when place in an operational context. We have conjectured that negativity can be given a formal interpretation of non-classicality when one can find a classical representation for quantum mechanical subtheory or operational restriction which gives rise to a resource theory. In section A.2, we identify such a correspondence between quantum computational models and quasi-probability representations such that negativity in the representation is a necessary condition for quantum computation. In particular, we have identified a correspondence between the “magic state model” [30] (which is well motivated by recent progress in quantum error correction and topological quantum computation) and a particular quasi-probability representation: the “discrete Wigner function”

[96]. In this case it is shown that negativity in this representation is a necessary condition for quantum computation. Analogously, we define a mixed state magic-state-like model of computation for continuous variables – those encountered in quantum optics, for example. Here we find the same phenomenon – negativity of Wigner's original function is necessary for continuous variable quantum computation via a continuous variable generalization of the magic state model.

The technical content of reference [215] was mostly due to the efforts of Victor Veitch and the details will not be reproduced here. The refined discussion of negativity as a resource has coalesced in ongoing discussions with Victor Veitch, Joseph Emerson and Daniel Gottesman.

A.1 Negativity as "Quantumness" *per se*

There have been a variety of approaches to the problem of characterizing what is non-classical about quantum theory. In the previous section one such notion of non-classicality was considered: the requirement of "negative probability", or simply *negativity*. However, the negativity theorem leaves open the question of the interpretation of negativity in any *particular* representation. In this section we discuss the ways in which negativity can be applied as a criterion for quantumness with respect to particular choices of representation. First we review the traditional ideas of quantumness, namely *contextuality* and *nonlocality*[1], in relation to negativity. Then we will discuss the deficiencies of broadly defining negativity as quantumness.

A.1.1 Traditional "Quantumness": Contextuality and Nonlocality

The traditional definition of contextuality evolved from a theorem which appears in a paper by Kochen and Specker [131]. The Kochen-Specker theorem concerns the standard quantum formalism: physical systems are assigned states in a complex Hilbert space $\mathcal{H}$ and measurements are made of observables represented by Hermitian operators. The theorem establishes a contradiction between a set of plausible assumptions which together imply that quantum systems possess a consistent set of pre-measurement values for observable quantities. Let $\mathcal{H}$ be the Hilbert space associated with a quantum system and $A \in \mathbb{H}(\mathcal{H})$

[1] These are diverse and rich fields of study in their own right. A starting point for the interested reader on contextuality is reference [109] and quantum nonlocality is reference [177].

be the operator associated with some observable. The function $f_\psi(A)$ represents the value of the observable when the system is in state ψ. One assumption used to derive the contradiction is that for any function F, $f_\psi(F(A)) = F(f_\psi(A))$. This is plausible because, for example, we would expect that the value of A^2 could be obtained in this way from the value of A.

The conclusion of Kochen-Specker theorem implies the following counterintuitive example [119]. Suppose three operators A, B, and C satisfy $[A, B] = 0 = [A, C]$, but $[B, C] \neq 0$. Then, the value of the observable A will depend on whether observable B or C is chosen to be measured as well. That is, assuming that physical systems do possess values which can be revealed via measurements, the value of A depends on the *context* of the measurement.

What the Kochen-Specker theorem establishes then is the mathematical framework of quantum theory does not allow for a *non-contextual* model for pre-measurement values. This fact is often expressed via the phrase "quantum theory is contextual".

The original notion of contextuality in lacking in the sense that it only applies to the standard formalism of quantum theory and does not apply to general operational models. This problem was addressed by Spekkens as discussed above. The notion of contextuality defined by Spekkens is more general; one can recover the original assumptions of Kochen-Specker by assuming that the projector valued measures in the spectral resolutions of observables are represented by dispersion free (0-1 valued) conditional probabilities (these are also called *sharp indicator functions*) [2]. Since the set of fewer assumptions already contains a contradiction when taken in conjunction, the addition of the assumption of Kochen-Specker is unnecessary. Thus, we need only consider the more general notion of contextuality we have already defined. This more general notion of contextuality has also recently been subject to experimental tests [201].

A hidden variable theory originally formulated by de Broglie and later by Bohm [26] is perhaps the most famous example of an ontological model of quantum theory. The model assumes that for a given experimental configuration, there exists a particle with well defined trajectory and a quantum state ψ. The hidden variable is the position of the particle in real space. That is, the classical state space is $\Lambda = \mathbb{R}^3 \times \mathcal{H}$. The Hilbert space is included in the state space as its serves as a wave which guides the particle. If at any time the particle is distributed according to quantum probability distribution $|\psi|^2$, it remains so. Thus, so long as it is assumed that the particle is prepared according to this distribution, the model provides the same predictions as the standard formulation of

[2] Cabello has also generalized the notion of contextuality to POVMs [36]. Again, however, the additional assumption of dispersion free condition probabilities is used. See [98] for an elaboration on this point. For a more broad discussion on contextuality see [159] and [106].

quantum theory.

Note that this model does not fit into the framework of quasi-probability representations. Exactly as it was for the Beltrametti-Bugajski model, the de Broglie-Bohm model does not consider the entire range of possible quantum states. Where a classical representation contains a convex-linear mapping $\rho \mapsto \mu(\lambda)$ for all $\rho \in \mathbb{D}(\mathcal{H})$, the de Broglie-Bohm model considers only a mapping with domain $\mathcal{H}$. Bell notes that [16] "in the de Broglie-Bohm theory a fundamental significance is given to the wavefunction, and it cannot be transferred to the density matrix."

Bell *does not* claim that the situation is such that the de Broglie-Bohm model *cannot* be extended to include density operators. The key words in his comment are "fundament significance". Indeed, the de Broglie-Bohm model *can* be extended to include density operators provided this extension is either contextual or contains negativity. In either case, the pure states (wavefunctions) retain their significance while the density operators possess non-classical features. As an example, the de Broglie-Bohm model could be such that $(\rho, c_{\mathcal{P}}) \mapsto \mu_{c_{\mathcal{P}}}(\lambda)$ where each preparation consists of a density operator ρ supplied with a context $c_{\mathcal{P}}$ which specifies a particular convex decomposition of ρ into pure states. Such a model would be preparation contextual.

The non-locality debate was initiated by a paper by Einstein, Podolsky and Rosen (EPR) [65] where it was argued that quantum mechanics is *incomplete* (each element of physical reality does not have a counterpart in quantum theory) if special relativity remains valid. The latter means physical causation must be local or events cannot have causes outside of their past light cones. Using a particular spatially separated quantum system, and some standard quantum theory, EPR concluded that quantum mechanics is either incomplete or nonlocal (or both!). Locality was such a desired property of any theory that quantum mechanics was concluded to be incomplete. That is, there must be elements of physical reality (hidden variables) which quantum mechanics does not account for.

The argument of EPR was reformulated by Bohm [26] for two qubits. The argument is built around the following hypothetical experiment. Two parties, Alice and Bob, are at distant locations with a source midway between them creating quantum systems described by the quantum state

$$\psi = \frac{1}{\sqrt{2}}(\phi_1 \otimes \phi_2 - \phi_2 \otimes \phi_1), \tag{A.1}$$

where $\{\phi_1, \phi_2\}$ is an orthonormal basis for a qubit. One particle is sent to Alice and the other to Bob. Alice performs the projective two-outcome measurement $\{P_1, P_2\}$ on the particle which was sent to her. The state in equation (A.1) is such that Alice, once she performs her measurement, she can predict with certainty the outcome Bob receives when he performs the same measurement at his side of the experiment *regardless of whether or not the measurement events are spacelike separated (i.e. nonlocal)*. For example, Alice could perform the measurement $\{\phi_1\phi_1^*, \phi_2\phi_2^*\}$. According to the collapse postulate, if Alice registers the first outcome, Bob particle will immediately collapse to ϕ_2 and he is certain to obtain the second outcome if he were to make the same measurement. Therefore, unless there exists hidden variables which pre-determine the possible outcomes when the particles are created, quantum theory is *nonlocal*.

Bell later investigated the possibility of finding the hidden variables Einstein thought to exist [16]. He noted immediately that the de Broglie-Bohm theory was such a theory yet in contained an astonishingly nonlocal character. He was able to prove that any hidden variable theory of quantum phenomena must possess nonlocal features. This is now called Bell's theorem.

The proof is by contraction and follows the general line of reasoning which lead to the negativity theorem: build a mathematical model with assumptions that can be identified with (or motivated by) some notion of classicality then prove that quantum theory does not satisfy these assumptions. Consider the EPR experimental setup. Alice and Bob can each perform a two-outcome measurement with outcomes labeled A and B, respectively. Without loss of generality, the outcomes can be assigned numerical values $A, B = \pm 1$.

Suppose there exist a classical state space Λ (i.e. a set of hidden variables or, as we have called it above, an ontology) which serves to determine the outcomes A and B. Probabilistic knowledge of the state is represented by a density $\mu(\lambda) \geq 0$ which is normalized

$$\int_\Lambda d\lambda \mu(\lambda) = 1.$$

The different measurements Alice and Bob can perform are parameterized by detector settings a and b, respectively. Locality is enforced by assuming that the outcomes A and B depend only the local detector settings and the global state. That is $A = A(a, \lambda)$ is allowed but $A = A(a, b, \lambda)$ is not. Define the correlation function

$$C(a, b) = \int_\Lambda d\lambda A(a, \lambda) B(b, \lambda) \mu(\lambda). \tag{A.2}$$

Bell's theorem states that the correlations obtained in the EPR experiment (i.e. a particular quantum experiment) cannot satisfy this equation. The proof follows by deriving an inequality from equation (A.2) such as

$$|C(a, b) - C(a, c)| \leq 1 + C(b, c). \tag{A.3}$$

This inequality holds for any hidden variable model which satisfies the locality assumption. For the quantum state in equation (A.1), the inequality is violated. This is the contradiction between the quantum theory and a local hidden variable model which proves Bell's theorem.

It was noted that the assumptions which go into the hidden variable models first considered by Bell imply those models are *deterministic*. That is, the theorem did not exclude models which suggested quantum theory only provides *stochastic* (or probabilistic) information of the possible outcomes of measurements. Bell later extended the theorem to include such models. For the EPR experimental setup, let the conditional probability of outcome $A = 1$, for Alice, given the state (hidden variable) is $\lambda \in \Lambda$ be denoted $M_A(\lambda)$ and similarly define $M_B(\lambda)$ for Bob. Now denote the conditional joint probability of the simultaneous outcomes $A, B = 1$ by $M_{AB}(\lambda)$. Fine [75] defines a *stochastic hidden variable model* as one which satisfies

$$\Pr(A = 1) = \int_\Lambda d\lambda M_A(\lambda)\mu(\lambda) \tag{A.4}$$

and

$$\Pr(A = 1, B = 1) = \int_\Lambda d\lambda M_{AB}(\lambda)\mu(\lambda). \tag{A.5}$$

If $M_{AB}(\lambda) = M_A(\lambda)M_B(\lambda)$, then the model is *factorizable*. Bell claimed this also encoded the assumption of locality. Again, it can be shown that quantum theory is in contradiction with an inequality derived from these assumptions. Fine showed that a factorizable stochastic hidden variable model exists for the EPR-type correlation experiment if and only if a deterministic hidden variable model exists for the experiment. Since the latter is ruled out, the former is also ruled out.

A.1.2 Negativity of Quantum States

There is a strong tradition in physics of considering negativity of the Wigner function as an indicator of non-classical features of quantum states. The non-classical features attributed to negativity of *the original* Wigner function include quantum nonlocality [16, 123], quantum chaos [168] and quantum coherence [168]. In quantum optics, however, tradition has been to use the Q- and P-functions (section B.1.2) to define quantumness [149]. The P-function of ρ is defined implicitly through

$$\rho = \int d^2\alpha P(\alpha)|\alpha\rangle\langle\alpha|.$$

If P has the properties of a probability distribution, then the state is a mixture of coherent states. Coherent states are minimum uncertainty states and this fact is often cited when it is stated that such a state is "the most classical" of the quantum states of light. More specifically, if P is a probability distribution then the quantum field cannot display genuine quantum optical effects and can be simulated by a stochastic classical electromagnetic field [220]. Technically, however, P is not a function but a distribution which can be highly singular. Thus P functions which are not classical distributions are difficult to experimentally prepare and verify; although, recent progress has been made [127].

Effort has been extended beyond qualitatively defining negativity as quantumness to *quantifying* quantumness via negativity. In terms of the Wigner function, the *volume* of the negative parts of the represented quantum state has been suggested as the appropriate measure of quantumness [126]. The *distance* (in some some preferred norm on $\mathbb{H}(\mathcal{H})$) to the convex subset of positive Wigner functions was suggested to quantify quantumness in reference [153]. This was also done in references [88, 89] for a finite analogs of the P- and Q-functions rather than the Wigner function.

The main difficulty with interpreting negativity in a particular quasi-probability representation as a criterion for or definition of quantumness is the non-uniqueness of that particular quasi-probability representation. We can always find a new representation in which any given state admits a non-negative quasi-probability representation. Recall, in fact, that in some representations all states are non-negative. Thus, negativity of some state ρ in one particular arbitrary representation is a meaningless notion of quantumness *per se*.

An alternative approach to establishing a connection between quantumness and negativity is to start by assuming some criterion for quantumness and then finding a choice of representation in which this criterion is expressed via negativity. This approach has been applied in the context of multi-partite systems for which entanglement is presumed to provide a criterion for quantumness. Entanglement is a kind of correlation between two quantum systems which cannot be achieved for classical variables and is one of the central ingredients in quantum information theory [3]. Recall that a density matrix ρ is *entangled* if it *cannot* be written as a convex combination of the form $\rho = \sum_k p_k \rho_k^{(1)} \otimes \rho_k^{(2)}$, for all k, where $\rho_k^{(1)}$ and $\rho_k^{(2)}$ are states on the individual subsystems. Consider a *product-state frame* constructed out of frames for two subsystems. That is, consider the frame $\{F^{(1)}(\lambda) \otimes F^{(2)}(\lambda')\}$, where each $\{F^{(j)}(\lambda)\}$ is a set of density matrices composing a frame. Then, if we represent a quantum state using the dual frame, we a have a quasi-probability representation in which states with negativity are entangled. Explicit constructions of such

[3] For a recent review of entanglement, see [115].

quasi-probability representations were developed by Schack and Caves [187]. An optimization procedure to find the representation with the *minimum* amount of negativity was given in reference [202].

An obvious limitation with the above approach is that entanglement cannot capture any notion of quantumness for single quantum systems. A second, more subtle, issue is that identification of entanglement as "the" crucial non-classical resources is problematic in certain branches of quantum information science. The most striking example questioning the role of entanglement in quantum information theory is DQC1 (deterministic quantum computing with one clean qubit). DQC1 [128] is a model of computation which refers to any algorithm which satisfies the following (or a modification not requiring exponentially more resources)[4]:

1. its input consists of a single pure state in the first *control* register and the remaining n registers are in the maximally mixed state $\rho = 2^{-n}\mathbb{1}$;
2. the input state is subjected to a unitary U_n controlled by the state of the first register;
3. the output is a statistical estimate of $2^{-n}\text{Tr}(U_n)$ (achieved by measuring the average of control bit in the Z basis).

DQC1 appears to be a non-trivial computational model which has been shown to have exponential advantages over (known) classical algorithms in the the follow areas: simulation of quantum systems [128], quadratically signed weight enumerators [129], evaluating the local density of states [66], estimating the average fidelity decay under quantum maps [174] and estimating the value of Jones polynomials [197].

In the DQC1 model, the bipartite split between the control qubit and the rest contains no entanglement and in reference [59] it was shown that there is a vanishingly small amount of entanglement across any other bipartite splitting. This suggests it is unlikely that entanglement is responsible for the speed-up provided by DQC1 [58]. Conceptually, computation is a local task with complex dynamics and may not require the non-local, Bell-inequality-violating correlations of entanglement [135]. A sentiment issued in reference [135] and reiterated recently by Vedral [214] is that no one single criteria can capture quantumness and perhaps even the resources necessary for the quantum advantage must be studied on a case-by-case basis.

[4] This model of computation has served as the basis for various definitions of complexity classes, also called DQC1. There are many open questions in this line of research and the interested reader should consult references [197, 9, 196].

An important consideration for all of the above approaches is that the notion of a quantum state, considered in isolation, is operationally meaningless. Comparison with experiment always requires specifying both a state (a preparation procedure) and a measurement. Consider two experiments, one which prepares a product state and measures the state by projecting onto an entangled basis, and a second which prepares an entangled state and measures that state in a product basis. Both experiments produce the same statistical predictions, but only the second is considered non-classical when considering the state in isolation. As emphasized in references [200, 69] we can overcome this obvious deficiency if we consider the whole operational set-up – states *and* measurements. In this way, the existence of a positive quasi-probability representation implies the existence of a non-contextual ontological model and vice versa.

The formalism of references [69, 70, 73] shows the necessity of negativity when considering a representation of the full quantum formalism. That is, the negativity theorem (theorem 4.2.7) applies to quasi-probability representations of quantum theory as a whole. However, the negativity theorem may not apply if we consider a specific experiment, device, or protocol which may not faithfully reproduce the full power of quantum theory. This work suggests the following promising approach: define a classical representation *of an experiment* as the existence of a frame and dual for which the convex hull of the experimentally accessible states and measurements have positive representation. Then, we can conclude that negativity, taken to mean the absence of any representation satisfying the above conditions, corresponds to quantumness.

The above criterion for classicality was considered in reference [186] to question the quantum nature of proposed NMR quantum computers. However, as noted there, the immediate objection is the following: the states and measurements can be represented by classical probabilities while the transformation between them may not be represented by classical stochastic maps. That is, a truly classical model must represent each applied transformation in a experiment as classical stochastic mapping. In reference [186], such stochastic maps were identified for the set of NMR experiments reported at the time [5].

The scope of quantum theory that has been consider thus far can be thought of as *kinematical*; only the description of experimental configurations is of concern. The traditional approach to quantum theory (quantum *mechanics*) focuses on how and why quantum systems change in time. Using the Wigner function formalism to describe the dynamical transformations predicted by quantum mechanics yields the dynamical law

$$\frac{\partial \mu_\rho}{\partial t} = \{H, \mu_\rho\} + \sum_{n=1}^{\infty} \frac{1}{2^{2n}(2n+1)!} \frac{\partial^{2n+1} H}{\partial q^{2n+1}} \frac{\partial^{2n+1} \mu_\rho}{\partial p^{2n+1}}, \tag{A.6}$$

[5] Note, however, that the reasonable requirement of an *efficient* classical model was not met.

where μ_ρ is the Wigner function and H is the classical Hamiltonian and $\{H, \mu_\rho\}$ is the classical Poisson bracket. Notice then that Equation (A.6) is of the form "classical evolution" + "quantum correction terms". Using this formalism, one can then do more than discuss which experimental procedures are classical. Now one can discuss the *transitions* between quantum and classical descriptions, a process known as *decoherence* [168, 100, 122]. The representation of the dynamics was also studied for the spherical phase space in [229, 124]. The goal is to compare the representation of the quantum dynamics (be it the Schrodinger equation or the more general *master equation*) to the natural classical dynamics of the representation's phase space. The challenge for finite dimensional systems is that no natural notion of discrete phase space exists for classical system. This problem has been recently studied by Livine [143] by introducing a *discrete differential calculus* for the discrete phase spaces of Wootters. However, beyond these few examples, transformations and dynamics have not been studied anywhere near to the extent that states have for quasi-probability representations and presents itself as a open problem.

A.2 Negativity as "Quantumness" via Quantum Computation

It is plausible that to each operationally restricted scenario, a unique resource can be identified which enables quantumness. For example, for the circuit model of quantum computation, is the demonstration by Vidal that large amounts of entanglement is necessary, but not sufficient, for quantum computational speedup [216]. Unfortunately, this result does not generalize to other models of quantum computation. A particularly important alternative model is the magic state model of Bravyi and Kitaev [30], which is well motivated by recent progress in quantum error correction, fault-tolerance and topological quantum computation. We will see here that, for odd dimensional systems, a necessary resource for universal computation in the magic state model corresponds to negativity in a distinguished discrete Wigner function [96]. In the second half of this section we shown that the analogous result holds for continuous variable quantum computation and the original Wigner function.

A.2.1 Negativity is a Necessary Condition for Magic State Distillation

The magic state distillation model requires only Clifford operations and the preparation of a mixed ancilla state [30]. The full set of ideal operations are taken to be: prepare a qudit in the state $|0\rangle$; apply unitary operators from the Clifford group; and measure an eigenvalue of a Pauli operator on any qudit. It is well known that these operations are insufficient for universal quantum computation [2], so something further is required. To this end an additional process is allowed: the preparation of an ancillary qudit in a mixed state ρ. Using only Clifford operations some, but not all, ancilla states may be distilled to non-stabilizer "magic states". These may be consumed to implement non-Clifford gates, enabling universal quantum computation. The major open problem is to determine the set of ancilla states which enable computation outside of the Clifford group. We will call a state ρ magic state distillable if for any $\epsilon > 0$ it is possible to produce a state with fidelity $1 - \epsilon$ to some pure, non-stabilizer state using only repeat preparations of ρ and ideal operations [6].

In [215] we derived a necessary condition for distillability which defines a region in the space of mixed quantum states. The only previously known boundary on this region is the trivial one consisting of the convex combinations of stabilizer states. Surprisingly, the region we derive here is strictly larger than this set, which implies the existence of bound states for magic state distillation [39]. These are states which cannot be prepared using Clifford resources but which do not enable universal quantum computation.

Our technique is to represent magic state distillation routines as stochastic processes over a discrete phase space. Intuitively, if the dynamics of the quantum system admit a representation as a classical statistical process then it should not be sufficient for universal quantum computation. To represent the magic state model as a classical probability theory we seek a quasi-probability representation where stabilizer states and projective measurements onto stabilizer states have non-negative representation and Clifford transformations correspond to stochastic processes. There exists a quasi-probability representation with these properties: the discrete Wigner function picked out by Gross [96] from the broad class defined by Gibbons *et al* [86] (this formalism is review in section B.2.5).

What was shown in [215] is that a necessary condition for a state ρ to enable universal computation in the magic state model is that it has negative quasi-probability representation. We also showed that in power of prime dimension the phase point operators

[6] A more sophisticated mapping might also make use of classical randomness, classical feedforward and more complicated input states than $\rho^{\otimes n}$. It turns out that no advantage can be gained from these techniques, and this is proven formally for the qubit case in [38].

correspond to a proper subset of the facets of the polytope with stabilizer states as its vertices. This implies the existence of states with non-negative representation which are not convex combinations of stabilizer states, the bound states for magic state distillation. This also generalizes the results of van Dam and Howard [211] who have used techniques of this type to derive a bound on the amount of depolarizing noise a state can withstand before entering the stabilizer polytope for systems of prime dimension.

For completeness, let us briefly review the important properties of the Clifford group for qudits and the discrete Gross-Wigner representation (which is an instance of those reviewed in B.2.5). The Clifford group on n qudits is the group of unitary operators which permute the phase space point operators. These are the only allowed unitary operations in magic state distillation. For quantum systems of odd Hilbert space dimension we define the discrete Gross-Wigner function using the phase point operators $\{A_\alpha\}$

$$A_{(0,0)} \;=\; \sum_\alpha T_\alpha,\ A_\alpha = T_\alpha A_{(0,0)} T_\alpha^\dagger,$$

where $\{T_\alpha\}$ are the Heisenberg-Weyl operators and $\alpha \in \mathbb{F}_d \times \mathbb{F}_d$ is an indexing vector. There are d^2 such operators for d-dimensional Hilbert space, they are informationally complete and orthogonal in the sense $\mathrm{Tr}(A_\alpha A_\beta) = d\delta(\alpha, \beta)$. These operators have several important features [96, 86]: (1) (Discrete Hudson's theorem) if $|S\rangle$ is a stabilizer state then $\mathrm{Tr}(A_\alpha |S\rangle\langle S|) \geq 0\ \forall \alpha$, and the stabilizer states are the only pure states satisfying this property; (2) the Clifford operators preserve the set of states satisfying $\mathrm{Tr}(\rho A_\alpha) \geq 0$; (3) for $\rho = \sum_\alpha p_\alpha A_\alpha$ and $\sigma = \sum_\alpha q_\alpha A_\alpha$ the trace inner product is $\mathrm{Tr}(\rho\sigma) = d \sum_\alpha p_\alpha q_\alpha$; and (4) the phase point operations in dimension d^n are tensor products of n copies of the d dimension phase point operators (B.35).

A magic state distillation routine consumes n copies of an ancilla state ρ to prepare a state ρ_{out} that has greater fidelity with respect to some pure, non-stabilizer state. At each step of a distillation protocol a Clifford operation is applied to $\rho^{\otimes n}$ and an error syndrome measurement is made on the last $n-1$ systems. This later step is equivalent to making a projective measurement of stabilizer states on the last $n-1$ qudits and post selecting on the outcome. We have shown that operations of this form preserve the non-negativity of the representation of ρ.

Formally, define:

$$\begin{aligned} F(\rho) &= \mathrm{min}_\alpha \mathrm{Tr}(A_\alpha \rho), \\ F(\rho_{\text{out}}) &= \mathrm{min}_\alpha \frac{\mathrm{Tr}(A_\alpha \otimes |0\rangle\langle 0|^{\otimes n-1} U \rho^{\otimes n} U^\dagger)}{\mathrm{Tr}(\mathbb{I} \otimes |0\rangle\langle 0|^{\otimes n-1} U \rho^{\otimes n} U^\dagger)}. \end{aligned}$$

The latter quantity is the negativity of the representation of ρ_{out}, the state achieved by applying the Clifford operation U to $\rho^{\otimes n}$ and doing a measurement with post selection on an arbitrary stabilizer state, where the arbitrary freedom is rolled into the Clifford operation U. A state ρ has non-negative representation if and only if $F(\rho) \geq 0$.

By showing $F(\rho) \geq 0 \implies F(\rho_{\text{out}}) \geq 0$ for any choice of distillation protocol we establish that states with non-negative representation are not distillable. This holds even allowing for repeated distillation operations of different types. The main theorem of [215] is as follows:

Theorem A.2.1. *A necessary condition for a state ρ to be distillable is that ρ has negative representation.*

As promised, this results comes with a peculiar bonus feature: there exists non-negative states which lie outside the stabilizer polytope. These states, which cannot be prepared via stabilizer operations but also cannot be distilled, have been dubbed *bound states* [39]. The set of convex combinations of stabilizer states is a convex polytope with the stabilizer states as vertices. The stabilizer polytope may be thought of as a bounded convex polytope living in $\mathbb{R}^{d^2-1}$, the space of d dimensional mixed quantum states. This means that there are states that may not be written as a convex combination of stabilizer states which nevertheless satisfy $\text{Tr}(A_\alpha \rho) \geq 0$ for all phase point operators. That is, there are non-negative states which are not in the convex hull of stabilizer states. These are bound states for magic state distillation. An explicit example of such a state for the qutrit is given in [96]. This volume of bound states is depicted in figure A.1.

A.2.2 Negativity is a Necessary Condition for Continuous Variable Quantum Computation

Quantum information is usually thought of in terms of qubits: 2-level systems or *quantum* bits. But we can do equally non-classical computation via qu*trits*, qu*dits* (finite d-level systems) or even qu*modes* – which are systems of continuous variables. The most commonly considered system of this latter type is a quantized electromagnetic field. Although we will discuss continuous variable quantum computation in the context of optics, we note that the mathematical models apply equally well to other such systems: vibration modes of a solid, Bose-Einstein condensates and large ensembles of atomic systems, to name a few.

Many of the seminal results in discrete variable quantum computation have analogs in the continuous variable setting. Perhaps the most important example is the "continuous

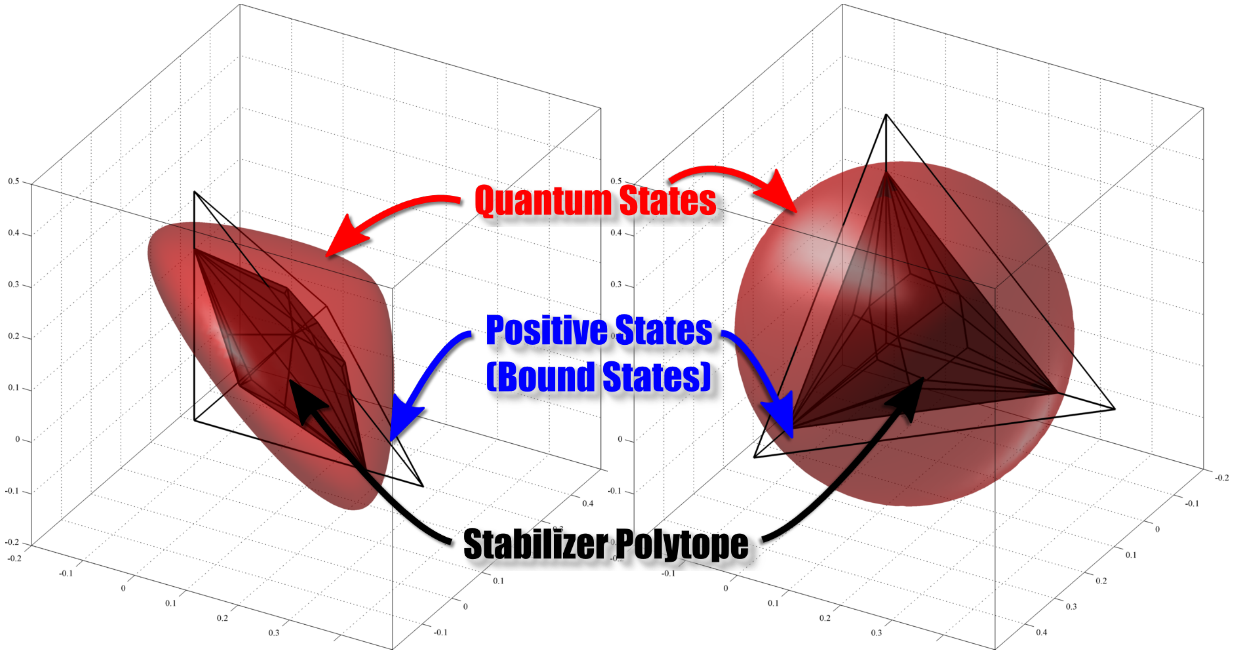

Figure A.1: Bound states for magic state distillation. Shown is a slice through the 8 dimensional qutrit state space from two different perspective. The point $(1/3, 1/3, 1/3)$ in the shown slice of this representation is the identity state. The inner polytope is the stabilizer polytope and the larger polytope is that of the positive Wigner function. The red object is the boundary of the qutrit state space.

variable Gottesman-Knill theorem", which states that a computation restricted to the subset of quantum theory containing only *Gaussian states* is classically efficiently simulatable [15]. The Gottesman-Knill theorem [92], on the other hand, states that a quantum computation restricted to *stabilizer states* is classically efficiently simulatable. This seems very surprising – at least at first sight – when juxtaposed with the following two facts: all known error-correction schemes consist mostly of elements of this stabilizer formalism and only a single type of unitary operation needs to be added to the formalism to enable universal quantum computation. The caveat here is that the standard stabilizer formalism requires pure states. An intuitively satisfying results is that allowing classical randomness to the preparation (allowing states in the convex hull of stabilizer states) affords no improvement in computational power of the model [34].

A natural extension then is to conceive of a resource theory for the stabilizer formalism.

Indeed, Bravy and Kitaev showed that repeated preparations of certain non-stabilizer mixed ancilla states can be *distilled* using only elements of the stabilizer formalism to enable universal quantum computation [30]. With these results in hand, we have two conditions for a resource state to allow for universal quantum computation via distillation: (1) it is necessary that the resource state be a non-stabilizer state, and (2) it is sufficient that the resource have fidelity above some threshold value with one of a set of pure "magic states". Unfortunately, these results hold only for qubits. But as we might expect, distillation is possible for systems of larger dimension. Indeed, Anwar *et al.* recently provided a sufficient condition by way of an explicit protocol [10] – while the necessary condition that stabilizer states are insufficient remains the same. But, as we crank up the dimension of the individual systems, the unexplored region between these two conditions becomes vast. Recently, we found a prodigious extension of the necessary boundary beyond the stabilizer states, as discussion in the previous section and reference [215]. This implies the existence of a large class of *bound states*: states which cannot be prepared via the stabilizer formalism but do not enable universal quantum computation.

Here we provide the analogous result for continuous variable systems. That is, we define a quantum computational resource theory for Gaussian quantum theory and show that there exists a large class of non-Gaussian bound states. Moreover, we show that the classical efficient simulatability of Gaussian quantum theory extends to this boundary. This is appealing because the boundary is given by those states with positive Wigner function and thus we argue that this gives the first operational interpretation of negativity of the Wigner function as a non-classical feature of the quantum state.

First, we will briefly review continuous variables à la quantum optics. Standard quantization of an electromagnetic field, with N modes indexed by k, results in a set of harmonic oscillators [220]. The dimensionless Hamiltonian of a single mode k reads

$$H_k = \frac{1}{2}(P_k^2 + Q_k^2),$$

operators satisfy the canonical commutation relations $[Q_k, P_j] = i\delta_{kj}$, $[Q_k, Q_j] = 0$, and $[P_k, P_j] = 0$ It is also common to treat the oscillator in terms of the *creation and annihilation operators*

$$\begin{aligned} a_k &= \frac{1}{\sqrt{2}}(Q_k + iP_k), \\ a_k^\dagger &= \frac{1}{\sqrt{2}}(Q_k - iP_k). \end{aligned}$$

A transformation is usually considered to be contained in linear optics if its Hamiltonian commutes with $H = \sum_{k=1}^N H_k$. i.e. it can be written $H_{\rm LO} = \boldsymbol{a}^\dagger \cdot M\boldsymbol{a}$ for some $N \times N$

Hermitian matrix M. Such transformations preserve the overall photon number and are called *passive.* In the Heisenberg picture it can be represented by

$$a_k \mapsto \sum_{j=1}^{N} U_{kj} a_j,$$

where U is an $N \times N$ unitary matrix. Note that any such unitary can be efficiently decomposed as a sequence of beam splitters and phase shifters [176].

But note also for linear optics Hamiltonians there are no terms which mix the Q's and P's. Although the Hamiltonian is a quadratic polynomial is the joint vector $\boldsymbol{x} = \boldsymbol{q} \oplus \boldsymbol{p}$, it is not the most general one. For that, we define *linear optics and squeezing* (which will refer to as just linear optics herein) to be the set of transformation whose Hamiltonians, in terms of $\boldsymbol{x}$, are exactly those that can be written as a quadratic polynomial:

$$H_{\text{LOS}} = \boldsymbol{x} \cdot S\boldsymbol{x} + V \cdot \boldsymbol{x},$$

where S is $2N \times 2N$ real symmetric matrix and V is a real $2N$-dimensional vector. In terms of the creation and annihilation operators in the Heisenberg picture, these transformation induce mappings of the form

$$a_k \mapsto \sum_j A_{kj} a_j + B_{kj} a_j^{\dagger} + \gamma_j,$$

where the matrices A and B satisfy $AB^{\text{T}} = (AB^{\text{T}})^{\text{T}}$ and $AA^{\dagger} = BB^{\dagger} + \mathbb{1}$ to preserve the commutation relations.

With these definitions we can now conveniently define a non-linear process. That is, we define *non-linear optics* as the set of transformations whose Hamiltonians are polynomials (in the elements of $\boldsymbol{x}$) whose degree is strictly greater than 2. A classical example of such a non-linear process is given by a *Kerr non-linearity,*

$$H_{\text{Kerr}} = (Q^2 + P^2)^2,$$

which was used to show to that quadratic Hamiltonians and any non-linear optics Hamiltonian generate the algebra of all polynomial Hamiltonians [144].

The eigenbasis of k'th harmonic oscillator Hamiltonian is $\{|n\rangle_k\}$ with $a_k |n\rangle_k = \sqrt{n}\, |n-1\rangle_k$ and $a^{\dagger} |n\rangle_k = \sqrt{n+1}\, |n+1\rangle_k$. The ground state $|0\rangle_k$ is called the *vacuum state* and in the quantum optics context, $|1\rangle_k$ is a *single photon* state, $|2\rangle_k$ is a *two photon* state, and so on. In general, they are called either *number states* or *Fock states.* The eigenstates of the annihilation operator, called *coherent states*, also play a particularly important role in quantum optics. In the number state basis, the state defined as $a\left|\alpha\right\rangle_k = \alpha\left|\alpha\right\rangle_k$ is explicitly

$$\left|\alpha\right\rangle_k = e^{-\frac{|\alpha|^2}{2}} \sum_n \frac{\alpha^n}{\sqrt{n!}} \left|n\right\rangle_k ,$$

where α is in general a complex number. It is convenient to write the coherent states as *displaced* vacuum states: $\left|\alpha\right\rangle_k = D_k(\alpha)\left|0\right\rangle_k$, where

$$D(\alpha) = e^{\alpha a_k^\dagger - \alpha^* a_k}$$

is called the *displacement operator*. Through this operator we can define one more class of states: *squeezed states*. First, we define a squeezed vacuum state as

$$\left|r, \phi\right\rangle_k = e^{r(e^{-2i\phi}a_k^2 - e^{2i\phi}a_k^{\dagger 2})}\left|0\right\rangle_k ,$$

for some real numbers r, ϕ. Then the general squeezed states are displacements: $\left|\alpha, r, \phi\right\rangle_k = D_k(\alpha)\left|r, \phi\right\rangle_k$.

Phase space representations have a strong tradition in quantum optics. The phase space distributions of quantum optics are reviewed in section B.1. Here we will focus on the Wigner function. Hudson's theorem [116] was the first attempt to characterize the positive Wigner functions and it was later generalized to the following [198]. Let ψ be a pure quantum state of N oscillators. Then its Wigner function is positive if and only if

$$\psi(\boldsymbol{Q}) = e^{-\frac{1}{2}(\boldsymbol{Q}\cdot A\boldsymbol{Q} + B\cdot\boldsymbol{Q} + c)},$$

where A is an $N \times N$ Hermitian matrix, B is an N-dimensional complex vector and c is a normalization constant. In our quantum optics terminology, these turn out to be either coherent states or squeezed states. Plugging these state into the definition of the Wigner function yield multivariate Gaussian distributions in phase space.

Convex combinations of these states (incoherent mixtures of them) will also have positive Wigner function since the mapping is linear. Early on, these were conjectured to be the only such mixed states with positive Wigner function. However, an explicit example of a quantum state with positive Wigner function that is not a convex combination of Gaussian states is [32]:

$$\rho = \frac{1}{2}(|0\rangle\langle 0| + |1\rangle\langle 1|).$$

This state is depicted in figure A.2.

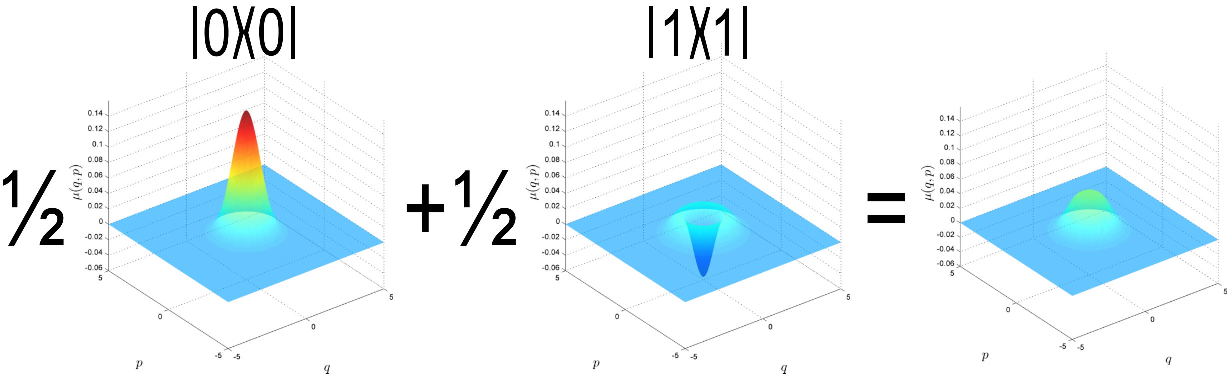

Figure A.2: A state with positive Wigner function that is not a convex combination of Gaussian states.

The question of mixed states was given a full treatment in reference [203] and later in [32]. Both references independently found that a theorem in classical probability attributed to Bochner [25] and generalization thereof can be used to characterize both the valid Wigner functions and the subset of positive ones. The point here is that there exist a large class of states with positive Wigner function that are not convex combinations of Gaussian states. So far, these states have received little attention. Next we show how they play a role more prominent than a mere mathematical curiosity.

Gaussian quantum information is defined to be the following set of operators (see, for example, [221]): M mode Gaussian input state; linear optics Hamiltonians; and, measurements with (or without) post-selection onto Gaussian states. If the ancilla system is a Gaussian state then the model above classically efficiently simulatable [15] (i.e. there exists an efficient, in the number of modes M, algorithm reproducing the output probabilities of the measurement results). Thus, some non-Gaussian resource is needed. We allow, in analogy with [30], that an arbitrary number N of copies of some state ρ to be acted upon via Gaussian quantum mechanics and ask if performing a measurement on the last $N-1$ copies of ρ results in a state arbitrarily close to some pure state for the remaining copy. Such a procedure is called *distillation* and if it succeeds, we call ρ *distillable.*

Our main result is as follows: a necessary condition for a state ρ to be distillable is that it have negative Wigner function. First we argue that any state with positive Wigner function remains so under linear optics. This happens to have a convenient classical interpretation as linear optics Hamiltonians induce affine transformations on the underlying phase space. That is, any linear optics unitary can be written as a linear transformation followed by a phase space translation. This fact goes all the way back to Moyal [160] who showed that quadratic Hamiltonians induce the following dynamics on the Wigner function:

$$\frac{\partial P(Q,P;t)}{\partial t} = \frac{\partial H(Q,P)}{\partial Q}\frac{\partial P(Q,P;t)}{\partial P} - \frac{\partial P(Q,P;t)}{\partial Q}\frac{\partial H(Q,P)}{\partial P}.$$

This is a completely classical dynamical trajectory given by Hamiltonian flow in Wigner representation. Via Liouville's theorem, the probability is conserved along trajectories and thus the Wigner function remains positive.

Now it remains to show that N copies of such a state remains positive after an arbitrary linear optics transformation followed by a projective measurement onto Gaussian states. As we have argued, the positivity of the initial state is preserved after the transformation. Since the Wigner mapping is self-dual, Gaussian measurements are also represented as positive functions. The final state is the result of the overlap of two positive functions and is hence positive.

The above argument also implies the following: Gaussian quantum information supplemented with ancilla states with positive Wigner function is classically efficiently simulatable. Since the Wigner function is positive, we are free to interpret it as a probability *distribution*. Now, although it requires an exponential number of resources to deterministically evolve the Wigner *distribution*, we can sample from it via Monte Carlo simulations and evolve the phase space point under the Hamiltonian evolution – as is done in the simulation of classical dynamical systems. Thus, we can efficiently simulate the outcomes of Gaussian measurements of physical systems evolving under linear optical operations when the initial state is represented by a positive Wigner function. At first sight this seems counter-intuitive since we are taught that *points* in the quantum phase space are meaningless since they maximally violate the uncertainty principle. This is a restriction, however, that the classical physical system carrying out the simulation need not be bound by.

Reference [14] nicely summarized what was known at the time about continuous variable quantum computation. The table presented there is reproduced below in Table A.1 with some more recent results. The field began with Lloyd and Braunstein's observation that non-linear optical processes are sufficient for UQC. Later, it was shown that linear optics is sufficient provided photon counting measurements are available [130, 93]. The continuous variable analog of the *measurement-based* model shows that preparation of single photon state preparation is also sufficient [154]. More recently, the result of Aaronson and Arkhipov [1] shows that preparing and measuring single photon states (without post-selection) is equivalent to problem that is thought to be hard classically – but it still manages to (probably) not be universal for quantum computation.

It is possible that the Aaronson and Arkhipov model may be intermediately between classically efficiently simulatable and universal for quantum computation. Another suspected model of this type is the DQC1 model of Knill and Laflamme [128]. The key point for this latter model is that uses highly mixed states. Mixed states have not been formally consider for continuous variable quantum computation. Here we have generalized the Bravyi and Kitaev magic state model to the continuous variable domain. We have shown, via the Wigner phase space formalism, that negative representation is necessary for universal quantum computation. Moreover, any computation that uses states possessing a positive Wigner function is classically efficiently simulatable. It would be quite interesting if this condition turned out also to be sufficient as this would provide a sharp boundary between quantum and classical systems with regard to their computational power.

As we have seen in the previous section, an analogous result holds also for collections of systems of odd dimension. However, such a class of bound states does not appear to exists for the qubit case. This provides one more piece of evidence that ∞ is indeed closer to 3 as it is to 2.

Preparations	**Gates**	**Measurement**	**Efficiently simulatable classically**
Vacua	Linear optics	Gaussian	✓ [15, 13]
Vacua	Non-linear optics	Gaussian	✗ [144]
Single photons	Linear optics (no squeezing)	Photon counting (with post-selection)	✗ [130]
Vacua	Linear optics	Gaussian and Photon counting (with post-selection)	✗ [93]
Single photons	Linear optics	Gaussian	✗ [99]
Single photons	Linear optics (no squeezing)	Photon counting	✗ [1]
Independent positive Wigner distributions	Linear optics	Gaussian	✓ (this work)

Table A.1: Simulation results for continuous variable quantum computation. An extension of the table appearing in [14].

Appendix B

Quasi-probability Function Review

This chapter reproduces the portions of reference [67] which reviewed the known quasi-probability representations of quantum theory along with examples of their application to problems in quantum information science. Section B.1 is devoted to the infinite dimensional setting while section B.2 reviews the finite dimensional analogs.

B.1 Quasi-probability in Infinite Dimensional Hilbert Space

Here will we review the quasi-probability distributions which have been defined for quantum states living in an infinite dimensional Hilbert space - the canonical example being a particle moving in one dimension. Since there are a myriad of excellent reviews of the Wigner function and other phase space distributions [1], our discussion of them will be brief. We will mainly focus on those details which have inspired analogous methods for finite dimensional Hilbert spaces.

First we start with the familiar Wigner function in section B.1.1. The other phase space distributions, such as the Husimi function, are bundled up in section B.1.2.

[1] See reference [112] for a classic and reference [136] for a more recent review of phase space quasi-probability distributions.

B.1.1 Wigner Phase Space Representation

The position operator Q and momentum operator P are the central objects in the abstract formalism of infinite dimensional quantum theory. The operators satisfy the canonical commutation relations

$$[Q, P] = i.$$

We are looking for a joint probability distribution $\mu_\rho(p, q)$ for the state of the quantum system. From the postulates of quantum mechanics we have a rule for calculating expectation values. In particular, we can compute the characteristic function

$$\phi(\xi, \eta) := \left\langle e^{i(\xi q + \eta p)} \right\rangle = \mathrm{Tr}(e^{i(\xi Q + \eta P)} \rho).$$

Since the characteristic function is just the Fourier transform of the joint probability distribution, we simply invert to obtain

$$\mu_\rho(p, q) = \frac{1}{(2\pi)^2} \iint_{\mathbb{R}^2} \mathrm{Tr}(e^{i(\xi Q + \eta P)} \rho) e^{-i(\xi q + \eta p)} d\xi d\eta, \tag{B.1}$$

which is the celebrated *Wigner function* of ρ [224]. The Wigner function is both positive and negative in general. However, it otherwise behaves as a classical probability density on the classical phase space. For these reasons, the Wigner function and others like it came to be called *quasi-probability* functions.

The Wigner function is the unique representation satisfying the properties [20]

Wig(1) For all ρ, $\mu_\rho(q, p)$ is real.

Wig(2) For all ρ_1 and ρ_2,

$$\mathrm{Tr}(\rho_1 \rho_2) = 2\pi \int_{\mathbb{R}^2} dq dp \; \mu_{\rho_1}(q, p) \mu_{\rho_2}(q, p).$$

Wig(3) For all ρ, integrating μ_ρ along the line $aq + bp = c$ in phase space yields the probability that a measurement of the observable $aQ + bP$ has the result c.

We can write the Wigner function as

$$\mu_\rho(p, q) = \mathrm{Tr}\left[F(p, q)\rho\right],$$

where

$$F(q, p) := \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} d\xi d\eta \; e^{i\xi(Q-q) + i\eta(P-p)}. \tag{B.2}$$

Thus the properties Wig(1)-(3) can be transformed into properties on a set of operators $F(q, p)$ which uniquely specify the set in Equation (B.3). These properties are

Wig(4) $F(q,p)$ is Hermitian.

Wig(5) $2\pi\text{Tr}(F(q,p)F(q',p')) = \delta(q-q')\delta(p-p')$.

Wig(6) Let P_c be the projector onto the eigenstate of $aQ+bP$ with eigenvalue c. Then,

$$\int_{\mathbb{R}^2} dqdp\ F(q,p)\delta(aq+bp-c) = P_c.$$

These six properties are often the basis for generalizing the Wigner function to finite dimensional Hilbert spaces, as we will soon see.

Applications: quantum teleportation

The applications of the Wigner function are far reaching and not limited to to physics [2]. A modern application can be found in reference [42] where Caves and Wódkiewicz use the Wigner function to obtain a hidden variable model of the continuous-variable teleportation protocol [210, 29]. Later, in section B.2.6, we will discuss the much simpler discrete-variable teleportation protocol. Here, then, we will avoid the details of the protocol and focus on the result. It suffices to know the following: there are three quantum systems; the goal of the protocol is to transfer a quantum state from system 1 to system 3; the transfer is mediated through the special correlations between system 2 and system 3. The success of the protocol is measured by the average *fidelity*: a measure of the closeness of the initial state $\rho = |\psi\rangle\langle\psi|$ of system 1 and the average final state ρ_{out} of system 3.

Following Caves and Wódkiewicz we define $\nu = q + ip$ and index the Wigner function as $\mu_\rho(\nu)$. This is convenient since the protocol is tailored to a quantum optical implementation where the outcomes of measurements are usually expressed as complex numbers. Initially, the state of system 2 and 3 is described by the joint Wigner function $\mu_{2,3}(\alpha,\beta)$.

In terms of the Wigner functions, the average fidelity is

$$\mathcal{F} = \pi \int d^2\nu d^2\beta \mu_\rho(\nu)\mu_{\rho_{\text{out}}}(\beta).$$

This intuitively measures closeness by quantifying the overlap of the Wigner functions on the classical phase space. The output state, determined by the details of the protocol, is

$$\mu_{\rho_{\text{out}}}(\beta) = \int d^2\nu d^2\alpha \mu_\rho(\beta-\nu)\mu_{2,3}(\alpha,\nu-\overline{\alpha}).$$

[2] For example, see reference [63] for a recent review of the applications of the Wigner function in signal processing.

The initial Wigner function $\mu_\rho(\nu)$ and the joint Wigner function $\mu_{2,3}(\alpha,\beta)$ are determined by the particular implementation of the protocol. The standard quantum optical implementation is done using coherent states of light. It is easy to show that such states have positive Wigner functions [3]. Thus, the Wigner function provides a classical phase space picture of the entire protocol.

A first step toward performing an experiment requiring genuine quantum resources might be to avoid the above classical description by teleporting a non-coherent quantum state. Caves and Wódkiewicz have devised a classical explanation for this case as well. The new model involves a randomization procedure which transforms the initial non-coherent state into a coherent one thus giving it a positive Wigner function. However, it can be shown that within such a model, the fidelity is bounded: $\mathcal{F} < 2/3$. So $2/3$ emerges as a "gold-standard" since teleporting a non-coherent state with fidelity $\mathcal{F} \geq 2/3$ avoids this classical phase space description.

B.1.2 Other Phase Space Representations

Another class of solutions to the ordering problem is the association $e^{i\xi q+i\eta p} \mapsto e^{i\xi Q+i\eta P} f(\xi,\eta)$ for some arbitrary function f [4].

Consider again the classical particle phase space $\mathbb{R}^2$ and the continuous set of operators

$$F(q,p) := \frac{1}{(2\pi)^2}\int_{\mathbb{R}^2} d\xi d\eta \; e^{i\xi(q-Q)+i\eta(p-P)} f(\xi,\eta). \tag{B.3}$$

The f dependent distributions

$$\mu_\rho^f(q,p) := \mathrm{Tr}(\rho F(q,p)) \tag{B.4}$$

define quasi-probability functions on phase space alternative to the Wigner function, which is simply the $f = 1$ special case of this more general formalism.

Besides the Wigner function, the most popular choices of f are

$$f(\xi,\eta) = e^{\pm\frac{1}{4}(\xi^2+\eta^2)},$$

which give, via equation (B.4), the Glauber-Sudarshan [90, 206] and Husimi [117] functions, respectively. These two mappings are sometimes referred to as the P- and Q-representations

[3] More difficult is to show that coherent states are the *only* states with positive Wigner functions [116].

[4] This is very closely related to the s-ordered Cahill-Glauber formalism [37]. See Table 1 of [136] for a concise review of the traditional choices for f.

(not to be confused with position and momentum representations). We will follow the usual convention by introducing the annihilation operator

$$a = \frac{1}{\sqrt{2}}(Q + iP) \tag{B.5}$$

and the *coherent states* defined via $a\,|\alpha\rangle = \alpha\,|\alpha\rangle$, where we write $\alpha = q + ip$. Then the Husimi function can be conveniently written

$$Q(\alpha) := \mu_\rho^f(q,p) = \frac{1}{\pi}\langle \alpha, \rho\alpha\rangle. \tag{B.6}$$

The Glauber-Sudarshan function $\rho \to P(\alpha)$ can be expressed implicitly through the identity

$$\rho = \int d^2\alpha P(\alpha)|\alpha\rangle\langle\alpha|, \tag{B.7}$$

where $d^2\alpha = (1/2)dqdp$. Notice that this immediate implies the following *duality* condition between the P- and Q-representation:

$$\mathrm{Tr}(\rho\rho') = \int d^2\alpha P_\rho(\alpha)Q_{\rho'}(\alpha). \tag{B.8}$$

Application: quantumness witness

Here we will discuss a more recent application of the P and Q functions of interest in quantum information and foundations [5]. We will concern ourself with a particular notion of "non-classicality" defined in reference [7].

Consider two observables represented by the self-adjoint operators R and S. Another observable $W(R,S)$ written as an ordered power series of R and S is a *quantumness witness* if it possesses at least one negative eigenvalue and the function $w(r,s)$ obtained by replacing R and S with its spectral elements is positive: $w(r,s) \geq 0$ for all $r \in \mathrm{spec}(R)$ and $s \in \mathrm{spec}(S)$.

As an example consider $R, S \geq 0$ and define $A := R$, $B := R + S$ and $V := W(R,S) = S^2 + RS + SR$. Then, V possess a negative eigenvalue and $w(r,s) = s^2 + 2rs \geq 0$ for $r, s \geq 0$. As proven in reference [8], $0 \leq A \leq B$ implies $A^2 \leq B^2$ if and only if the algebra of observables is commutative, which is generally agreed upon as a necessary requirement

[5] The P and Q functions are powerful visualization tools prominently used in the areas of quantum optics and quantum chaos [207]. See [188] for an overview of the applications in quantum optics.

for classicality. However, the algebra of quantum observables is such that their exists $0 \leq A \leq B$ but $A^2 \geq B^2$. Thus, if we can measure V and find $\langle V \rangle \leq 0$, then we have shown that the system being measured cannot admit a classical model and we have found a signature of quantumness.

Now consider $R = Q$ and $S = P$, the usual position and momentum operators. Recalling the parameterization in terms of the annihilation operator in equation (B.5), we define

$$W(a) := \sum_{m,n} c_{mn} (a^\dagger)^m a^n$$

such that

$$w(\alpha) = \sum_{m,n} c_{mn} \overline{\alpha}^m \alpha^n \geq 0$$

and W possesses at least one negative eigenvalue (a concrete example is $(a^\dagger)^2 a^2 - 2ma^\dagger a + m^2$ for $m \geq 1$). Then, in terms of the Q and P-representation defined above, we have

$$\langle W \rangle = \mathrm{Tr}(\rho W) = \int d^2\alpha \langle \alpha, W\alpha, P \rangle(\alpha) = \int d^2\alpha w(\alpha) P(\alpha). \tag{B.9}$$

In optics especially, coherent states are considered classical. From equation (B.7) we see that if $P(\alpha) \geq 0$, the quantum state is a statistical mixture of coherent states and hence just as classical. So if ρ is a classical state, equation (B.9) tells us $\langle W \rangle \geq 0$. Therefore, if we measure $\langle W \rangle \leq 0$, we can rule out the classical model of statistical mixtures of coherent states; we can say $W(a)$ detects the quantumness of the states.

B.2 Quasi-probability in Finite Dimensional Hilbert Space

Nearly all definitions of quasi-probability distributions for finite dimensional Hilbert spaces have been motivated by the Wigner function. The earliest such effort was by Stratonovich and is reviewed in section B.2.1. The Stratonovich phase space is a sphere and hence continuous. Later, many authors have define Wigner function analogs on *discrete* phase spaces. A sampling, with a bias towards those which have found application in quantum information theory, is given in sections B.2.2-B.2.6.

There also exist quasi-probability distributions which where introduced to solve various problems far removed from proposing a finite dimensional analog of the Wigner function.

Sections B.2.7-B.2.10 review those quasi-probability distributions which do not have a canonical phase space structure and hence form a somewhat weaker analogy to the Wigner function.

We note that there exists many other quasi-probability distributions defined on discrete phase spaces which are not reviewed here [103, 51, 74, 83, 165, 147, 101, 179, 161, 218, 44, 45, 96, 97].

B.2.1 Spherical Phase Space

Here we will be concerned with a set of postulates put forth by Stratonovich [204]. The aim of Stratonovich was to find a Wigner function type mapping, analogous to that of a infinite dimensional system on $\mathbb{R}^2$, of a d dimensional system on the sphere $\mathbb{S}^2$. The first postulate is linearity and is always satisfied if the Wigner functions on the sphere satisfy

$$\mu_\rho(\mathbf{n}) = \mathrm{Tr}(\rho\triangle(\mathbf{n})), \tag{B.10}$$

where $\mathbf{n}$ is a point on $\mathbb{S}^2$. The remaining postulates on this quasi-probability mapping are

$$\begin{aligned}
&\mu_\rho(\mathbf{n})^* = \mu_\rho(\mathbf{n}),\\
&\frac{d}{4\pi}\int_{\mathbb{S}^2} d\mathbf{n}\ \mu_\rho(\mathbf{n}) = 1,\\
&\frac{d}{4\pi}\int_{\mathbb{S}^2} d\mathbf{n}\ \mu_{\rho_1}(\mathbf{n})\mu_{\rho_2}(\mathbf{n}) = \mathrm{Tr}(\rho_1\rho_2),\\
&\mu_{(g\cdot\rho)}(\mathbf{n}) = \mu_\rho(\mathbf{n})^g,\ g \in \mathrm{SU}(2),
\end{aligned}$$

where $g\cdot\rho$ is the image of $U_g\rho U_g^\dagger$ and $U : \mathrm{SU}(2) \to \mathbb{U}\mathcal{H}$ is an irreducible unitary representation of the group SU(2). These postulates are analogous to Wig(1)-(3) for the Wigner function modulo the second normalization condition (which could have be included in the Wigner function properties).

The continuous set of operators $\triangle(\mathbf{n})$ is called a *kernel* and we note it plays the role of the more familiar *phase space point operators* in the latter. Requiring that Equation

(B.10) hold changes the postulates to new conditions on the kernel

$$\triangle(\mathbf{n})^\dagger = \triangle(\mathbf{n}), \tag{B.11}$$
$$\frac{d}{4\pi}\int_{\mathbb{S}^2} d\mathbf{n}\, \triangle(\mathbf{n}) = \mathbb{1}, \tag{B.12}$$
$$\frac{d}{4\pi}\int_{\mathbb{S}^2} d\mathbf{n}\, \mathrm{Tr}(\triangle(\mathbf{n})\triangle(\mathbf{m}))\triangle(\mathbf{n}) = \triangle(\mathbf{m}), \tag{B.13}$$
$$\triangle(g\cdot\mathbf{n}) = U_g\triangle(\mathbf{n})U_g^\dagger,\ g\in \mathrm{SU}(2). \tag{B.14}$$

These postulates are the spherical analogies of properties Wig(4)-(6) (again, modulo the normalization condition). Heiss and Weigert [108] provided a concise derivation of 2^{2s}, where $s = \frac{d-1}{2}$ is the *spin*, unique kernels satisfying these postulates [6]. They are

$$\triangle(\mathbf{n}) = \sum_{m=-s}^{s}\sum_{l=0}^{2s}\epsilon_l\frac{2l+1}{2s+1}C^{s\,l\,s}_{m\,0\,m}\phi_m(\mathbf{n})\phi^*_m(\mathbf{n}), \tag{B.15}$$

where C denotes the Clebsch-Gordon coefficients. Here, $\phi_m(\mathbf{n})$ are the eigenvectors of the operator $\mathbf{S}\cdot\mathbf{n}$, where $\mathbf{S} = (X, Y, Z)$; and $\epsilon_l = \pm1$, for $l = 1\ldots 2s$ and $\epsilon_0 = 1$.

Heiss and Weigert relax the postulates Equations (B.11)-(B.14) on the kernel $\triangle(\mathbf{n})$ to allow for a pair of kernels $\triangle^{\mathbf{n}}$ and $\triangle_{\mathbf{m}}$. The pair individually satisfy Equation (B.11), while one of them satisfies Equation (B.12) and the other Equation (B.14). Together, the pair must satisfy the generalization of Equation (B.13)

$$\frac{d}{4\pi}\int_{\mathbb{S}^2} d\mathbf{n}\mathrm{Tr}(\triangle^{\mathbf{n}}\triangle_{\mathbf{m}})\triangle^{\mathbf{n}} = \triangle_{\mathbf{m}}. \tag{B.16}$$

A pair of kernels, together satisfying Equation (B.16), is given by

$$\triangle_{\mathbf{n}} = \sum_{m=-s}^{s}\sum_{l=0}^{2s}\gamma_l\frac{2l+1}{2s+1}C^{s\,l\,s}_{m\,0\,m}\phi_m(\mathbf{n})\phi^*_m(\mathbf{n}),$$
$$\triangle^{\mathbf{n}} = \sum_{m=-s}^{s}\sum_{l=0}^{2s}\gamma_l^{-1}\frac{2l+1}{2s+1}C^{s\,l\,s}_{m\,0\,m}\phi_m(\mathbf{n})\phi^*_m(\mathbf{n}),$$

where each γ_l is a finite non-zero real number and $\gamma_0 = 1$. The original postulates are satisfied when $\gamma_l = \gamma_l^{-1} \equiv \epsilon_l$.

[6] This was also shown earlier [212, 31] - see also references [11, 192].

The major contribution of reference [108] is the derivation of a *discrete* kernel $\triangle_\nu := \triangle_{\mathbf{n}_\nu}$, for $\nu = 1 \dots d^2$ which satisfies the discretized postulates

$$\triangle_\nu^\dagger = \triangle_\nu, \tag{B.17}$$

$$\frac{1}{d}\sum_{\nu=1}^{d^2} \triangle^\nu = \mathbb{1}, \tag{B.18}$$

$$\frac{1}{d}\sum_{\nu=1}^{d^2} \mathrm{Tr}(\triangle_\nu \triangle^\mu)\triangle_\nu = \triangle^\mu, \tag{B.19}$$

$$\triangle_{g\cdot\nu} = U_g \triangle_\nu U_g^\dagger,\ g \in \mathrm{SU}(2). \tag{B.20}$$

The subset of points $\mathbf{n}_\nu$ is called a *constellation.* The linearity postulate is not explicitly stated since it is always satisfied under the assumption

$$\rho \to \mu_\rho(\nu) = \mathrm{Tr}(\rho\triangle_\nu). \tag{B.21}$$

Equation (B.19) is called a *duality* condition. That is, it is only satisfied if $\triangle_\nu$ and $\triangle^\mu$ are *dual bases* for $\mathbb{H}(\mathcal{H})$. In particular,

$$\frac{1}{d}\mathrm{Tr}(\triangle_\nu \triangle^\mu) = \delta_{\nu\mu}.$$

Although the explicit construction of a pair of discrete kernels satisfying Equations (B.17)-(B.20) might be computationally hard, their existence is a trivial exercise in linear algebra. Indeed, so long as $\triangle_\nu$ is a basis for $\mathbb{H}(\mathcal{H})$, its dual, $\triangle^\mu$, is uniquely determined by

$$\triangle^\mu = \sum_{\nu=1}^{d^2} \mathtt{G}^{-1}_{\nu\mu}\triangle_\nu,$$

where the Gram matrix $\mathtt{G}$ is given by

$$\mathtt{G}_{\nu\mu} = \mathrm{Tr}(\triangle_\nu \triangle_\mu).$$

The authors of reference [108] note that almost any constellation leads to a discrete kernel $\triangle_\nu$ forming a basis for $\mathbb{H}(\mathcal{H})$. The term *almost any* here means that a randomly selected discrete kernel will form, with probability 1, a basis for $\mathbb{H}(\mathcal{H})$.

Application: NMR quantum computation

The spherical quasi-probability functions for qubit systems ($d = 2$) were put to use by Schack and Caves for the purpose of obtaining a classical model of nuclear magnetic resonance (NMR) experiments designed to perform quantum information tasks [186]. For a single qubit we choose the kernels

$$\triangle_{\mathbf{n}} = \frac{1}{2}(\mathbb{1} + \mathbf{n} \cdot \sigma),$$
$$\triangle^{\mathbf{n}} = \frac{1}{4\pi}(\mathbb{1} + 3\mathbf{n} \cdot \sigma),$$

where $\sigma = (X, Y, Z)$ are the usual Pauli operators. In NMR experiments many qubits are employed to perform quantum information tasks such as error correction and teleportation. Suppose there are n qubits with total Hilbert space dimension 2^n. We choose an n-fold tensor product of the qubit kernels. Explicitly, they are

$$\triangle_{\mathbf{n}} = \frac{1}{2^n}\bigotimes_{j=1}^{n}(\mathbb{1} + \mathbf{n} \cdot \sigma), \tag{B.22}$$

$$\triangle^{\mathbf{n}} = \frac{1}{(4\pi)^n}\bigotimes_{j=1}^{n}(\mathbb{1} + 3\mathbf{n} \cdot \sigma). \tag{B.23}$$

The quasi-probability function is given by

$$\mu_\rho(\mathbf{n}) = \mathrm{Tr}(\rho\triangle^{\mathbf{n}}).$$

As expected, in general, this function is both positive and negative.

The quantum state of an NMR experiment is of the form

$$\rho = (1 - \epsilon)\frac{1}{2^n}\mathbb{1} + \epsilon\rho_1, \tag{B.24}$$

where ρ_1 is arbitrary but often chosen to be a specific pure state. The parameter scales as

$$\epsilon \propto \frac{n}{2^n}.$$

So we have

$$\mu_\rho(\mathbf{n}) = \frac{1 - \epsilon}{(4\pi)^n} + \epsilon\mu_{\rho_1}(\mathbf{n}).$$

It is easy to determine the lower bound

$$\mu_\rho(\mathbf{n}) \geq \frac{1-\epsilon}{(4\pi)^n} - \frac{\epsilon 2^{2n-1}}{(4\pi)^n}.$$

Thus, provided

$$\epsilon \leq \frac{1}{1+2^{2n-1}},$$

$\mu_\rho(\mathbf{n}) \geq 0$ and we have a representation of NMR quantum states in terms of classical probability distributions on a classical phase space. In reference [186] the authors note that, for typical experimental values of the scaling parameter in ϵ, such a classical representation is valid for $n < 16$ qubits.

The spherical phase space representations have also been put to good use in visualizing decoherence [85] and photon squeezing [194].

B.2.2 Wootters Discrete Phase Space Representation

In reference [225], Wootters defined a discrete analog of the Wigner function. Associated with each Hilbert space $\mathcal{H}$ of finite dimension d is a *discrete phase space*. First assume d is prime. The *prime phase space*, Φ, is a $d \times d$ array of points $\alpha = (q,p) \in \mathbb{Z}_d \times \mathbb{Z}_d$.

A *line*, λ, is the set of d points satisfying the linear equation $aq + bp = c$, where all arithmetic is modulo d. Two lines are *parallel* if their linear equations differ in the value of c. The prime phase space Φ contains $d+1$ sets of d parallel lines called *striations*.

Assume the the Hilbert space $\mathcal{H}$ has composite dimension $d = d_1 d_2 \cdots d_k$. The discrete phase space of the entire d dimensional system is the Cartesian product of two-dimensional prime phase spaces of the subsystems. The phase space is thus a $(d_1 \times d_1) \times (d_2 \times d_2) \times \cdots (\times d_k \times d_k)$ array. Such as construction is formalized as follows: the *discrete phase space* is the multi-dimensional array $\Phi = \Phi_1 \times \Phi_2 \times \cdots \times \Phi_k$, where each Φ_i is a prime phase space. A *point* is the k-tuple $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ of points $\alpha_i = (q_i, p_i)$ in the prime phase spaces. A *line* is the k-tuple $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_k)$ of lines in the prime phase spaces. That is, a line is the set of d points satisfying the equation

$$(a_1 q_1 + b_1 p_1, a_2 q_2 + b_2 p_2, \ldots, a_k q_k + b_k p_k) = (c_1, c_2, \ldots, c_k),$$

which is symbolically written $aq + bp = c$. Two lines are *parallel* if their equations differ in the value c. As was the case for the prime phase spaces, parallel lines can be partitioned into sets, again called striations; the discrete phase space Φ contains $(d_1+1)(d_2+1)\cdots(d_k+1)$ sets of d parallel lines.

The construction of the discrete phase space is now been complete. To introduce Hilbert space into the discrete phase space formalism, Wootters chooses the following special basis for the space of Hermitian operators. The set of operators $\{A_\alpha : \alpha \in \Phi\}$ acting on an d dimensional Hilbert space are called *phase point operators* if the operators satisfy

Woo(4) For each point α, A_α is Hermitian.

Woo(5) For any two points α and β, $\mathrm{Tr}(A_\alpha A_\beta) = d\delta_{\alpha\beta}$.

Woo(6) For each line λ in a given striation, the operators $P_\lambda = \frac{1}{d}\sum_{\alpha\in\lambda} A_\alpha$ form a projective valued measurement (PVM): a set of d orthogonal projectors which sum to identity.

Notice that these properties of the phase point operators Woo(4)-(6) are discrete analogs of the properties Wig(4)-(6) of the function F defining the original Wigner function. This definition suggests that the lines in the discrete phase space should be labeled with states of the Hilbert space. Since each striation is associated with a PVM, each of the d lines in a striation is labeled with an orthogonal state. For each Φ, there is a unique set of phase point operators up to unitary equivalence.

Although the sets of phase point operators are unitarily equivalent, the induced labeling of the lines associated to the chosen set of phase point operators are not equivalent. This is clear from the fact that unitarily equivalent PVMs do not project onto the same basis.

The choice of phase point operators in reference [225] will be adopted. For d prime, the phase point operators are

$$A_\alpha = \frac{1}{d}\sum_{j,m=0}^{d-1} \omega^{pj-qm+\frac{jm}{2}} X^j Z^m, \tag{B.25}$$

where ω is a d'th root of unity and X and Z are the generalized Pauli operators. For composite d, the phase point operator in Φ associated with the point $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ is given by

$$A_\alpha = A_{\alpha_1} \otimes A_{\alpha_2} \otimes \cdots \otimes A_{\alpha_k}, \tag{B.26}$$

where each A_{α_i} is the phase point operator of the point α_i in Φ_i.

The d^2 phase point operators are linearly independent and form a basis for the space of Hermitian operators acting on an d dimensional Hilbert space. Thus, any density operator ρ can be decomposed as

$$\rho = \sum_{q,p} \mu_\rho(q,p) A(q,p),$$

where the real coefficients are explicitly given by

$$\mu_\rho(q,p) = \frac{1}{d} \mathrm{Tr}(\rho A(q,p)). \tag{B.27}$$

This discrete phase space function is the Wootters *discrete Wigner function.* This discrete quasi-probability function satisfies the following properties which are the discrete analogies of the properties Wig(1)-(3) the original continuous Wigner function satisfies.

Woo(1) For all ρ, $\mu_\rho(q,p)$ is real.

Woo(2) For all ρ_1 and ρ_2,

$$\mathrm{Tr}(\rho_1 \rho_2) = d \sum_{q,p} \mu_{\rho_1}(q,p) \mu_{\rho_2}(q,p).$$

Woo(3) For all ρ, summing μ_ρ along the line λ in phase space yields the probability that a measurement of the PVM associated with the striation which contains λ has the outcome associated with λ.

Application: entanglement characterization

In [78], Franco and Penna relate the negativity of Wootter's discrete Wigner function to entanglement. Recall that a bipartite density matrix ρ is *separable* if it can be written as a convex combination of the form

$$\rho = \sum_k p_k \rho_k^{(1)} \otimes \rho_k^{(2)},$$

for all k, where $\rho_k^{(1)}$ and $\rho_k^{(2)}$ are states on the individual subsystems. Let Φ_1 and Φ_2 be the DPS associated with $\rho_k^{(1)}$ and $\rho_k^{(2)}$, respectively.

The Wootters representation of a density matrix of the form $\rho = \rho^{(1)} \otimes \rho^{(2)}$ is given by $\mu_\rho(\alpha) = \mu_\rho^{(1)}(\alpha_1)\mu_\rho^{(2)}(\alpha_2)$, where $\alpha = (\alpha_1, \alpha_2) \in \Phi_1 \times \Phi_2$. This can be shown as follows:

$$\begin{aligned}\mu_\rho(\alpha_1, \alpha_2) &= \frac{1}{d}\mathrm{Tr}(\rho^{(1)} A_{\alpha_1} \otimes \rho^{(2)} A_{\alpha_2}) \\ &= \frac{1}{d}\sum_{\beta_1 \in \Phi_1, \beta_2 \in \Phi_2} \mu_\rho^{(1)}(\beta_1)\mu_\rho^{(2)}(\beta_2)\mathrm{Tr}(A_{\beta_1} A_{\alpha_1} \otimes A_{\beta_2} A_{\alpha_2}) \\ &= \frac{1}{d}\sum_{\beta_1 \in \Phi_1, \beta_2 \in \Phi_2} \mu_\rho^{(1)}(\beta_1)\mu_\rho^{(2)}(\beta_2) d\delta_{\beta_1\alpha_1}\delta_{\beta_2\alpha_2} \\ &= \mu_\rho^{(1)}(\alpha_1)\mu_\rho^{(2)}(\alpha_2).\end{aligned}$$

Thus, separability can be recast entirely in terms of the discrete phase space. That is, a discrete Wigner function is *separable* if it can be written

$$\mu_\rho(\alpha) = \sum_k p_k \mu_\rho^{(1)}(\alpha_1)_k \mu_\rho^{(2)}(\alpha_2)_k, \tag{B.28}$$

else it is *entangled*.

The two qubit product state $\mu_\rho(\alpha) = \mu_\rho^{(1)}(\alpha_1)\mu_\rho^{(2)}(\alpha_2)$ with $\mu_\rho^{(1)}(\alpha_1) = \frac{1}{2}$ for some α_1 and $\mu_\rho^{(2)}(\alpha_2) = \frac{1-\sqrt{3}}{4}$ for some α_2 will have the most negative value for a separable state, namely $\frac{1-\sqrt{3}}{8}$. Thus, if a two qubit Wigner function has a value strictly less that $\frac{1-\sqrt{3}}{8}$, it is entangled. Since entanglement is considered non-classical, negativity of the Wigner function (below some threshold) is associated with non-classicality. However, even if a Wigner function is positive on all of phase space, it can still be entangled. Therefore, Franco and Penna have found a new sufficient condition for entanglement in two qubits.

For a necessary condition, the authors of [78] turn to the positive partial transpose condition [170, 114]. The result is a two qubit state ρ is separable if and only if both the discrete Wigner function of ρ and the discrete Wigner function of ρ^{T_2} (the partial transpose) are non-negative everywhere on the discrete phase space.

Wootters discrete Wigner function has also found application in quantum teleportation [132]. The authors have found the discrete phase space representation of the teleportation protocol much clearer especially when considering quantum systems with much larger than qubit dimensional Hilbert spaces.

B.2.3 Extended Discrete Phase Space

In reference [53], Cohendet *et al* define a discrete analogue of the Wigner function which is valid for integer spin [7]. That is, $\dim(\mathcal{H}) = d$ is assumed to be odd. Whereas Wootters builds up a discrete phase space before defining a Wigner function, the authors of [53] implicitly define a discrete phase space through the definition of their Wigner function.

Consider the operators

$$W_{mn}\phi_k = \omega^{2n(k-m)}\phi_{k-2m},$$

with $m, n \in \mathbb{Z}_d$ and ϕ_k are the eigenvectors of Z. Then, the *discrete Wigner function* of a density operator ρ is

$$\mu_\rho^{\mathrm{odd}}(q,p) = \frac{1}{d}\mathrm{Tr}(\rho W_{qp}P), \tag{B.29}$$

where P is the parity operator.

The authors call the operators $\triangle_{qp} = W_{qp}P$ *Fano operators* and note that they satisfy

$$\begin{aligned} \triangle_{qp}^\dagger &= \triangle_{qp}, \\ \mathrm{Tr}(\triangle_{qp}\triangle_{q'p'}) &= d\delta_{qq'}\delta_{pp'}, \\ W_{xk}^\dagger\triangle_{qp}W_{xk} &= \triangle_{q-2x\;p-2k}. \end{aligned}$$

The Fano operators play a role similar to Wootters' phase point operators; they form a complete basis of the space of Hermitian operators. The phase space implicitly defined through the definition of the discrete Wigner function (B.29) is $\mathbb{Z}_d \times \mathbb{Z}_d$. When d is an odd prime, this phase space is equivalent to Wootters discrete phase space. In this case the Fano operators are $\triangle_{qp} = A_{(-q,p)}$. This can seen by writing the Wootters phase point operators as

$$A_{(q,p)} = \frac{1}{d}X^{2q}Z^{2p}P\omega^{2qp}.$$

Let $\sigma \in \{\pm 1\}$. The *extended* phase space is $\mathbb{Z}_d \times \mathbb{Z}_d \times \{\pm 1\}$. Define the new Wigner function

$$\mu_\rho(q,p,\sigma) = \frac{1}{4d}\left(\frac{2}{d} + \sigma\mu_\rho^{\mathrm{odd}}(q,p)\right).$$

This function is satisfies the positivity and normalization requirements of a true probability distribution.

[7] This difficulty was overcome in a later paper [52].

Application: Master equation for an integer spin

In the same paper, Cohendet *et al* show the quantum dynamical equation of motion can be represented in the extended phase space as a classical stochastic process. This is achieved by showing the time evolution of the discrete Wigner function is

$$\frac{\partial}{\partial t}\mu_\rho(q,p,\sigma;t) = \sum_{q',p',\sigma'} A(q,p,\sigma|q',p',\sigma')\mu_\rho(q',p',\sigma';t),$$

for a suitable choice of jump moments A. This is in the form of the master equation of a Markov process. The authors interpret this result as follows: "Quantum mechanics of an integer spin appears as the mixture of two classical schemes of a spin. However at random times the schemes are exchanged."

B.2.4 Even Dimensional Discrete Wigner Functions

In reference [138], Leonhardt defines discrete analogues of the Wigner function for both odd and even dimensional Hilbert spaces. In a later paper [139], Leonhardt discusses the need for separate definitions for the odd and even dimension cases. Naively applying his definition, or that of Cohendet *et al*, of the discrete Wigner function for odd dimensions to even dimensions yields unsatisfactory results. The reason for this is the discrete Wigner function carries redundant information for even dimensions which is insufficient to specify the state uniquely. The solution is to enlarge the phase space until the information in the phase space function becomes sufficient to specify the state uniquely.

Suppose $\dim(\mathcal{H}) = d$ is odd. Leonhardt defines the discrete Wigner function as

$$\mu_\rho^{\text{Leo}}(q,p) = \frac{1}{d}\text{Tr}(\rho X^{2q} Z^{2p} P \omega^{2qp}).$$

Leonhardt's definition of an odd dimensional discrete Wigner function is unitarily equivalent to the Cohendet *et al* definition. That is, $\mu_\rho^{\text{Leo}}(q,p) = \mu_\rho^{\text{odd}}(-q,p)$. To define a discrete Wigner function for even dimensions, Leonhardt takes half-integer values of q and p. This amounts to enlarging the phase space to $\mathbb{Z}_{2d} \times \mathbb{Z}_{2d}$. Thus the *even dimensional* discrete Wigner function is

$$\mu_\rho^{\text{even}}(q,p) = \frac{1}{2d}\text{Tr}(\rho X^q Z^p P \omega^{\frac{qp}{2}}),$$

where the operators

$$\triangle_{qp}^{\text{even}} = \frac{1}{2d} X^q Z^p P \omega^{\frac{qp}{2}}$$

could be called the even dimensional Fano or phase point operators. Of course, these operators do not satisfy all the criteria which the Fano operators (in the case of Cohendet *et al*) or the phase point operators (in the case of Wootters) satisfy; they are not orthogonal, for example. Moreover, they are not even linearly independent which can easily be inferred since there are $4d^2$ of them and a set of linearly independent operators contains a maximum of d^2 operators.

Application: quantum computation

Leonhardt's discrete Wigner function has been used to visualize and gain insights for algorithms expected to be performed on a quantum computer [21, 156, 157]. For each step in a quantum algorithm the state $\rho(t)$ of the quantum computer is update via some unitary transformation

$$\rho(t+1) = U\rho(t)U^\dagger.$$

This can be represented in the discrete phase space as

$$\mu_\rho(q,p;t) = \sum_{q'p'} Z(p,q|p',q')\mu_\rho(q',p';t),$$

where Z can be easily obtained from U. This resembles the update map for the probabilities of classical stochastic variables. However, the properties of Z imply that not all admissible maps are classical; they do not connect single points in phase space and hence are "nonlocal". In reference [156] the authors identify a family of classical maps which can be efficiently implemented on a quantum computer. The authors admit that the ultimate usefulness of this approach is uncertain but speculate that the phase space representation may inspire improvement and innovation in quantum algorithms. It certainly makes for some inspiring pictures!

The Leonhardt phase space formalism has also been applied to study decoherence in quantum walks [145]. For large system, numerics are often employed to study the main features. The phase space method offers an intuitive and visual alternative. It allows one to visually see the quantum interference and its disappearance under decoherence. Related to these is a hybrid approach between the Wootters and Leonhardt discrete phase spaces used to analyze various aspects of quantum teleportation [167].

B.2.5 Finite Fields Discrete Phase Space Representation

Recall that when $\dim(\mathcal{H}) = d$ is prime, Wootters defines the discrete phase space as a $d \times d$ lattice indexed by the group $\mathbb{Z}_d$. In reference [226], Wootters generalizes his original construction of a discrete phase space to allow the $d \times d$ lattice to be indexed by a finite field $\mathbb{F}_d$ which exists when $d = p^n$ is an integer power of a prime number. This approach is discussed at length in the paper [86] authored by Gibbons, Hoffman and Wootters (GHW).

Similar to his earlier approach, Wootters defines the *phase space*, Φ_d, as a $d \times d$ array of points $\alpha = (q, p) \in \mathbb{F}_d \times \mathbb{F}_d$. A *line*, λ, is the set of d points satisfying the linear equation $aq + bp = c$, where all arithmetic is done in $\mathbb{F}_d$. Two lines are *parallel* if their linear equations differ in the value of c.

The mathematical structure of $\mathbb{F}_d$ is appealing because lines defined as above have the following useful properties: (i) given any two points, exactly one line contains both points, (ii) given a point α and a line λ not containing α, there is exactly one line parallel to λ that contains α, and (iii) two nonparallel lines intersect at exactly one point. Note that these are usual properties of lines in Euclidean space. As before, the d^2 points of the phase space Φ_d can be partitioned into $d+1$ sets of d parallel lines called *striations*. The line containing the point (q, p) and the origin $(0, 0)$ is called a *ray* and consists of the points (sq, sp), where s is a parameter taking values in $\mathbb{F}_d$. We choose each ray, specified by the equation $aq + bp = 0$, to be the representative of the striation it belongs to.

A translation in phase space, $\mathcal{T}_{\alpha_0}$, adds a constant vector, $\alpha_0 = (q_0, p_0)$, to every phase space point: $\mathcal{T}_{\alpha_0}\alpha = \alpha + \alpha_0$. Each line, λ, in a striation is invariant under a translation by any point contained in its ray, parameterized by the points (sq, sp). That is,

$$\tau_{(sq,sp)}\lambda = \lambda. \tag{B.30}$$

The discrete Wigner function is

$$\mu_\rho^{\text{field}}(q, p) = \frac{1}{d}\text{Tr}(\rho A_{(q,p)}),$$

where now the Hermitian *phase point operators* satisfy the following properties for a projector valued function Q, called a *quantum net*, to be defined later.

GHW(4) For each point α, A is Hermitian.

GHW(5) For any two points α and β, $\text{Tr}(A_\alpha A_\beta) = d\delta_{\alpha\beta}$.

GHW(6) For any line λ, $\displaystyle\sum_{\alpha\in\lambda} A_\alpha = dQ(\lambda)$.

The projector valued function Q assigns quantum states to lines in phase space. This mapping is required to satisfy the special property of *translational covariance*, which is defined after a short, but necessary, mathematical digression. Notice first that properties GHW(4) and GHW(5) are identical to Woo(4) and Woo(5). Also note that if GHW(6) is to be analogous to Woo(6), the property of translation covariance must be such that the set $\{Q(\lambda)\}$ when λ ranges over a striation forms a PVM.

The set of elements $E = \{e_0, ..., e_{n-1}\} \subset \mathbb{F}_d$ is called a *field basis* for $\mathbb{F}_d$ if any element, x, in $\mathbb{F}_d$ can be written

$$x = \sum_{i=0}^{n-1} x_i e_i, \tag{B.31}$$

where each x_i is an element of the prime field $\mathbb{Z}_p$. The *field trace*[8] of any field element is given by

$$\mathrm{tr}(x) = \sum_{i=0}^{n-1} x^{p^i}. \tag{B.32}$$

There exists a unique field basis, $\tilde{E} = \{\tilde{e}_0, ...\tilde{e}_{n-1}\}$, such that $\mathrm{tr}(\tilde{e}_i e_j) = \delta_{ij}$. We call $\tilde{E}$ the *dual* of E.

The construction presented in reference [86] is physically significant for a system of n objects (called *particles*) having a p dimensional Hilbert space. A translation operator, T_α associated with a point in phase space $\alpha = (q, p)$ must act independently on each particle in order to preserve the tensor product structure of the composite system's Hilbert space. We expand each component of the point α into its field basis decomposition as in Equation (B.31)

$$q = \sum_{i=0}^{n-1} q_i e_i \tag{B.33}$$

and

$$p = \sum_{i=0}^{n-1} p_i f \tilde{e}_i, \tag{B.34}$$

with f any element of $\mathbb{F}_d$. Note that the basis we choose for p is a multiple of the dual of that chosen for q. Now, the translation operator associated with the point (q, p) is

$$T_{(q,p)} = \bigotimes_{i=0}^{n-1} X^{q_i} Z^{p_i}, \tag{B.35}$$

Since X and Z are unitary, T_α is unitary.

[8] Note that we will distinguish the field trace, $\mathrm{tr}(\cdot)$, from the usual trace of a Hilbert space operator, $\mathrm{Tr}(\cdot)$, by the case of the first letter.

We assign with each line in phase space a pure quantum state. The quantum net Q is defined such that for each line, λ, $Q(\lambda)$ is the operator which projects onto the pure state associated with λ. As a consequence of the choice of basis for p in Equation (B.34), the state assigned to the line $\tau_\alpha\lambda$ is obtained through

$$Q(\tau_\alpha\lambda) = T_\alpha Q(\lambda)T_\alpha^\dagger. \tag{B.36}$$

This is the condition of translational covariance and it implies that each striation is associated with an orthonormal basis of the Hilbert space. To see this, recall the property in Equation (B.30). From Equation (B.36), this implies that, for each $s \in \mathbb{F}_d$, $T_{(sq,sp)}$ must commute with $Q(\lambda)$, where the line λ is any line in the striation defined by the ray consisting of the points (sq, sp). That is, the states associated to the lines of the striation must be common eigenstates of the unitary translation operator $T_{(sq,sp)}$, for each $s \in \mathbb{F}_d$. Thus, the states are orthogonal and form a basis for the Hilbert space. That is, their projectors form a PVM which makes GHW(6) identical to Woo(6) when d is prime.

In reference [86], the author's note that, although the association between states and vertical and horizontal lines is fixed, the quantum net is not unique. In fact, there are d^{d+1} quantum nets which satisfy Equation (B.30). When d is prime, one of these quantum nets corresponds exactly to the original discrete Wigner function defined by Wootters in Section B.2.2.

Application: quantum computation

As conjectured by Galvão [84], the authors of reference [55] have shown the only quantum states having a non-negative discrete Wigner function [9] are convex combinations of stabilizer states, which are simultaneous eigenstates of the generalized Pauli operators [92]. Working only with stabilizer states is "classical" in the sense that that one can represent them with only a polynomial number of classical bits whereas an arbitrary quantum state requires a exponential number of bits [163].

Strengthening the connection between negativity and non-classicality, it was also shown that the unitary operators preserving the non-negativity of the discrete Wigner function are a subset of the Clifford group, which are those unitaries which preserve Pauli operators under a conjugate mapping. According to the Gottesman-Knill theorem, a quantum computation using only operators from the Clifford group and stabilizer states can be efficiently simulated on a classical computer [92]. Thus, as noted in reference [84] for a particular

[9] Note that it is assumed the discrete Wigner function is non-negative for all definitions - that is, for all quantum nets.

computational model, negativity of the Wigner function is necessary for quantum computational speedup.

This discrete Wigner function was also used to analyze quantum error correcting codes in reference [169]. The aim was to gain insights and intuition for various quantum maps by studying their pictorial representation in the discrete phase space.

B.2.6 Discrete Cahill-Glauber Formalism

In reference [182], Ruzzi *et al* have discretized the Cahill-Glauber phase-space formalism. The set of operators $\{S(\eta, \xi)\}$, where $\eta, \xi \in [-l, l]$ and $l = \frac{d-1}{2}$ (d odd), is called the *Schwinger basis* and explicitly given by

$$S(\eta, \xi) = \frac{1}{\sqrt{d}} X^\eta Z^\xi \omega^{\frac{\eta\xi}{2}}.$$

These d^2 operators form an orthonormal basis for the space of linear operators. In analogy with the Cahill-Glauber formalism, the basis is generalized to

$$S^{(s)}(\eta, \xi) = S(\eta, \xi)\mathcal{K}(\eta, \xi)^{(-s)},$$

where $|s| \leq 1$ is any complex number and $\mathcal{K}(\eta, \xi)$ is a (relatively) complicated expression of Jacobi ϑ-functions (see the Appendix of reference [182]). Next we take the Fourier transform

$$T^{(s)}(q, p) = \frac{1}{\sqrt{d}} \sum_{\eta,\xi=-l}^{l} S^{(s)}(\eta, \xi)\omega^{-(\eta q + \xi p)}.$$

The set operators $\{T^{(s)}(q, p)\}$ is the discrete analog of the s-ordered mapping kernel of the Cahill-Glauber formalism. Moreover, the authors of reference [182] have shown that the continuous limit of this set is indeed the Cahill-Glauber mapping kernel.

Suppose s is real. Then , the operators $\{T^{(s)}(q, p)\}$ enjoy the following familiar properties:

$$T^{(s)}(q, p)^\dagger = T^{(s)}(q, p),$$
$$\mathrm{Tr}(T^{(s)}(q, p)T^{(-s)}(q', p')) = d\delta_{qq'}\delta_{pp'}.$$

Thus, similarly to the discrete kernel of Heiss and Weigert, $\{T^{(s)}(q, p)\}$ and $\{T^{(-s)}(q, p)\}$ are dual bases for the space of Hermitian operators.

In the now familiar way, we can define a quasi-probability function on the (q, p) phase space as

$$\mu_\rho^{(s)}(q, p) = \mathrm{Tr}(T^{(s)}(q, p)\rho). \tag{B.37}$$

Cahill and Glauber showed, for their s-ordered formalism, that $s = 0$ corresponds to the Wigner function; $s = 1$ corresponds to the Husimi function; and $s = -1$ corresponds to the Glauber-Sundarshan function. Using equation (B.37), we call, for example, the function obtained when $s = 0$ the discrete Wigner function.

Application: quantum teleportation

Marchiolli *et al* have applied this formalism to quantum tomography and teleportation [150]. The teleportation protocol was analyzed for arbitrary s but, for brevity, we will consider the $s = 0$ case (which is now assumed so the superscript can be ignored). The teleportation protocol utilizes entanglement to transfer a quantum state between two parties through the exchange of only a small amount of classical information [163]. Consider the tripartite system

$$\rho = \rho^{(1)} \otimes \rho^{(2,3)},$$

where one party possess subsystem 1 and 2 and the other possess subsystem 3. The goal is for $\rho^{(1)}$ to be transferred from subsystem 1 to subsystem 3 without simply swapping them. It is essential that the shared state $\rho^{(2,3)}$ be entangled. In particular, assume it is a maximally entangled pure Bell-state. We choose, following Wootters [225], to construct the global phase space to be a Cartesian product of the phase spaces of the individual subsystems. The discrete Wigner function of the whole system is then

$$\mu_\rho(q_1, q_2, q_3; p_1, p_2, p_3) = \mu_{\rho^{(1)}}(q_1, p_1)\mu_{\rho^{(2,3)}}(q_2, q_3; p_2, p_3).$$

A Bell-measurement is performed on the first two subsystems, which in phase space is interpreted as a measurement of the total momentum and relative coordinate of the subsystem composed of subsystems 1 and 2. Marginalizing over subsystems 1 and 2 gives

$$\mu_{\rho^{(3)}}(q_3, p_3) = \mu_{\rho^{(1)}}(q_3 - \alpha, p_3 + \beta),$$

where α and β parameterize the result of the Bell-measurement (note $\rho^{(3)}$ can be identified as the reduced state of subsystem 3). Thus, the final state of the subsystem 3 is simply a displacement in the phase space and communicating only the measurement result (α, β) leads to recovery of the initial state.

The discrete Husimi function ($s = 1$) was used to define a discrete analog of squeezed states [151] and to analyze spin tunneling effects in a particular toy model of interacting fermions [152].

B.2.7 Probability Tables

In 1986, before introducing the discrete Wigner function, Wootters represented the quantum state as a "probability table" which was simply a list of outcome probabilities for a complete set of measurements [227]. The complete set of interest was that of *mutually unbiased bases* (MUBs). We call n bases $\{\psi_k^n\}$ mutually unbiased if they satisfy

$$\left|\langle\psi_{k'}^{n'},\psi_k^n\rangle\right|^2=\delta_{kk'}\delta_{nn'}+\frac{1}{d}(1-\delta_{nn'}). \tag{B.38}$$

Wootters noted for d prime, a set of $n=d+1$ MUBs could be explicitly constructed via a prescription in reference [120]. Wootters also posed many questions of MUBs, some of which have now been answered. It is now known that for any dimension $3\leq n\leq d+1$, where the upper bound can be achieved, by construction, for any dimension which is a power of a prime [10].

Here we will consider the case when d is prime and all probability tables for non prime dimensions can be built up from those for their prime factors, in much the same was as was done in section B.2.2 for the discrete phase spaces.

Consider the generalized Pauli operator Z and its eigenbasis $\{\phi_k\}$ and the projectors onto these vectors $P_k:=\phi_k\phi_k^*$. Define the finite Fourier transform

$$F=\frac{1}{\sqrt{d}}\sum_{k,k'=0}^{d-1}\omega^{kk'}\phi_k\phi_{k'}^* \tag{B.39}$$

and the operator

$$V=\sum_{k=0}^{d-1}\omega^{\frac{k^2}{2}}FP_kF^\dagger. \tag{B.40}$$

Here, as before, division by two represents the multiplicative inverse of the element 2. For Hilbert space dimension $d=2$, this operator requires the special definition

$$V=\frac{1}{2}\begin{pmatrix}1+i & 1-i\\ 1+i & 1-i\end{pmatrix}. \tag{B.41}$$

Now we can construct $d+1$ MUBs via

$$\begin{aligned}\psi_k^0&=\phi_k,\\ \psi_k^n&=V^n\phi_k,\ \ n=1,\ldots,d.\end{aligned}$$

[10] For a recent review of the MUB problem see [54].

We will denote the projectors onto these basis vectors $P(n,k) := \psi_k^n \psi_k^{n*}$. Then the probability of obtaining the kth outcome when measuring in the nth basis is

$$\mu_\rho(n,k) = \mathrm{Tr}(\rho P(n,k)). \tag{B.42}$$

This can be view as a matrix in which the columns index the measurements while the rows index the outcomes. This can also be viewed as a mapping whose inverse is given by

$$\rho = \sum_{n=0}^{d} \sum_{k=0}^{d-1} \mu_\rho(n,k) P(n,k) - \mathbb{1}. \tag{B.43}$$

Application: quantum mechanics without amplitudes

The purpose of reference [227] was not to introduce a new representation of the quantum state *per se*, but to show that the whole of operational formalism of quantum mechanics can be done rather simply without complex numbers.

Wootters notes first:

> It is obviously possible to devise a formulation of quantum mechanics without probability amplitudes. One is never forced to use any quantities in one's theory other than the raw results of measurements. However, there is no reason to expect such a formulation to be anything other than extremely ugly.

To our surprise, the rule for transitioning between the probability tables turn out to be remarkably simple. In quantum mechanics, for the transition between states ρ to ρ', the probability of this transition is $\Pr(\rho \to \rho') = \mathrm{Tr}(\rho\rho')$. If we work with the probability tables and call the tables μ and μ', we have

$$\Pr(\mu \to \mu') = \sum_{n=0}^{d} \sum_{k=0}^{d-1} \mu(n,k)\mu'(n,k) - 1. \tag{B.44}$$

Unfortunately, as Wootters notes, it is not easy to ignore the density matrix altogether. We have yet to specify which probability tables are valid and which do not correspond to quantum states. The simplest characterization of valid probability tables is to say those for which equation (B.43) is a unit trace positive semi-definite matrix. This is unsatisfying as we would like a characterization independent of the density matrix.

B.2.8 Hardy's Vector Representation

In reference [104] Hardy showed that five axioms are sufficient to imply a special vector representation which is equivalent to an operational form of quantum theory. We first describe the vector representation.

Consider a basis for a d dimensional Hilbert space $\{\phi_k\}$ (the eigenbasis of Z, say) and the following set of d^2 projectors:

$$P_{kj} := \begin{cases} \phi_k\phi_k^* & \text{if } k = j \\ (\phi_k + \phi_j)(\phi_k + \phi_j)^* & \text{if } k < j \\ (\phi_k + i\phi_j)(\phi_k + i\phi_j)^* & \text{if } k > j \end{cases} . \tag{B.45}$$

These projectors span the space of linear operators on the Hilbert space spanned by $\{\phi_k\}$. Now we vectorize by choosing an arbitrary but fixed ordering convention. For definiteness, we choose to stack the rows on top of one another. To this end, define $\alpha := dk + j$ and $P(\alpha) := P_{kj}$. Then, the vector representation of the state ρ is given by

$$\mu_\rho(\alpha) = \mathrm{Tr}(\rho P(\alpha)). \tag{B.46}$$

Now the outcome of any quantum measurement can be assigned a positive operator E. Call this "outcome E". Define the vector $\xi_E(\alpha)$ implicitly through

$$E = \sum_\alpha \xi_E(\alpha) P(\alpha).$$

Then, the probability of "outcome E" is given by

$$\Pr(\text{"outcome } E\text{"}) = \sum_\alpha \xi_E(\alpha)\mu_\rho(\alpha),$$

which, in vector notation, we can write as the dot product $\overrightarrow{\xi} \cdot \overrightarrow{\mu}$.

We define the sets M and Ξ as the set of vectors obtainable through the mappings $\rho \mapsto \mu$ and $E \mapsto \xi$ defined above. More precise statements, in the form of inequalities, which make no recourse to the usual quantum mechanical objects, can be made to define these sets. Assuming this has been done, we can rephrase the axioms of quantum mechanics, without mention of Hermitian operators and the like, in this vector representation succinctly as follows: states are represent by vectors $\overrightarrow{\mu} \in M$; measurement outcomes are represented by vectors $\overrightarrow{\xi} \in \Xi$; the probability of "outcome $\overrightarrow{\xi}$" in state $\overrightarrow{\mu}$ is given by $\Pr(\text{"outcome } \overrightarrow{\xi}\text{"}) = \overrightarrow{\xi} \cdot \overrightarrow{\mu}$.

Application: quantum axiomatics

As was the case in the previous section, this vector representation was not introduced as such. In references [104, 105], Hardy has shown that five axioms are sufficient to imply the real vector formalism of quantum mechanics. The frequency interpretation of probability was given its own axiom. However, if we take our everyday intuitive notion of probability [173], we no longer require this first axiom, which is independent of the rest [184].

We will make use of the following definitions:

- The number of *degrees of freedom*, K, is defined as the minimum number of yes-no measurements whose outcome probabilities are needed to determine the state (of belief in the mind of a reasonable agent), or, more roughly, as the number of real parameters required to specify the state.

- The *dimension*, d, is defined as the maximum number of states that can be reliably distinguished from one another in a single shot measurement.

Axiom 1 defines probability as limiting frequencies and is not required [184]. The remainder of the axioms are as follows:

2 *Simplicity.* K is determined by a function of N (i.e. $K = K(d)$) where $d = 1, 2, \ldots$ and where, for each given d, K takes the minimum value consistent with the axioms.

3 *Subspaces.* A system whose state is constrained to belong to an n dimensional subspace (i.e. have support on only n of a set of d possible distinguishable states) behaves like a system of dimension n.

4 *Composite systems.* A composite system consisting of subsystems A and B satisfies $d = d_A d_B$ and $K = K_A K_B$.

5 *Continuity.* There exists a continuous reversible transformation on a system between any two pure states of that system.

These four axioms are sufficient for a derivation of the vector representation of quantum theory defined above. This axiomatization is also important for contrasting quantum theory with classical probability theory. As Hardy has shown, discrete classical probability theory (of dice, coins and so on) can be derived from only axioms 2, 3 and 4. That is, the only difference between quantum and classical theory is the existence of a *continuous* reversible transformation between pure states.

B.2.9 The Real Density Matrix

In reference [107] Havel defined the "real density matrix" which, not surprisingly, is a particular real-valued representation of the quantum state.

For $d = 2$, define the 2×2 matrix of Pauli operators as

$$P = \begin{pmatrix} \mathbb{1} & X \\ Y & Z \end{pmatrix}. \tag{B.47}$$

For $d = 2^n$, denote the bits in the binary expansion of k as

$$k = \sum_{a=1}^{n} k_a \times 2^{n-a},$$

and similarly for j. Then, the $d \times d$ matrix of Pauli operators is given by

$$P_{kj} = P_{k_1 j_1} \otimes \cdots \otimes P_{k_n j_n}. \tag{B.48}$$

These d^2 operators are orthogonal, and hence form a basis for the space of linear operators on the d dimensional Hilbert space. Therefore, each density matrix can be expressed as

$$\rho = \frac{1}{d} \sum_{k,j=0}^{d-1} \sigma_{kj} P_{kj},$$

where the coefficients σ_{kj}, explicitly given by

$$\sigma_{kj} = \mathrm{Tr}(\rho P_{kj}), \tag{B.49}$$

form the *real density matrix*.

Application: NMR pedagogy

Since the observables measured in NMR experiments are elements of the matrix of Pauli operators (B.48), the elements of the real density matrix are the experimentally measurable values. There is no need to reconstruct the density matrix. This is also a convenient fix to the problem of reporting or visualizing a quantum state. Since the density matrix contains d^2 complex values, it is often graphically displayed as two $d \times d$ matrices of the real and imaginary parts. Not only is this redundant, it is conceptually awkward. On the other hand, the real density matrix can be displayed as a single $d \times d$ matrix of real values. Havel offers the real density matrix as useful teaching device in such situations.

B.2.10 Symmetric Representations

Consider the unitary group

$$U_{(p,q)} = \omega^{\frac{pq}{2}} X^p Z^q, \tag{B.50}$$

where $(p,q) \in \mathbb{Z}_d \times \mathbb{Z}_d$. In references [228, 178], the authors conjecture that the set $\{U_{(p,q)}\phi\}$ for some fiducial $\phi \in \mathcal{H}$ forms a *symmetric informationally complete positive operator valued measure* (SIC-POVM). The defining condition of a SIC-POVM is a set of d^2 vectors $\{\phi_k\}$ such that

$$|\langle \phi_k, \phi_j \rangle|^2 = \frac{\delta_{kj} d + 1}{d+1}. \tag{B.51}$$

The set is called symmetric since the vectors have equal overlap. The POVM is formed by taking the projectors onto the one-dimensional subspaces spanned by the vectors. It is informationally complete since these d^2 projectors span the space of linear operators acting on $\mathcal{H}$.

As of writing, it is an open question whether SIC-POVMs exist in every dimension. Although numerical evidence suggests this to be the case [191].

For the remainder of this section we assume, for any dimension d, a SIC-POVM exists. Define the operators $P_k := \frac{1}{d}\phi_k\phi_k^*$. Then, define the *symmetric-representation* of a quantum state ρ as

$$\mu_\rho(k) = \mathrm{Tr}(\rho P_k). \tag{B.52}$$

This is a probability distribution and in particular it is *the* probability distribution for the POVM measurement formed by the effects $\{P_k\}$. As we have noted, this is an informationally complete measurement. Therefore, the density matrix can be reconstructed from the probabilities via

$$\rho = d(d+1) \sum_{k=0}^{d^2-1} \mu_\rho(k) P_k - \mathbb{1}. \tag{B.53}$$

When viewed as a mapping, this representation is a bijection from the convex set of density matrices to a convex subset of the d^2-dimensional probability simplex.

Application: Quantum Bayesianism

Quantum Bayesianism [185, 40, 80, 82, 81] is an interpretation of quantum theory which sheds new light on not only the tradition "foundational" problems (the "measurement problem", for example) but also many concepts in quantum theory, such as the "unknown quantum state"[41]. A key realization is the mathematical and conceptual sufficiency of viewing quantum states as the probability distribution via the Born rule for a fixed POVM $\{E_k\}$. The only remaining freedom is which one.

One ideal is to have the POVM elements orthogonal: $\mathrm{Tr}(P_k P_j) = \delta_{kj}$. The statement that is not possible is equivalent to theorem 4.2.4. Next, then, we desire them to be as close to orthogonal as possible. Formally, we want to minimize the quantity

$$F = \sum_{kj} (\mathrm{Tr}(P_k P_j) - \delta_{kj})^2.$$

This expression is minimized if and only if the $\{P_k\}$ form a SIC-POVM. Using the reconstruction formula in equation (B.53) it can be shown that, in terms of this SIC-representation, the Born rule for a measurement $\{E_j\}$ given state $\mu(k)$ is

$$\mathrm{Pr}(\text{outcome } j) = \sum_k \left(\mu(k) - \frac{1}{d} \right) \xi(j|k), \tag{B.54}$$

where $\xi(j|k) = \mathrm{Tr}(E_j P_k)$. In the same sense as the SIC-POVM being as close to orthogonal as possible, equation (B.54) is as close as possible to the classical Law of Total Probability. Effort is being made to use equation (B.54) as a starting point for a natural set of axioms which would single out quantum theory.

References

[1] Scott Aaronson and Alex Arkhipov. The Computational Complexity of Linear Optics. November 2010.

[2] Scott Aaronson and Daniel Gottesman. Improved simulation of stabilizer circuits. *Physical Review A*, 70(5):052328+, November 2004.

[3] J. Aczél. A mixed theory of information. V. How to keep the (inset) expert honest. *Journal of Mathematical Analysis and Applications*, 75(2):447–453, June 1980.

[4] J. Aczél. Measuring information beyond communication theory Some probably useful and some almost certainly useless generalizations. *Information Processing & Management*, 20(3):383–395, 1984.

[5] J. Aczél and Z. Daróczy. *On measures of information and their characterization.* Academic Press, 1975.

[6] J. Aczél and J. Pfanzagl. Remarks on the measurement of subjective probability and information. *Metrika*, 11(1):91–105, December 1967.

[7] Robert Alicki, Marco Piani, and Nicholas V. Ryn. Quantumness witnesses. *Journal of Physics A: Mathematical and Theoretical*, 41(49):495303+, 2008.

[8] Robert Alicki and Nicholas V. Ryn. A simple test of quantumness for a single system. *Journal of Physics A: Mathematical and Theoretical*, 41(6):062001+, 2008.

[9] Andris Ambainis, Leonard J. Schulman, and Umesh Vazirani. Computing with highly mixed states. *Journal of the ACM*, 53(3):507–531, March 2006.

[10] Hussain Anwar, Earl T. Campbell, and Dan E. Browne. Qutrit Magic State Distillation. February 2012.

[11] F. T. Arecchi, Eric Courtens, Robert Gilmore, and Harry Thomas. Atomic Coherent States in Quantum Optics. *Physical Review A*, 6(6):2211–2237, December 1972.

[12] E. Bagan, M. A. Ballester, R. D. Gill, Mu noz Tapia, and O. Romero Isart. Separable Measurement Estimation of Density Matrices and its Fidelity Gap with Collective Protocols. *Physical Review Letters*, 97(13):130501+, September 2006.

[13] Stephen D. Bartlett and Barry C. Sanders. Efficient Classical Simulation of Optical Quantum Circuits. October 2002.

[14] Stephen D. Bartlett and Barry C. Sanders. Requirement for quantum computation. *Journal of Modern Optics*, 50(15-17):2331–2340, October 2003.

[15] Stephen D. Bartlett, Barry C. Sanders, Samuel L. Braunstein, and Kae Nemoto. Efficient Classical Simulation of Continuous Variable Quantum Information Processes. *Physical Review Letters*, 88(9):097904+, February 2002.

[16] J. S. Bell. *Speakable and Unspeakable in Quantum Mechanics: Collected Papers on Quantum Philosophy.* Cambridge University Press, June 2004.

[17] E. G. Beltrametti and S. Bugajski. A classical extension of quantum mechanics. *Journal of Physics A: Mathematical and General*, 28(12):3329+, June 1995.

[18] James O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer, 2nd edition, 1985.

[19] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory.* Wiley, Chichester, 1994.

[20] J. Bertrand and P. Bertrand. A tomographic approach to Wigner's function. *Foundations of Physics*, 17(4):397–405, April 1987.

[21] Pablo Bianucci, Cesar Miquela, Juan P. Paz, and Marcos Saraceno. Discrete Wigner functions and the phase space representation of quantum computers. *Physics Letters A*, 297(5-6):353–358, May 2002.

[22] A. Bisio, G. Chiribella, G. M. D'Ariano, S. Facchini, and P. Perinotti. Optimal Quantum Tomography of States, Measurements, and Transformations. *Physical Review Letters*, 102(1):010404+, January 2009.

[23] Robin Blume-Kohout. Hedged maximum likelihood quantum state estimation. *Physical Review Letters*, 105:200504+, November 2010.

[24] Robin Blume-Kohout and Patrick Hayden. Accurate quantum state estimation via "Keeping the experimentalist honest". March 2006.

[25] S. Bochner. Monotone funktionen, Stieltjessche integrate, und harmonischeanalyse. *Math. Ann.*, 108:378–410, 1933.

[26] David Bohm. *Quantum theory.* Dover Publications, May 1989.

[27] N. Boulant, T. F. Havel, M. A. Pravia, and D. G. Cory. Robust method for estimating the Lindblad operators of a dissipative quantum process from measurements of the density operator at multiple time points. *Physical Review A*, 67(4):042322+, April 2003.

[28] Dietrich Braess, Jürgen Forster, Tomas Sauer, and Hans U. Simon. *How to Achieve Minimax Expected Kullback-Leibler Distance from an Unknown Finite Distribution*, volume 2533 of *Lecture Notes in Computer Science*, chapter 30, pages 380–394. Springer Berlin Heidelberg, Berlin, Heidelberg, November 2002.

[29] Samuel L. Braunstein and H. J. Kimble. Teleportation of Continuous Quantum Variables. *Physical Review Letters*, 80(4):869–872, January 1998.

[30] Sergey Bravyi and Alexei Kitaev. Universal quantum computation with ideal Clifford gates and noisy ancillas. *Physical Review A*, 71(2):022316+, February 2005.

[31] C. Brif and A. Mann. Phase-space formulation of quantum mechanics and quantum-state reconstruction for physical systems with Lie-group symmetries. *Physical Review A*, 59(2):971–987, February 1999.

[32] T. Bröcker and R. F. Werner. Mixed states with positive Wigner functions. *Journal of Mathematical Physics*, 36(1):62–75, 1995.

[33] Sławomir Bugajski. Classical frames for a quantum theory A bird's-eye view. *International Journal of Theoretical Physics*, 32(6):969–977, June 1993.

[34] Harry Buhrman, Richard Cleve, Monique Laurent, Noah Linden, Alexander Schrijver, and Falk Unger. New Limits on Fault-Tolerant Quantum Computation. pages 411–419, October 2006.

[35] P. Busch, K. E. Hellwig, and W. Stulpe. On classical representations of finite-dimensional quantum mechanics. *International Journal of Theoretical Physics*, 32(3):399–405, March 1993.

[36] Adán Cabello. Kochen-Specker Theorem for a Single Qubit using Positive Operator-Valued Measures. *Physical Review Letters*, 90(19):190401+, May 2003.

[37] K. E. Cahill and R. J. Glauber. Density Operators and Quasiprobability Distributions. *Physical Review Online Archive (Prola)*, 177(5):1882–1902, January 1969.

[38] Earl T. Campbell and Dan E. Browne. On the Structure of Protocols for Magic State Distillation Theory of Quantum Computation, Communication, and Cryptography. volume 5906 of *Lecture Notes in Computer Science*, chapter 3, pages 20–32. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2009.

[39] Earl T. Campbell and Dan E. Browne. Bound States for Magic State Distillation in Fault-Tolerant Quantum Computation. *Physical Review Letters*, 104:030503+, January 2010.

[40] Carlton M. Caves, Christopher A. Fuchs, and Rüdiger Schack. Quantum probabilities as Bayesian probabilities. *Physical Review A*, 65(2):022305+, January 2002.

[41] Carlton M. Caves, Christopher A. Fuchs, and Rudiger Schack. Unknown quantum states: The quantum de Finetti representation. *Journal of Mathematical Physics*, 43(9):4537+, 2002.

[42] Carlton M. Caves and Krzysztof Wódkiewicz. Classical Phase-Space Descriptions of Continuous-Variable Teleportation. *Physical Review Letters*, 93(4):040506+, July 2004.

[43] Jose L. Cereceda. Local hidden-variable models and negative-probability measures. December 2000.

[44] S. Chaturvedi, E. Ercolessi, G. Marmo, G. Morandi, N. Mukunda, and R. Simon. Wigner distributions for finite dimensional quantum systems: An algebraic approach. *Pramana*, 65(6):981–993, December 2005.

[45] S. Chaturvedi, E. Ercolessi, G. Marmo, G. Morandi, N. Mukunda, and R. Simon. Wigner&ndash;Weyl correspondence in quantum mechanics for continuous and discrete systems&mdash;a Dirac-inspired view. *Journal of Physics A: Mathematical and General*, 39(6):1405–1423, 2006.

[46] Andrew M. Childs, Isaac L. Chuang, and Debbie W. Leung. Realization of quantum process tomography in NMR. *Physical Review A*, 64(1):012314+, June 2001.

[47] Andrew M. Childs, John Preskill, and Joseph Renes. Quantum information and precision measurement. *Journal of Modern Optics*, 47(2):155–176, 2000.

[48] Ole Christensen. *An Introduction to Frames and Riesz Bases.* Birkhäuser, Boston, 2003.

[49] Isaac L. Chuang and M. A. Nielsen. Prescription for experimental determination of the dynamics of a quantum black box. *Journal of Modern Optics*, 44(11):2455–2467, November 1997.

[50] Bertrand S. Clarke and Andrew R. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41(1):37–60, August 1994.

[51] Leon Cohen and Marlan Scully. Joint Wigner distribution for spin-1/2 particles. *Foundations of Physics*, 16(4):295–310, April 1986.

[52] O. Cohendet, P. Combe, and Sirugue M. Collin. Fokker-Planck equation associated with the Wigner function of a quantum system with a finite number of states. *Journal of Physics A: Mathematical and General*, 23(11):2001–2011, 1990.

[53] O. Cohendet, P. Combe, M. Sirugue, and Sirugue M. Collin. A stochastic treatment of the dynamics of an integer spin. *Journal of Physics A: Mathematical and General*, 21(13):2875–2883, 1988.

[54] M. Combescure. The Mutually Unbiased Bases Revisited. May 2006.

[55] Cecilia Cormick, Ernesto F. Galvão, Daniel Gottesman, Juan P. Paz, and Arthur O. Pittenger. Classicality in discrete Wigner functions. *Physical Review A (Atomic, Molecular, and Optical Physics)*, 73(1):012301+, 2006.

[56] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley-Interscience, second edition, July 2006.

[57] G. M. D'Ariano and P. Lo Presti. Quantum Tomography for Measuring Experimentally the Matrix Elements of an Arbitrary Quantum Operation. *Physical Review Letters*, 86(19):4195–4198, May 2001.

[58] Animesh Datta. *Studies on the Role of Entanglement in Mixed-state Quantum Computation.* PhD thesis, University of New Mexico, July 2008.

[59] Animesh Datta, Steven T. Flammia, and Carlton M. Caves. Entanglement and the power of one qubit. *Physical Review A*, 72(4):042316+, October 2005.

[60] Igor Devetak, Aram W. Harrow, and Andreas J. Winter. A Resource Framework for Quantum Shannon Theory. *IEEE Transactions on Information Theory*, 54(10):4587–4618, October 2008.

[61] Matthew J. Donald. On the relative entropy. *Communications in Mathematical Physics*, 105(1):13–34–34, March 1986.

[62] Arnaud Doucet and Adam M. Johansen. *A Tutorial on Particle Filtering and Smoothing: Fifteen Years Later*. Oxford University Press, 2009.

[63] Daniela Dragoman. Applications of the Wigner Distribution Function in Signal Processing. *EURASIP Journal on Applied Signal Processing*, 2005:1520–1534, 2005.

[64] H. Dreier, A. Dinklage, R. Fischer, M. Hirsch, and P. Kornejew. Bayesian experimental design of a multichannel interferometer for Wendelstein 7-X. *Review of Scientific Instruments*, 79(10):10E712+, 2008.

[65] A. Einstein, B. Podolsky, and N. Rosen. Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review Online Archive (Prola)*, 47(10):777–780, May 1935.

[66] Joseph Emerson, Seth Lloyd, David Poulin, and David Cory. Estimation of the local density of states on a quantum computer. *Physical Review A*, 69(5):050305+, May 2004.

[67] Christopher Ferrie. Quasi-probability representations of quantum theory with applications to quantum information science. *Reports on Progress in Physics*, 74(11):116001+, November 2011.

[68] Christopher Ferrie and Robin Blume-Kohout. Estimating the bias of a noisy coin. In *BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS IN SCIENCE AND ENGINEERING: 31st International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, pages 14–21. AIP, January 2012.

[69] Christopher Ferrie and Joseph Emerson. Frame representations of quantum mechanics and the necessity of negativity in quasi-probability representations. *Journal of Physics A: Mathematical and Theoretical*, 41(35):352001+, 2008.

[70] Christopher Ferrie and Joseph Emerson. Framed Hilbert space: hanging the quasi-probability pictures of quantum theory. *New Journal of Physics*, 11(6):063040+, 2009.

[71] Christopher Ferrie, Christopher E. Granade, and D. G. Cory. Adaptive Hamiltonian estimation using Bayesian experimental design. In *BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS IN SCIENCE AND ENGINEERING: 31st International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, pages 165–173. AIP, November 2012.

[72] Christopher Ferrie, Christopher E. Granade, and D. G. Cory. How to best sample a periodic probability distribution, or on the accuracy of Hamiltonian finding strategies. *Quantum Information Processing*, April 2012.

[73] Christopher Ferrie, Ryan Morris, and Joseph Emerson. Necessity of negativity in quantum theory. *Physical Review A*, 82(4):044103+, October 2010.

[74] R. P. Feynman. *Negative Probability*, pages 235–248. Routledge & Kegan Paul Ltd, London & New York, 1987.

[75] Arthur Fine. Some local models for correlation experiments. *Synthese*, 50(2):279–294, February 1982.

[76] P. Fischer. On the inequality p i f (p i) p i f (q i). *Metrika*, 18(1):199–208, December 1972.

[77] Jarom'ir Fiurášek and Zdeněk Hradil. Maximum-likelihood estimation of quantum processes. *Physical Review A*, 63(2):020101+, January 2001.

[78] Riccardo Franco and Vittorio Penna. Discrete Wigner distribution for two qubits: a characterization of entanglement properties. *Journal of Physics A: Mathematical and General*, 39(20):5907–5919, May 2006.

[79] Christopher A. Fuchs. Distinguishability and Accessible Information in Quantum Theory. January 1996.

[80] Christopher A. Fuchs. Quantum Mechanics as Quantum Information (and only a little more). May 2002.

[81] Christopher A. Fuchs. QBism, the Perimeter of Quantum Bayesianism. March 2010.

[82] Christopher A. Fuchs and Ruediger Schack. Quantum-Bayesian Coherence. June 2009.

[83] D. Galetti and A. F. R. de Toledo Piza. An extended Weyl-Wigner transformation for special finite spaces. *Physica A: Statistical and Theoretical Physics*, 149(1-2):267–282, March 1988.

[84] Ernesto F. Galvão. Discrete Wigner functions and quantum computational speedup. *Physical Review A*, 71(4):042302+, April 2005.

[85] Shohini Ghose, Rene Stock, Poul Jessen, Roshan Lal, and Andrew Silberfarb. Chaos, entanglement, and decoherence in the quantum kicked top. *Physical Review A*, 78(4):042318+, October 2008.

[86] Kathleen S. Gibbons, Matthew J. Hoffman, and William K. Wootters. Discrete phase space based on finite fields. *Physical Review A*, 70(6):062101+, December 2004.

[87] Richard D. Gill and Boris Y. Levit. Applications of the van Trees Inequality: A Bayesian Cramér-Rao Bound. *Bernoulli*, 1(1/2), 1995.

[88] Olivier Giraud, Petr Braun, and Daniel Braun. Classicality of spin states. *Physical Review A*, 78(4):042112+, October 2008.

[89] Olivier Giraud, Petr A. Braun, and Daniel Braun. Quantifying Quantumness and the Quest for Queens of Quantum. February 2010.

[90] Roy J. Glauber. Coherent and Incoherent States of the Radiation Field. *Physical Review Online Archive (Prola)*, 131(6):2766–2788, September 1963.

[91] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, 140(2):107–113, April 1993.

[92] Daniel Gottesman. *Stabilizer Codes and Quantum Error Correction.* PhD thesis, California Institute of Technology, May 1997.

[93] Daniel Gottesman, Alexei Kitaev, and John Preskill. Encoding a qubit in an oscillator. *Physical Review A*, 64(1):012310+, June 2001.

[94] Gilad Gour and Robert W. Spekkens. The resource theory of quantum reference frames: manipulations and monotones. *New Journal of Physics*, 10(3):033023+, March 2008.

[95] Philip Goyal. Prior Probabilities: An Information-Theoretic Approach. *AIP Conference Proceedings*, 803(1):366–373, 2005.

[96] D. Gross. Hudson's theorem for finite-dimensional quantum systems. *Journal of Mathematical Physics*, 47(12):122107+, 2006.

[97] D. Gross. Non-negative Wigner functions in prime dimensions. *Applied Physics B*, 86(3):367–370, February 2007.

[98] Andrzej Grudka and Pawel Kurzyński. Is There Contextuality for a Single Qubit? *Physical Review Letters*, 100(16):160401+, April 2008.

[99] Mile Gu, Christian Weedbrook, Nicolas C. Menicucci, Timothy C. Ralph, and Peter van Loock. Quantum computing with continuous-variable clusters. *Physical Review A*, 79:062318+, June 2009.

[100] S. Habib, K. Shizume, and W. H. Zurek. Decoherence, Chaos, and the Correspondence Principle. *Physical Review Letters*, 80:4361–4365, May 1998.

[101] T. Hakioglu. Finite-dimensional Schwinger basis, deformed symmetries, Wigner function, and an algebraic approach to quantum phase. *Journal of Physics A: Mathematical and General*, 31(33):6975–6994, 1998.

[102] Y. Han. Explicit solutions for negative-probability measures for all entangled states. *Physics Letters A*, 221(5):283–286, October 1996.

[103] J. Hannay and M. V. Berry. Quantization of linear maps on a torus-fresnel diffraction by a periodic grating. *Physica D: Nonlinear Phenomena*, 1(3):267–290, September 1980.

[104] Lucien Hardy. Quantum Theory From Five Reasonable Axioms. September 2001.

[105] Lucien Hardy. Why Quantum Theory? November 2001.

[106] Nicholas Harrigan and Terry Rudolph. Ontological models and the interpretation of contextuality. September 2007.

[107] Timothy F. Havel. The Real Density Matrix. *Quantum Information Processing*, 1(6):511–538, December 2002.

[108] Stephan Heiss and Stefan Weigert. Discrete Moyal-type representations for a spin. *Physical Review A*, 63(1):012105+, December 2000.

[109] Carsten Held. The Kochen-Specker Theorem. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2008 edition, 2008.

[110] K. E. Hellwig. Coexistent effects in quantum mechanics. *International Journal of Theoretical Physics*, 2(2):147–155, June 1969.

[111] Fumio Hiai and Dénes Petz. The proper formula for relative entropy and its asymptotics in quantum probability. *Communications in Mathematical Physics*, 143(1):99–114–114, December 1991.

[112] M. Hillery, R. F. O'Connell, M. O. Scully, and E. P. Wigner. Distribution functions in physics: Fundamentals. *Physics Reports*, 106(3):121–167, April 1984.

[113] Bruce Hoadley. Asymptotic Properties of Maximum Likelihood Estimators for the Independent Not Identically Distributed Case. *The Annals of Mathematical Statistics*, 42(6), 1971.

[114] P. Horodecki. Separability criterion and inseparable mixed states with positive partial transposition. *Physics Letters A*, 232(5):333–339, August 1997.

[115] Ryszard Horodecki, Pawel Horodecki, Michal Horodecki, and Karol Horodecki. Quantum entanglement. *Reviews of Modern Physics*, 81(2):865–942, June 2009.

[116] R. Hudson. When is the wigner quasi-probability density non-negative? *Reports on Mathematical Physics*, 6(2):249–252, October 1974.

[117] K. Husimi. Some formal properties of the density matrix. *Proceedings of the Physico-Mathematical Society of Japan*, 22:264+, 1940.

[118] Ferenc Huszár and Neil M. T. Houlsby. Adaptive Bayesian Quantum Tomography. July 2011.

[119] C. Isham. *Lectures on Quantum Theory*. Imperial College Press, 1995.

[120] I. D. Ivonovic. Geometrical description of quantal state determination. *Journal of Physics A: Mathematical and General*, 14(12):3241–3245, December 1981.

[121] Harold Jeffreys. *Theory of probability*. Oxford University Press, second edition, 1939.

[122] E. Joos. *Decoherence through interaction with the environment*, chapter 3, pages 41–180. Springer-Verlag, Berlin, 2 edition, 2003.

[123] Amir Kalev, Ady Mann, Pier A. Mello, and Michael Revzen. Inadequacy of a classical interpretation of quantum projective measurements via Wigner functions. *Physical Review A (Atomic, Molecular, and Optical Physics)*, 79(1):014104+, 2009.

[124] Yuri P. Kalmykov, William T. Coffey, and Serguey V. Titov. Phase space equilibrium distribution function for spins. *Journal of Physics A: Mathematical and Theoretical*, 41(10):105302+, 2008.

[125] Peter J. Kempthorne. Numerical Specification of Discrete Least Favorable Prior Distributions. *SIAM Journal on Scientific and Statistical Computing*, 8(2):171–184, 1987.

[126] Anatole Kenfack and Karol Życzkowski. Negativity of the Wigner function as an indicator of non-classicality. *Journal of Optics B: Quantum and Semiclassical Optics*, 6(10):396+, October 2004.

[127] T. Kiesel, W. Vogel, V. Parigi, A. Zavatta, and M. Bellini. Experimental determination of a nonclassical Glauber-Sudarshan P function. *Physical Review A*, 78(2):021804+, August 2008.

[128] E. Knill and R. Laflamme. Power of One Bit of Quantum Information. *Physical Review Letters*, 81(25):5672–5675, December 1998.

[129] E. Knill and R. Laflamme. Quantum computing and quadratically signed weight enumerators. *Information Processing Letters*, 79(4):173–179, August 2001.

[130] E. Knill, R. Laflamme, and G. J. Milburn. A scheme for efficient quantum computation with linear optics. *Nature*, 409(6816):46–52, January 2001.

[131] S. Kochen and E. Specker. The Problem of Hidden Variables in Quantum Mechanics. *Journal of Mathematics and Mechanics*, 17:59–87, 1967.

[132] M. Koniorczyk, V. Bužek, and J. Janszky. Wigner-function description of quantum teleportation in arbitrary dimensions and a continuous limit. *Physical Review A*, 64(3):034301+, August 2001.

[133] R. E. Krichevskiy. Laplace's law of succession and universal encoding. *IEEE Transactions on Information Theory*, 44(1):296–303, January 1998.

[134] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.

[135] Raymond Laflamme, David G. Cory, Camille Negrevergne, and Lorenza Viola. NMR Quantum Information Processing and Entanglement. *Quantum Information and Computation*, 2:166+, October 2002.

[136] H. Lee. Theory and application of the quantum phase-space distribution functions. *Physics Reports*, 259(3):147–211, August 1995.

[137] E. L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, 2nd edition, September 1998.

[138] Ulf Leonhardt. Quantum-State Tomography and Discrete Wigner Function. *Physical Review Letters*, 74(21):4101–4105, May 1995.

[139] Ulf Leonhardt. Discrete Wigner function and quantum-state tomography. *Physical Review A*, 53(5):2998–3013, May 1996.

[140] Debbie W. Leung. Choi's proof as a recipe for quantum process tomography. *Journal of Mathematical Physics*, 44(2):528–533, 2003.

[141] D. V. Lindley. On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, December 1956.

[142] J. Liu and M. West. *Combined parameter and state estimation in simulation-based filtering*. Springer-Verlag, 2000.

[143] Etera R. Livine. Notes on qubit phase space and discrete symplectic structures. *Journal of Physics A: Mathematical and Theoretical*, 43(7):075303+, February 2010.

[144] Seth Lloyd and Samuel L. Braunstein. Quantum Computation over Continuous Variables. *Physical Review Letters*, 82(8):1784–1787, February 1999.

[145] Cecilia C. López and Juan P. Paz. Phase-space approach to the study of decoherence in quantum walks. *Physical Review A*, 68(5):052305+, November 2003.

[146] Thomas J. Loredo. Bayesian Adaptive Exploration. *AIP Conference Proceedings*, 707(1):330–346, 2004.

[147] A. Luis and J. Perina. Discrete Wigner function for finite-dimensional systems. *Journal of Physics A: Mathematical and General*, 31(5):1423–1441, 1998.

[148] Mark W. Maciejewski, Harry Z. Qui, Iulian Rujan, Mehdi Mobli, and Jeffrey C. Hoch. Nonuniform sampling and spectral aliasing. *Journal of Magnetic Resonance*, 199(1):88–93, July 2009.

[149] L. Mandel. Non-Classical States of the Electromagnetic Field. *Physica Scripta*, 1986(T12):34+, January 1986.

[150] Marcelo A. Marchiolli, Maurizio Ruzzi, and Diógenes Galetti. Extended Cahill-Glauber formalism for finite-dimensional spaces. II. Applications in quantum tomography and quantum teleportation. *Physical Review A*, 72(4):042308+, October 2005.

[151] Marcelo A. Marchiolli, Maurizio Ruzzi, and Diógenes Galetti. Discrete squeezed states for finite-dimensional spaces. *Physical Review A (Atomic, Molecular, and Optical Physics)*, 76(3):032102+, 2007.

[152] Marcelo A. Marchiolli, Evandro C. Silva, and Diógenes Galetti. Quasiprobability distribution functions for finite-dimensional discrete phase spaces: Spin-tunneling effects in a toy model. *Physical Review A (Atomic, Molecular, and Optical Physics)*, 79(2):022114+, 2009.

[153] A. Mari, K. Kieling, B. Melholt Nielsen, E. S. Polzik, and J. Eisert. Directly estimating non-classicality. May 2010.

[154] Nicolas C. Menicucci, Peter van Loock, Mile Gu, Christian Weedbrook, Timothy C. Ralph, and Michael A. Nielsen. Universal Quantum Computation with Continuous-Variable Cluster States. *Physical Review Letters*, 97:110501+, September 2006.

[155] Yann Merlé and France Mentré. Stochastic optimization algorithms of a Bayesian design criterion for Bayesian parameter estimation of nonlinear regression models: Application in pharmacokinetics. *Mathematical Biosciences*, 144(1):45–70, August 1997.

[156] César Miquel, Juan P. Paz, and Marcos Saraceno. Quantum computers in phase space. *Physical Review A*, 65(6):062309+, June 2002.

[157] Cesar Miquel, Juan P. Paz, Marcos Saraceno, Emanuel Knill, Raymond Laflamme, and Camille Negrevergne. Interpretation of tomography and spectroscopy as dual forms of quantum computation. *Nature*, 418(6893):59–62, July 2002.

[158] M. W. Mitchell, C. W. Ellenor, S. Schneider, and A. M. Steinberg. Diagnosis, Prescription, and Prognosis of a Bell-State Filter by Quantum Process Tomography. *Physical Review Letters*, 91(12):120402+, September 2003.

[159] Ryan Morris. Topics in Quantum Foundations: Ontological Models, and Distinguishability as a Resource. Master's thesis, University of Waterloo, 2009.

[160] J. E. Moyal. Quantum mechanics as a statistical theory. *Mathematical Proceedings of the Cambridge Philosophical Society*, 45(01):99–124, 1949.

[161] N. Mukunda. Wigner distributions for non-Abelian finite groups of odd order. *Physics Letters A*, 321(3):160–166, February 2004.

[162] M. A. Nielsen, E. Knill, and R. Laflamme. Complete quantum teleportation using nuclear magnetic resonance. *Nature*, 396(6706):52–55, November 1998.

[163] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information.* Cambridge University Press, 1 edition, October 2004.

[164] Daniel Oi and Sophie Schirmer. Quantum system characterization with limited resources. February 2012.

[165] T. Opatrný, D. G. Welsch, and V. Bužek. Parametrized discrete phase-space functions. *Physical Review A*, 53(6):3822–3835, June 1996.

[166] Matteo Paris and Jaroslav Rehacek, editors. *Quantum State Estimation*, volume 649 of *Lecture Notes in Physics*. Springer, 2004.

[167] Juan P. Paz. Discrete Wigner functions and the phase-space representation of quantum teleportation. *Physical Review A*, 65(6):062311+, June 2002.

[168] Juan P. Paz, Salman Habib, and Wojciech H. Zurek. Reduction of the wave packet: Preferred observable and decoherence time scale. *Physical Review D*, 47(2):488–501, January 1993.

[169] Juan P. Paz, Augusto J. Roncaglia, and Marcos Saraceno. Qubits in phase space: Wigner-function approach to quantum-error correction and the mean-king problem. *Physical Review A*, 72(1):012309+, July 2005.

[170] Asher Peres. Separability Criterion for Density Matrices. *Physical Review Letters*, 77(8):1413–1415, August 1996.

[171] P. G. L. Porta Mana. Probability tables. March 2004.

[172] Piero G. L. Porta Mana. Why can states and measurement outcomes be represented as vectors? September 2004.

[173] Piero G. L. Porta Mana. *Studies in plausibility theory, with applications to physics.* PhD thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2007.

[174] David Poulin, Robin B. Kohout, Raymond Laflamme, and Harold Ollivier. Exponential Speedup with a Single Bit of Quantum Information: Measuring the Average Fidelity Decay. *Physical Review Letters*, 92(17):177906+, April 2004.

[175] J. F. Poyatos, J. I. Cirac, and P. Zoller. Complete Characterization of a Quantum Process: The Two-Bit Quantum Gate. *Physical Review Letters*, 78(2):390–393, January 1997.

[176] Michael Reck, Anton Zeilinger, Herbert J. Bernstein, and Philip Bertani. Experimental realization of any discrete unitary operator. *Physical Review Letters*, 73(1):58–61, July 1994.

[177] Michael Redhead. *Incompleteness, Nonlocality, and Realism.* Oxford University Press, 1989.

[178] Joseph M. Renes, Robin B. Kohout, A. J. Scott, and Carlton M. Caves. Symmetric informationally complete quantum measurements. *Journal of Mathematical Physics*, 45(6):2171–2180, 2004.

[179] A. Rivas. The Weyl Representation on the Torus. *Annals of Physics*, 276(2):223–256, September 1999.

[180] Tony Rothman and E. C. G. Sudarshan. Hidden Variables or Positive Probabilities? *International Journal of Theoretical Physics*, 40(8):1525–1543, August 2001.

[181] A. L. Rukhin. Minimax Estimation of the Binomial Parameter Under Entropy Loss. *Statistics and Decisions*, Supplement Issue No. 3:69–81, 1993.

[182] M. Ruzzi, M. A. Marchiolli, and D. Galetti. Extended Cahill-Glauber formalism for finite-dimensional spaces: I. Fundamentals. *Journal of Physics A: Mathematical and General*, 38(27):6239–6251, July 2005.

[183] Massimiliano F. Sacchi. Maximum-likelihood reconstruction of completely positive maps. *Physical Review A*, 63(5):054104+, April 2001.

[184] Rüdiger Schack. Quantum Theory from Four of Hardy's Axioms. *Foundations of Physics*, 33(10):1461–1468, October 2003.

[185] Rüdiger Schack, Todd A. Brun, and Carlton M. Caves. Quantum Bayes rule. *Physical Review A*, 64(1):014305+, June 2001.

[186] Rüdiger Schack and Carlton M. Caves. Classical model for bulk-ensemble NMR quantum computation. *Physical Review A*, 60(6):4354–4362, December 1999.

[187] Rüdiger Schack and Carlton M. Caves. Explicit product ensembles for separable quantum states. *Journal of Modern Optics*, 47(2):387–399, 2000.

[188] Wolfgang P. Schleich. *Quantum Optics in Phase Space*. Wiley-VCH, 1 edition, February 2001.

[189] Benjamin Schumacher and Michael D. Westmoreland. Relative entropy in quantum information theory. In Samuel J. Lomonaco and Howard E. Brandt, editors, *Quantum Computation and Information: Ams Special Session Quantum Computation and Information*, volume 305, pages 265–290. AMS Bookstore, January 2000.

[190] A. J. Scott. Tight informationally complete quantum measurements. *Journal of Physics A: Mathematical and General*, 39(43):13507–13530, October 2006.

[191] A. J. Scott and M. Grassl. Symmetric informationally complete positive-operator-valued measures: A new computer study. *Journal of Mathematical Physics*, 51(4):042203+, 2010.

[192] M. Scully and K. Wódkiewicz. Spin quasi-distribution functions. *Foundations of Physics*, 24(1):85–107, January 1994.

[193] Alexandr Sergeevich, Anushya Chandran, Joshua Combes, Stephen D. Bartlett, and Howard M. Wiseman. Characterization of a qubit Hamiltonian using adaptive measurements in a fixed basis. February 2011.

[194] L. K. Shalm, R. B. A. Adamson, and A. M. Steinberg. Squeezing and over-squeezing of triphotons. *Nature*, 457(7225):67–70, January 2009.

[195] Claude E. Shannon and Warren Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, September 1998.

[196] Dan Shepherd. Computation with Unitaries and One Pure Qubit. November 2006.

[197] Peter W. Shor and Stephen P. Jordan. Estimating Jones polynomials is a complete problem for one clean qubit. *Quantum Information and Computation*, 8:681+, February 2008.

[198] Francisco SotoEguibar and Pierre Claverie. When is the Wigner function of multidimensional systems nonnegative? *Journal of Mathematical Physics*, 24(1):97–100, 1983.

[199] R. W. Spekkens. Contextuality for preparations, transformations, and unsharp measurements. *Physical Review A*, 71(5):052108+, May 2005.

[200] Robert W. Spekkens. Negativity and Contextuality are Equivalent Notions of Nonclassicality. *Physical Review Letters*, 101(2):020401+, 2008.

[201] Robert W. Spekkens, D. H. Buzacott, A. J. Keehn, Ben Toner, and G. J. Pryde. Preparation Contextuality Powers Parity-Oblivious Multiplexing. *Physical Review Letters*, 102(1):010401+, January 2009.

[202] J. Sperling and W. Vogel. Representation of entanglement by negative quasiprobabilities. *Physical Review A (Atomic, Molecular, and Optical Physics)*, 79(4):042337+, 2009.

[203] M. D. Srinivas and E. Wolf. Some nonclassical features of phase-space representations of quantum mechanics. *Physical Review D*, 11(6):1477+, March 1975.

[204] R. L. Stratonovich. On distributions in representation space. *J. Exp. Theor. Phys.*, 4:891+, 1957.

[205] Werner Stulpe. *Classical Representations of Quantum Mechanics Related to Statistically Complete Observables.* Wissenschaft und Technik Verlag, Berlin, October 1997.

[206] E. C. G. Sudarshan. Equivalence of Semiclassical and Quantum Mechanical Descriptions of Statistical Light Beams. *Physical Review Letters*, 10(7):277–279, April 1963.

[207] Kin'ya Takahashi. Distribution Functions in Classical and Quantum Mechanics. *rogress of Theoretical Physics Supplement*, 98:109–156, 1989.

[208] Fuyuhiko Tanaka and Fumiyasu Komaki. Bayesian predictive density operators for exchangeable quantum-statistical models. *Physical Review A*, 71(5):052323+, May 2005.

[209] P. Tichavsky, C. H. Muravchik, and A. Nehorai. Posterior Cramer-Rao bounds for discrete-time nonlinear filtering. *Signal Processing, IEEE Transactions on*, 46(5):1386–1396, May 1998.

[210] Lev Vaidman. Teleportation of quantum states. *Physical Review A*, 49(2):1473–1476, February 1994.

[211] Wim van Dam and Mark Howard. Noise thresholds for higher-dimensional systems using the discrete Wigner function. *Physical Review A*, 83:032310+, March 2011.

[212] J. Varilly and J. Graciabondia. The moyal representation for spin. *Annals of Physics*, 190(1):107–148, February 1989.

[213] V. Vedral. The role of relative entropy in quantum information theory. *Reviews of Modern Physics*, 74(1):197–234, March 2002.

[214] V. Vedral. The Elusive Source of Quantum Speedup. *Foundations of Physics*, 40:1141–1154, 2010.

[215] Victor Veitch, Christopher Ferrie, and Joseph Emerson. Negative Quasi-Probability Representation is a Necessary Resource for Magic State Distillation. January 2012.

[216] Guifré Vidal. Efficient Classical Simulation of Slightly Entangled Quantum Computations. *Physical Review Letters*, 91:147902+, October 2003.

[217] U. von Toussaint, T. Schwarz-Selinger, M. Mayer, and S. Gori. Optimizing NRA depth profiling using Bayesian experimental design. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 268(11-12):2115–2118, June 2010.

[218] A. Vourdas. Quantum systems with finite Hilbert space. *Reports on Progress in Physics*, 67(3):267–320, 2004.

[219] Joel J. Wallman and Stephen D. Bartlett. Nonnegative subtheories and quasiprobability representations of qubits. March 2012.

[220] D. F. Walls and G. J. Milburn. *Quantum Optics*. Springer, February 1995.

[221] Christian Weedbrook, Stefano Pirandola, Raul Garcia-Patron, Nicolas J. Cerf, Timothy C. Ralph, Jeffrey H. Shapiro, and Seth Lloyd. Gaussian Quantum Information. October 2011.

[222] Yaakov S. Weinstein, Timothy F. Havel, Joseph Emerson, Nicolas Boulant, Marcos Saraceno, Seth Lloyd, and David G. Cory. Quantum process tomography of the quantum Fourier transform. *The Journal of Chemical Physics*, 121(13):6117–6133, 2004.

[223] Ruby C. Weng. A Bayesian Edgeworth expansion by Stein's Identity. *Bayesian Analysis*, 5:741–764, 2010.

[224] E. Wigner. On the Quantum Correction For Thermodynamic Equilibrium. *Physical Review Online Archive (Prola)*, 40(5):749–759, June 1932.

[225] W. Wootters. A Wigner-function formulation of finite-state quantum mechanics. *Annals of Physics*, 176(1):1–21, May 1987.

[226] W. K. Wootters. Picturing qubits in phase space. *IBM Journal of Research and Development*, 48(1):99+, 2004.

[227] William Wootters. Quantum mechanics without probability amplitudes. *Foundations of Physics*, 16(4):391–405, April 1986.

[228] Gerhard Zauner. Quantum designs: Foundations of a noncommutative design theory. *International Journal of Quantum Information*, 09(01):445+, 2011.

[229] D. Zueco and I. Calvo. Bopp operators and phase-space spin dynamics: application to rotational quantum Brownian motion. *Journal of Physics A: Mathematical and Theoretical*, 40(17):4635–4648, 2007.