Foundational aspects of nonlocality

Thesis by

Greg Ver Steeg

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

2009
(Defended May 26, 2009)

Copyright notice and joint work:

Chapter 4 contains material from [1] which is joint work with Nicolas Menicucci and is copyrighted by the American Physical Society. Chapter 2 contains material from [2], which is joint work with Stephanie Wehner and has been accepted for publication in Quantum Information and Computation, copyright Rinton Press.

Acknowledgements

I am grateful to my advisor, John Preskill, for being an excellent guide into the unknown. I thank Alexei Kitaev and Sean Carrol for being on my candidacy and thesis committees, Yangbei Chen for being on my candidacy committee, and Yaser Abu-Mostafa for being on my thesis committee and indulging my questions about statistical learning theory.

I greatly appreciate the excellent people I collaborated with during my time at Caltech. Thanks to Stephanie Wehner, Nick Menicucci, and Chris Adami for thinking about outlandish topics with me. For administrative support I am indebted to Ann Harvey especially, and also Donna Driscoll and Carol Silberstein.

Looking back to my undergraduate days, I couldn't have gotten here without my research advisor, Klaus Bartschat, for giving me fantastic research opportunities, Athan Petridis, for deep insights about physics and being a physicist, and my philosophy advisor, Allen Scult, for insights about everything else.

All my friends at Caltech have been fantastic, I especially want to thank my former roommates Michael Adams and Paul Cook.

Most importantly I thank my parents Galen and Jane and my sisters Amy and Alissa for being unconditionally supportive in spite of my eccentricities. Finally, Farah, I dedicate the end of this stage of my life to our beginning. May our correlations remain inexplicably strong.

Abstract

Nonlocality refers to correlations between spatially separated parties that are stronger than those explained by the existence of local hidden variables. Quantum mechanics is known to allow some nonlocal correlations between particles in a phenomena known as entanglement. We explore several aspects of nonlocality in general and how they relate to quantum mechanics.

First, we construct a hierarchy of theories with nonlocal correlations stronger than those allowed in quantum mechanics and derive several results about these theories. We show that these theories include codes that can store an amount of information exponential in the number of physical bits used. We use this result to demonstrate an unphysical consequence of theories with stronger-than-quantum correlations: learning even an approximate description of states in such theories would be practically impossible.

Next, we consider the difficult problem of determining whether specific correlations are nonlocal. We present a novel learning algorithm and show that it provides an outer bound on the set of local states, and can therefore be used to identify some nonlocal states.

Finally, we put nonlocal correlations to work by showing that the entanglement present in the vacuum of a quantum field can be used to detect spacetime curvature. We quantify how the entangling power of the quantum field varies as a function of spacetime curvature.

Contents

Chapter 1

Introduction

Physis kryptesthai philei.
"Nature loves to hide."
Heraclitus

1.1 Motivation

Even after a hundred years of progress, new issues regarding the foundations of quantum physics continue to arise. Many of these issues could not be appreciated without the framework provided by recent advances in information science. Before the advent of computers, the question of the information processing power of a physical theory held limited ramifications.

Why do we find quantum mechanics and not another theory? This may seem like a metaphysical question, but many independent inquiries suggest that even slight modifications to quantum mechanics lead to physically unacceptable consequences. What qualifies as unacceptable is a free lunch, whether it is in the form of computing[3], energy[4], or communication[5]. This suggests that the most important content of a physical theory has not to do with Hilbert spaces, but information properties that are neither trivial nor too powerful.

One of the most shocking examples of physics that is nontrivial while remaining just shy of a free lunch has come from the study of nonlocal correlations. We will focus on nonlocality as we attempt to address some foundational questions.

For instance, we know quantum mechanics is nonlocal, but we also know it is not as nonlocal as it could be [6]. Is there a deep reason this should be the case? This question is closely related to an approach which attempts to reformulate quantum mechanics in terms of a small number of information-based axioms, rather than abstract mathematical ones [7]. A collection of physically motivated, informational axioms might also suggest solutions to the vexing problem of combining gravity and quantum mechanics. Indeed, predictions on the fate of information in quantum mechanics versus general relativity are the source of much of the controversy on how to proceed[8].

In the work that follows we consider these themes in various guises. What nonlocal correlations are possible, and what are their information processing properties? How can we determine if correlations are nonlocal? How is nonlocality affected by the curvature of spacetime?

1.2 Operational backdrop

In the early days of quantum mechanics, counterfactual questions were a rich source of philosophic strife surrounding quantum mechanics: "What *would* the result have been if we had made a different measurement?" Our physical intuition tells us there must be an answer. A thing must be one way or another, whether we look at it or not. But in this case, nature hides nothing, unless it is that there is nothing to hide. If we give up our idea of truly objective measurements that have no effect on the measured system, then how can we construct meaningful experiments?

Now we define a general experimental setup which will stay with us throughout this thesis.

Definition 1.2.1. *The following general experimental setup we call a* measurement scenario*, the result of which is to generate a* probability distribution *over the outcomes and measurement choices. We have n parties, each of whom has a choice of m measurements choices, which may result in one of k outcomes. For the i-th party we refer to their measurement choice as $X_i \in \mathbb{Z}^m$ and the result of their measurement as $A_i \in \mathbb{Z}^k$. We assume that each measurement choice is made independently and that the measuring events for each party are spacelike separated. After repeating this experiment many times, the experimenters generate a probability distribution $p(A_1, \ldots, A_n|X_1, \ldots, X_n)$ fully describing the results of this experimental setup.*

This is a quite general experimental setup, but there are some ingredients to it that would have been unthinkable a hundred years ago.

- Measurement choice: The idea that some measurements are simply incompatible is a fundamentally quantum mechanical idea. Because making a measurement necessarily disturbs the results of subsequent measurements, we must choose which properties we would like to measure accurately in any particular experiment.

- Choice independence: Normally, it would not make a difference how our measurements are chosen, but if something or someone is behind the scenes manipulating which measurements our "independent" experimenters are making, this would allow the simulation of nonclassical correlations with only classical resources. Although this remains a possible explanation for nonlocal correlations, the next ingredient makes this highly unlikely.

- Spacelike separation: This is not a strictly necessary ingredient, but it does guarantee, at least insofar as we trust the laws of special relativity, no information about one measurement choice can propagate to one of the other experimenters to cause some effect. As an added bonus, we need not concern ourselves about the timing of the different measurements; according to special relativity the order in which the measurements are made is totally relative.

- Probabilistic description: In Laplace's day, it seemed that randomness in our measurements could be banished by examining a system in finer detail, until all parts were understood and everything behaved deterministically. Quantum mechanics tells us that some randomness is intrinsic. In these cases, a probabilistic description is our only hope.

1.3 Definition of concepts

In the context of the measurement scenario just described, we now define precisely what is meant by the concepts used heavily throughout this work: *local*, *nonlocal*, and *no-signalling.*

Definition 1.3.1. *A probability distribution for measurement scenario 1.2.1 is* local *if it can be described by a local hidden variable model. That is, their exists a (possibly infinite) hidden variable $R \in \mathbb{Z}^N$ and probability distributions $p(A_i|X_i, R), p(R)$, so that*

$$p(A_1, \ldots, A_n|X_1, \ldots, X_n) = \sum_R p(R) \prod_{i=1}^{n} p(A_i|X_i, R).$$

In theory, the hidden variable could be continuous, though this makes no difference if we only consider experimental setups with finite n, m, k. For a continuous formulation, see [9]. The content of this definition is just that any measurement outcome should depend only on the measurement choice and some (classical) information shared by all the parties.

Definition 1.3.2. *A probability distribution for measurement scenario 1.2.1 is* nonlocal *if it can not be described by a local hidden variable model, as defined by 1.3.1.*

Surprisingly, we shall see that there are situations in which a measurement result does depend on the measurement choice and outcome of another party, but not in a way that can be used to send a signal.

Definition 1.3.3. *A probability distribution for measurement scenario 1.2.1 is* no-signaling *if it satisfies the following equality constraints. For any partitioning of the parties $\{1, \ldots, n\}$ into $\{i_1, \ldots, i_s\}$ and $\{j_1, \ldots, j_t\}$ with $s + t = n$,*

$$\begin{gathered}\sum_{A_{i_1},\ldots,A_{i_s}} p(A_1, \ldots, A_n|X_{i_1}, \ldots, X_{i_s}, X_{j_1}, \ldots, X_{j_t}) \\ = \sum_{A_{i_1},\ldots,A_{i_s}} p(A_1, \ldots, A_n|X'_{i_1}, \ldots, X'_{i_s}, X_{j_1}, \ldots, X_{j_t}) \\ \forall A_{j_1}, \ldots, A_{j_t}, X_{i_1}, \ldots, X_{i_s}, X'_{i_1}, \ldots, X'_{i_s}, X_{j_1}, \ldots, X_{j_t}\end{gathered}$$

Intuitively, this just says that the probability of finding any measurement result for one party is independent of the measurement choices of another party *as long as we average over their possible results.* Special relativity has been so extensively confirmed, that it would be surprising indeed if any physical theory failed to satisfy this requirement.

1.4 Bell inequalities

How can we identify nonlocal correlations in a probability distribution? For now we will simply point out the existence of linear inequalities called Bell inequalities which must be satisfied for any local probability distribution [10]. We will discuss the conditions distinguishing local and nonlocal distributions in more detail in Chapter 3, along with a more geometric interpretation of Bell inequalities, see 3.1.

In this work, we will generally speak of Bell inequalities as any inequality which, when violated, implies nonlocal correlations. This is in contrast to what we will refer to as *tight* Bell inequalities, some collection of inequalities which are satisfied if and only if a probability distribution is local.

1.4.1 CHSH inequality

The simplest situation one can imagine in which nonlocality arises is with only two parties, each making one of two measurements, with two possible outcomes. In this case, up to the symmetry of relabeling the parties, the outcomes, or the measurements, there is only one (tight) Bell inequality, referred to as the CHSH inequality[11]. If the outcomes A, B and measurement settings X, Y all take values in $\{0, 1\}$, the inequality can be expressed very simply using $\delta_{a,b}$, equal to one if $a = b$ or zero otherwise.

$$\eta_L = \frac{1}{4} \sum_{A,B,X,Y \in \{0,1\}} p(AB|XY)\ \delta_{A\oplus B, X\cdot Y} \leq \frac{3}{4} \tag{1.1}$$

For a no-signalling distribution, the maximum achievable $\eta_{NS} = 1$ [12], while for quantum mechanics, the maximum is $\eta_Q = \frac{1}{2} + \frac{\sqrt{2}}{2}$[13]. This is the origin of the claim that quantum mechanics is nonlocal, but not as nonlocal as it could be without violating special relativity.

1.5 Nonlocality versus entanglement

Although *entangled* states in quantum mechanics are generally considered to be nonlocal objects, in the sense that two spatially disjoint subsystems must be regarded as parts of one larger system, this is subtly different than nonlocality as we have defined it.

Consider the density matrix ρ_{AB} for a quantum system that exists in a tensor product Hilbert space $\mathcal{H}^A \otimes \mathcal{H}^B$. This state could represent, for instance, two spacelike separated particles. The quantum systems A and B are considered to be *separable* if the density matrix describing ρ_{AB} can be written,

$$\rho_{AB} = \sum_i \lambda_i \rho_A^{(i)} \otimes \rho_B^{(i)},$$

that is, as a mixture of tensor products states. If two systems are not separable they are considered *entangled* [14].

Although it is easy to see that any measurements on a separable state must lead to a local probability distribution, it is less clear that an entangled one is necessarily nonlocal. That is, are there always a set of measurement choices on an entangled quantum state that produce a probability distribution violating some Bell inequality [15]? Surprisingly, for a class of states called Werner states, the answer is no[16]. On the other hand, even in cases where a Bell inequality is not violated by some entangled state, it may be the case that several copies of such an entangled state can be used to produce a violation. Such states are said to exhibit hidden nonlocality[17]. In this thesis we will generally be more concerned with nonlocality than entanglement, though in Chapter 4 we will make use of results for amplifying small amounts of entanglement to produce Bell inequality violations.

1.6 Overview of the thesis

The groundwork for the subsequent three chapters is mostly independent, though all relate to the theme of nonlocality in some way. Here we discuss the main results of each chapter with a particular emphasis on the individual contributions of the author and their connection to the motivation of this thesis.

1.6.1 Information processing in nonlocal theories

The first chapter considers nonlocal correlations in a general way without reference to any particular physical theory. This allows us to consider the information processing capabilities that nonlocality allows.

There are two main ideas in this chapter. The first concerns the structure of nonlocal theories and is mostly discussed in Sections 2.2, 2.3, and 2.5. Here we argue that the typical paradigm for generalized no-signaling theories (we refer to these as GNST) has been artificially restricted to include only choices among local measurements[18, 7, 6]. We propose a generalization, which we refer to as p-nonlocal theories, that allows simultaneous measurement for any commuting observables. We show that this structure implies additional restrictions on the space of allowed states in any such theory. The differences in information processing power of these two types of theories is discussed in the second half.

The second main idea concerns the information properties of nonlocal theories and is presented in Sections 2.4, 2.5, 2.6, and 2.7. Despite quantum physics' computing strengths, and the seemingly vast size of Hilbert space, quantum states are only marginally better at a task known as random access coding[19]. That is, if you want to store many bits so that a small number can be recovered with high probability, quantum physics performs only marginally better than classical physics. Aaronson turned this "lemon" result into "lemonade" by pointing out that this also insures that quantum states are "learnable"[20]. Suppose you had an unknown quantum state of a mere 100 qubits. This state is described by 2^{100} complex amplitudes which would require $O(2^{100})$ separate experiments to determine. We can never hope to do so many experiments to confirm the identity of our state, even in the lifetime of the universe. What Aaronson's learnability result implies is that after performing a number of experiments linear in the number of qubits, we can create an approximate description of the quantum state that will accurately describe the results of most future measurements with high probability.

We consider the extension of this result to any theory that includes more general nonlocal correlations. First we demonstrate that these theories must include powerful random access codes, capable of storing an exponential amount of information of which any single bit can be recovered with high probability. This power comes with a cost. We show the converse of Aaronson's result; powerful random access codes imply poor learnability. Therefore, for theories with more general nonlocal correlations we have the unphysical consequence that we are required to do a number of experiments exponentially large in the size of the system to predict future measurements.

1.6.2 Tests of nonlocality

This chapter begins by introducing the practical difficulties of discovering tests of nonlocality and suggests several remedies to overcome them. In particular, the number of inequalities that must be checked to determine whether correlations are nonlocal are at least exponential in the problem parameters. Therefore we propose to replace an exponential number of linear inequalities with a polynomial number of nonlinear inequalities. We consider several alternatives based on convex optimization techniques for accomplishing this and compare them.

We then consider some connections with statistical learning theory. We point out a parallel between Bell inequalities and Bayesian networks with hidden nodes, which suggests an alternate use for "tests of nonlocality" as "tests of graph structure." We also consider the applicability of our techniques based on convex geometry to the problem of learning classifiers from clusters of sample data points.

Finally we consider some novel alternative methods for finding Bell inequalities based on algebraic geometry, though we ultimately find these methods too inefficient to be useful.

1.6.3 Nonlocality in curved space

In the final chapter, we consider nonlocality to be a physical resource, present even in the vacuum of a quantum field theory. We consider the *entangling power* of a field to be its ability to entangle two previously unentangled spacelike separated quantum detectors. We explore how the curvature of spacetime affects the entangling power of the field in the particular case where the curvature corresponds to an inflating universe. We show that although, locally, a flat, heated universe and an inflating one look identical, entanglement can be used to distinguish the two situations.

This demonstrates another connection between spacetime curvature and quantum information beyond the usual example of black holes. Hopefully, a better understanding of how spacetime curvature relates to information properties will someday lead to a better synthesis between gravity and quantum mechanics.

Chapter 2

Information processing in nonlocal theories

The man bent over his guitar,
A shearsman of sorts. The day was green.

They said, "You have a blue guitar,
You do not play things as they are."

The man replied, "Things as they are
Are changed upon the blue guitar."

And they said then, "But play, you must,
A tune beyond us, yet ourselves,

A tune upon the blue guitar
Of things exactly as they are."

Wallace Stevens

Exploring the implications of fundamentally incompatible measurements has altered our paradigms about physical theories. Nonlocal correlations and uncertainty relations both express relationships among the expected outcomes of incompatible measurements. What relationships are possible and what are their physical consequences? Quantum mechanics imposes very stringent restrictions [21], and we would very much like to understand their extent and implications. To this end, it is instructive to remove some of these restrictions and investigate how our ability to perform information processing tasks changes as a result.

A popular approach has been to consider only local measurements and the consequences of relaxing the possible violation of the CHSH inequality while obeying the no-signaling principle. By instead focusing on relaxing the restrictions on uncertainty relations, which hold for any incompatible measurements, we eliminate this unnecessary fixation on local measurements. We show that Tsirelson's bound is actually a direct consequence of the uncertainty relations of [22], and relaxing these relations still leads to a theory which maximally violates the CHSH inequality while respecting the no-signaling principle. We explore the consequences of allowing more general types of measurements and point out information processing differences between the two approaches. We use these results to show that states "more nonlocal" than quantum mechanics would be "hard to learn".

2.1 Background

The existence of nonlocal correlations in quantum mechanics that are stronger than those allowed by local realism [23], but yet strictly weaker than those consistent with the no-signaling principle [12] poses an enigma to the understanding of the foundations of quantum physics. What are the properties of quantum mechanics that disallow these stronger correlations [7]? And, what possibilities would be opened by the existence of these correlations? Much of the work exploring these questions has focused on the "box paradigm" that was initially inspired by the CHSH inequality [11]. This particular Bell inequality [23] can be cast into a form of a simple game between two players, Alice and Bob. When the game starts, Alice and Bob are presented with randomly and independently chosen questions $s \in \{0,1\}$ and $t \in \{0,1\}$, respectively. They win if and only if they manage to return answers $a \in \{0,1\}$ and $b \in \{0,1\}$ such that $s \cdot t = a \oplus b$. Alice and Bob may thereby agree on any strategy before the game starts, but may not communicate afterwards. Classically, that is in any model based on local realism, this strategy consists of shared randomness. It has been shown [11] that for any such strategy we have

$$\gamma := \frac{1}{4} \sum_{s,t \in \{0,1\}} \Pr[s \cdot t = a_s \oplus b_t] \leq \frac{3}{4},$$

where $\Pr[s \cdot t = a_s \oplus b_t]$ is the probability that Alice and Bob return winning answers a_s and b_t when presented with questions s and t. Quantumly, Alice and Bob may choose any shared quantum state together with local measurements as part of their strategy. This allows them to violate the inequality above, but curiously only up to a value

$$\gamma \leq \frac{1}{2} + \frac{1}{2\sqrt{2}},$$

known as Tsirelson's bound [13, 24]. We will see later that there exists a state $|\Psi\rangle_{AB}$ shared by Alice and Bob that achieves this bound when Alice and Bob perform measurements given by the observables $A_0 = B_0 = X$ and $A_1 = B_1 = Z$ where we use A_s and B_t to denote the measurement corresponding to questions s and t respectively. The non-signaling principle that disallows faster than light communication between Alice and Bob alone does not impose such a restrictive bound. Hence, Popescu and Rohrlich [12, 25, 26] raised the question why nature is not more "nonlocal"? That is, why does quantum mechanics not allow for a stronger violation of the CHSH inequality up to the maximal value of 1? To gain more insight into this question, they constructed a toy-theory based on so-called PR-boxes [6]. Each such box takes inputs $s, t \in \{0,1\}$ from Alice and Bob respectively and simply outputs randomly chosen measurement outcomes a_s,b_t such that $s \cdot t = a_s \oplus b_t$. Each such box can be used exactly once, and no notion of post-measurement states exists. Note that Alice and Bob still cannot use this box to transmit any information. However, since we have for all s and t that

$\Pr[s \cdot t = a_s \oplus b_t] = 1$, Tsirelson's bound is clearly is violated. It is interesting to consider how our ability to perform information processing tasks changes, if PR-boxes indeed existed. For example, it has been shown that Alice and Bob can use such PR-boxes to compute any Boolean function $f : \{0,1\}^{2n} \rightarrow \{0,1\}$ of their individual inputs $x \in \{0,1\}^n$ and $y \in \{0,1\}^n$ by communicating only a *single* bit [5], which is even true when the boxes have slight imperfections [27].

Much interest has since been devoted to the study of such PR-boxes and their generalizations known as nonlocal boxes [28, 29, 30, 31, 32, 33, 34]. In particular, they have been incorporated in a very nice way into generalized non-signaling theories (GNST) due to Barrett [18] (the relation of such theories to generalizations of quantum theory is due to Hardy [7]) as a means of exploring foundational questions in quantum information. Intuitively, such theories allow for "boxes" involving many more inputs for one or more players/systems, and also allow for some transformations between such boxes. Both theories seek out physically motivated properties that single out quantum mechanics from other theories such as the classical world. These theories have also found interesting applications in deriving new bounds for quantum mechanics itself, e.g., monogamy of entanglement [35].

In such a theory, n-partite states are characterized by the probabilities of obtaining certain outcomes when performing a fixed set of local fiducial measurements on each system. For example, to describe a nonlocal box, consider a bipartite system, where Alice holds the first and Bob the second system. We will label both Alice and Bob's measurements using X and Z in analogy to the quantum setting. For convenience we will also label the outcomes using $a, b \in \{0,1\}$, where the actual outcomes of X and Z in the quantum setting could be recovered as $(-1)^a$, and use $p(A|M)$ to denote the probability of obtaining outcomes A for measurements M. A nonlocal box is now given by the probabilities $p(0,0|X,X) = p(0,0|X,Z) = p(0,0|Z,X) = 1/2$, $p(1,1|X,X) = p(1,1|X,Z) = p(1,1|Z,X) = 1/2$, $p(0,1|Z,Z) = p(1,0|Z,Z) = 1/2$ and $p(A|M) = 0$ otherwise. We will describe such theories in more detail in Section 2.4. We will also refer to GNST using the commonly used term "box-world".

2.1.1 Relaxed uncertainty relations

Even when allowing more than two measurements and outcomes, such boxes remain very artificial constructs and it is not quite clear how they relate to quantum theory. In this note, we hope to provide a more intuitive understanding by showing that superstrong correlations can indeed be obtained by relaxing an uncertainty relation known to hold in quantum theory. Consider *any* anti-commuting observables $\Gamma_1, \ldots, \Gamma_{2n}$ satisfying

$$\{\Gamma_j, \Gamma_k\} = 0$$

whenever $j \neq k$ and

$$\Gamma_j^2 = \mathbb{I},$$

for any $j \in [2n]$, and let $\Gamma_0 = i\Gamma_1 \ldots \Gamma_{2n}$ (see Section 2.2 on how to construct such operators). It was shown in [22] that any quantum state obeys

$$\sum_{j=0}^{2n} \operatorname{Tr}\left(\Gamma_j \rho\right)^2 \leq 1, \tag{2.1}$$

which also lead to several entropic uncertainty relations for such observables. To see why Eq. (2.1) itself can be understood as an uncertainty relation note that $\operatorname{Tr}(\Gamma_j \rho)$ is the expectation value of measuring the observable Γ_j on ρ. The probability of obtaining a measurement outcome $b \in \{\pm 1\}$ can furthermore be written as $p(b|\Gamma_j) = 1/2 + b \operatorname{Tr}(\Gamma_j \rho)/2$. Hence, $\operatorname{Tr}(\Gamma_j \rho)$ can also be understood as the bias towards a particular measurement outcome. Eq. (2.1) now tells us that this bias cannot be arbitrarily large for all measurements Γ_j. Note that we could rewrite the condition of Eq. (2.1) as $||v||_2^2 \leq 1$ where $v = (\operatorname{Tr}(\Gamma_1 \rho), \ldots, \operatorname{Tr}(\Gamma_{2n} \rho))$. Whereas the uncertainty relations of [22] may appear unrelated to the problem of determining the strength of nonlocal correlations, we will see later that Tsirelson's bound for the CHSH inequality is in fact a consequence of Eq. (2.1), when we use the fact that local anti-commutation and maximal violations of the CHSH inequality are closely related [13, 36, 37]. Thus, as one might intuitively guess, bounds for the strength of nonlocal correlations are indeed closely related to uncertainty relations, and such connections have been observed in a different form by [31, 18].

What happens if we merely ask for $||v||_p^p \leq 1$, where $|| \cdot ||_p$ is the p-norm of the vector v? Since Eq. (2.1) must hold for any quantum state, that is for any positive semi-definite matrix ρ with $\operatorname{Tr}(\rho) = 1$, it is clear that this allows operators ρ which are no longer positive semi-definite. In the spirit of Barrett's GNST, we will however restrict ourselves to allowing a particular set of fiducial measurements only, for which the probabilities will remain positive and thus well-defined. In Section 2.3, we will describe a hierarchy of such "theories" in detail, and investigate their power with respect to nonlocal correlations and information processing problems. In particular, we will see that

- For the CHSH inequality, we can obtain at most

  $$\gamma = \frac{1}{2} + \frac{1}{2(2)^{1/p}} \text{ for } ||v||_p^p \leq 1.$$

  where in the limit of $p \rightarrow \infty$ the right-hand side becomes 1, and we have a state that acts analogous to a nonlocal box.

- Furthermore, any unique XOR-game can be played with perfect success for $p \rightarrow \infty$.

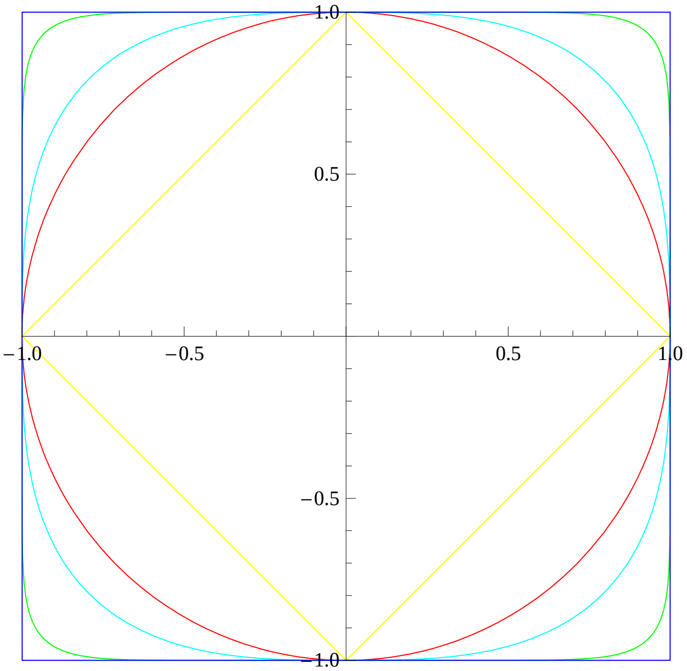


Figure 2.1: p-norm unit circles in dimension 2 for $p = 1, 2, 3, 10, 10000$

It is instructive to consider what our relaxed uncertainty relation means in the case of a single qubit. Note that for quantum mechanics we have $p = 2$ in which case Eq. (2.1) corresponds to the statement that v must lie inside the Bloch sphere. Allowing different values of p now constraints us to the corresponding p-spheres as depicted in Figure 2.1. It is interesting to consider that even though for $p > 2$ we obtain nonlocal correlations that are *stronger* than what quantum theory allows, we now have a *weaker* uncertainty relation than in quantum theory. It has previously been noted by Barrett [18] that GNST has no uncertainty relations for particular measurements. Our work makes this relation very intuitive. In particular, for the case of $p \to \infty$ corresponding to a nonlocal box we essentially place no restrictions on the bias $\text{Tr}(\Gamma_j \rho)$ at all. Since Eq. (2.1) leads to the entropic uncertainty relations on which the security of the protocols in the bounded-quantum-storage model [38, 39, 40] is based, it may be worth considering how certain cryptographic tasks change in the setting of nonlocal boxes. Indeed, it has recently been shown [41] that privacy amplification fails in a world based on nonlocal boxes. Whereas it is known that cryptographic tasks such as bit commitment and oblivious transfer are compatible with the no-signaling principle [29], little is known about them in general theories [42].

It should be noted that except for a single qubit, Eq. (2.1) is of course only a necessary and not a sufficient condition for $\rho \geq 0$. In higher dimensions, such relations are much more involved, but have been obtained for certain operators [43, 44, 45] and also some operators relating more closely to unbiased measurements [46]. Relaxing this particular uncertainty relation is thus only one way to go. Yet, due to the rich structure of the Clifford algebra of operators $\Gamma_1, \ldots, \Gamma_{2n}$ and their central importance for entropic uncertainty relations and so-called XOR nonlocal games (also known as two-party correlation inequalities) with 2 measurement outcomes, this small relaxation allows us to gain some insights into their role in quantum information processing tasks.

2.1.2 Information processing in generalized nonlocal theories

Inspired by these relaxations in terms of an operator ρ, we then construct a hierarchy of p-GNST theories exhibiting similar constraints. For such theories, we identify a single gbit with a single qubit obeying the relaxed uncertainty relations above. That is, we will think of a single gbit as allowing three fiducial measurements labeled X, Z and Y in analogy to the quantum case. Whereas this choice is of course again quite arbitrary, and heavily inspired by the quantum setting, it will allow us to gain a slightly better understanding of the relation of "box-world" and quantum theory later on. We show that the states we allow above, as well as states in p-GNST's have several properties that set them apart from quantum theory. In particular, we will see that

- In p-GNST, there exist superstrong random access encodings. For example, there exists an encoding of $N = 3^n$ bits into $(2n+1)^{3/p}n$ *gbits* such that we can retrieve any bit with probability $1 - \varepsilon$ for $\varepsilon = 2\exp(-(2n+1)^{1/p}/2)$. Quantumly on the other hand it is known that we require at least $(1 - h(1-\varepsilon))N$ *qubits* to encode N classical bits with the same recovery probability, where h denotes the binary Shannon entropy.

- As a consequence, in p-GNST there exists single server PIR scheme with $O(\text{polylog}(N))$ bits of communication for an N bit database with large N, whereas quantumly $\Omega(N)$ bits are needed.

- On the other hand, we show that in GNST it becomes much harder to learn a state in the sense of [20]. In fact, unlike in the quantum setting, we can essentially not ignore even a small part of the information we are given about a state.

It may not be surprising that such effects exist for Hermitian operators ρ, when all we essentially demand is that the condition $||v||_p^p \leq 1$ is obeyed for any set of anti-commuting measurements. However, it will be interesting to consider why for example the superstrong random access code encodings we find above are disallowed in quantum theory, but allowed in GNST.

2.1.3 Commuting measurements

Although the results of local measurements suffice to describe quantum states [7], our results suggest that building a toy-theory around local measurements acting on fixed systems alone (such as GNST) may miss part of the flavor when considering some applications. Quantum mechanics has a rich structure of commuting and anti-commuting measurements built in which make no particular reference to locality. Uncertainty relations impose restrictions for non-commuting measurements, such as for example the anti-commuting measurements $\Gamma_1, \ldots, \Gamma_{2n}$. However, we will see in Section 2.2.4 that also certain sets of commuting measurements cannot have arbitrary expectation values when measured on a particular state ρ. As a simple example, consider a 2 qubit system shared between Alice and Bob, and consider the measurement $X \otimes \mathbb{I}$, $\mathbb{I} \otimes X$ and $X \otimes X$. Suppose that we have $\text{Tr}((X \otimes \mathbb{I})\rho) = \text{Tr}((\mathbb{I} \otimes X)\rho) = 1$. This tells us that when Alice and Bob measure X locally, they obtain an outcome of "1" each with probability 1. However, the measurement of $X \otimes X$ can very intuitively be viewed as Alice and Bob performing a local measurement of X and taking the product of their outcomes. Hence, we do not expect a simultaneous assignment of $\text{Tr}((X \otimes X)\rho) = -1$ to be consistent with the previous two expectation values. We will formalize this intuition in Section 2.2.4, where we will derive a series of conditions such expectation values must obey which in spirit is similar to [21].

GNST does satisfy these conditions for measurements that commute because they act on different subsystems. It does not exhibit any inconsistencies otherwise, as no commutation relations are defined for measurements on the same system. The issue of such inconsistencies is further circumvented by the simple fact that a nonlocal box can only be used once, and there is no notion of subsequent measurements on the same system. This of course is perfectly adequate for studying the strength of nonlocal correlation between two space-like separated systems for example, and led to such perplexing results as [5]. We will however see that it is essentially this lack of additional constraints that allows us to form superstrong random access codes for example, and may indicate that using "box-world" to investigate the role of the strength of nonlocal correlations within quantum theory itself is possibly doomed to fail. It also indicates why defining a consistent notion of 'post-measurement' states for nonlocal boxes is quite difficult, since many constraints that would allow such a task to succeed are simply not present in box-world.

To see how box-world differs from quantum theory consider the measurements $M_1 = X \otimes Z$, $M_2 = Z \otimes X$ and $M_3 = -XZ \otimes XZ$. These are related in exactly the same way as the measurements we considered above, except that in GNST there is no notion that M_1 and M_2 commute. Yet, we intuitively expect similar conditions to hold as for the measurements above when trying to form an analogy to the quantum setting. Indeed, one can easily construct a unitary transformation that maps the measurements M_1, M_2, and M_3 into a form analogous to the above, where two of the measurements act on different systems. [1] In GNST, however, the separation into different systems is always a given, which may lead to difficulties when examining some problems which are not really concerned with correlations among two distant systems alone, but to information processing in general.

[1] Consider $U = (\mathbb{I} \otimes H)\text{CNOT}(\mathbb{I} \otimes H)$

2.1.4 Outline

Whereas we only examine a very small piece of the puzzle, our work hopes to shed some light on the relation between uncertainty relations, nonlocal correlations and the role of above mentioned consistency constraints in information processing. In Section 2.2 we first explain the basic concepts we need to refer to. commuting measurements in more detail. In Section 2.3 we then define a range of simple "theories" obtained by relaxing the uncertainty relation for anti-commuting observables. To highlight the analogy with nonlocal boxes, we then define a range of similar GNST-like theories in Section 2.4. In Sections 2.5, 2.6, and 2.7 we then investigate the power of such theories with respect to nonlocal correlations, random access codes, and information processing problems respectively. In Section 2.2.4 we then investigate why such effects are possible within GNST, but not in quantum theory. Table 2.7.4 summarizes similarities and differences among theories.

2.2 Preliminaries

2.2.1 Basic concepts

In the following, we write $[n] := \{1, \ldots, n\}$ and use X, Z and Y to denote the well-known Pauli matrices [14]. We also speak of a *string of Paulis* to refer to a matrix of the form

$$S_{ab} := X^{a_1}Z^{b_1} \otimes \ldots \otimes X^{a_n}Z^{b_n}, \tag{2.2}$$

with $a = (a_1, \ldots, a_n)$, $b = (b_1, \ldots, b_n)$ and $a_j, b_j \in \{0, 1\}$. We sometimes write the Pauli operator acting on subsystem j, with identity on the other subsystems as

$$X_j = \mathbb{I}^{\otimes j-1} \otimes X \otimes \mathbb{I}^{\otimes n-j-1}$$

.

The *Pauli basis expansion* of a density matrix ρ is given by $\rho = (\mathbb{I} + \sum_{a,b} s_{ab}S_{ab})/d$, where we call s_{ab} the *coefficient* of S_{ab}. Consider the form $f(a, b, a', b') = (a, b') + (a', b)$, where we write $(a, b) = \sum_j a_j b_j \mod 2$. It it straightforward to convince yourself that for any pair S_{ab} and $S_{a'b'}$ either $[S_{ab}, S_{a'b'}] = 0$ if $f(a, b, a', b') = 0$ or $\{S_{ab}, S_{a'b'}\} = 0$ if $f(a, b, a', b') = 1$. Whereas Eq. (2.1) holds for any choice of anti-commuting measurements, it is worth noting that in dimension $d = 2^n$ we can find at most $2n + 1$ anti-commuting operators given by

$$\begin{aligned} \Gamma_{2j-1} &= Y^{\otimes(j-1)} \otimes X \otimes \mathbb{I}^{\otimes(n-j)} \\ \Gamma_{2j} &= Y^{\otimes(j-1)} \otimes Z \otimes \mathbb{I}^{\otimes(n-j)}, \end{aligned}$$

for $j = 1, \ldots, n$ and $\Gamma_0 = i\Gamma_1 \ldots \Gamma_{2n}$. Note that for $n = 1$ we have $\Gamma_1 = X$, $\Gamma_2 = Z$, $\Gamma_0 = Y$ and Eq. (2.1) is equivalent to the Bloch sphere condition. We will also need the notion of a p-norm of a vector $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$ which is defined as

$$||v||_p := \left(\sum_{j=1}^{n} |v_j|^p \right)^{1/p} .$$

Note that for $p = 2$ this is just the Euclidean norm. Of particular interest to us will also be the ∞-norm defined as $||v||_\infty := \lim_{p\to\infty} ||v||_p$ which can also be written as

$$||v||_\infty = \max(|v_1|, \ldots, |v_n|).$$

2.2.2 Probability distributions

Unlike previous descriptions of general probabilistic theories, our notation must be versatile enough to accommodate arbitrary choices of simultaneous commuting measurements, even if they do not act on separate subsystems. In quantum mechanics we may choose to measure $X \otimes X$ along with either $X \otimes \mathbb{I}, \mathbb{I} \otimes X$, or $Z \otimes Z, XZ \otimes XZ$. We will see that including this flexibility in a more general theory leads to new constraints.

First, we want to consider some finite set of measurements $\mathcal{O} = \{M_1, \ldots, M_N\}$ where without loss of generality we assume that each measurement has the same finite set of outcomes $\mathcal{A}$ and the $\mathcal{O}$ is ordered lexiocraphically. Although we initially impose no structure on $\mathcal{O}$, in analogy to quantum mechanics we consider certain collections of measurements $C \subseteq \mathcal{O}$ to have some property which directly corresponds to simultaneous measurability. In particular, we will consider the set of possible experiments

$$\mathcal{E} := \{C \subseteq \mathcal{O} \ \wedge \ \forall M_i, M_j \in C \ \mathrm{sim}(M_i, M_j) = 0\},$$

where "sim" is a predicate indicating simultaneous measurability that remains to be specified. Of particular concern to us will be the probability distributions p over the outcomes $A \in \mathcal{A}^{\times|C|}$ of some set of simultaneously performed measurements $C \in \mathcal{E}$. We use $p(A|C)$ to denote the probability of obtaining outcomes $A = (A_1, A_2, \ldots, A_{|C|}) \in \mathcal{A}^{\times|C|}$ for measurements $C \subseteq \mathcal{O}$ where we wlog take C to be ordered lexicographically. For simplicity, we will also write $p(A_1, \ldots, A_n|M_1, \ldots, M_n) := p((A_1, \ldots, A_n)|\{M_1, \ldots, M_n\})$.

What conditions do the functions $p : \mathcal{A}^{\times|C|} \times C \to [0, 1]$ have to fulfill be a valid probability distribution for any experiment $C \in \mathcal{E}$? We require that the following conditions need to be satisfied for *any* probability distribution

(1) Normalization: $\forall C \in \mathcal{E}, \sum_{A \in \mathcal{A}^{\times |C|}} p(A|C) = 1$.

(2) Positivity: $\forall C \in \mathcal{E}, \forall A \in \mathcal{A}^{\times |C|}, p(A|C) \geq 0$.

The next condition may appear unfamiliar at first glance. Intuitively it says that the distributions of outcomes we obtain for commuting measurements are independent of what other commuting measurements we perform.

(3) Independence:

$$\forall C, C' \in \mathcal{E} \text{ with } C \subseteq C', \quad p((A_1, \ldots, A_{|C|})|C) = \sum_{A_{|C|+1}, \ldots, A_{|C'|} \in \mathcal{A}^{\times |C'|}} p((A_1, \ldots, A_{|C'|})|C'),$$

where, without loss of generality, we take the first $|C|$ outcomes to be associated with the measurements in C.

Throughout this text, we explore the result of choosing two different ways of choosing simultaneous measurements. First, we consider simultaneous measurements on distinct systems as reflected in the construction of nonlocal boxes. Second, we consider a more general notion of such measurements based on commutation relations as in quantum mechanics. Note that in the quantum case such sets of mutually commuting measurements induce a partitioning of the Hilbert space into different systems in the finite-dimensional setting [47, 21].

Consider the set of measurements $\mathcal{O}_P$ to be strings of Paulis on n-partite systems as defined in Section 2.2.1. The two different notions of simultaneous measurements can now be expressed in two different choices of $\text{sim}(M_i, M_j)$, leading to two different sets of realizable experiments. To capture the first notion, we let

$$\mathcal{E}_L := \{C \subseteq \mathcal{O}_P \ \wedge \ \forall M_i, M_j \in C \ \text{local}(M_i, M_j) = 0\},$$

where $\text{local}(M_i, M_j) = 0$ if and only if M_i and M_j act on different subsystems. For example, we have $\text{local}(X \otimes \mathbb{I}, \mathbb{I} \otimes Z) = 0$. Second, we let

$$\mathcal{E}_C := \{C \subseteq \mathcal{O}_P \ \wedge \ \forall M_i, M_j \in C \ [M_i, M_j] = 0\},$$

where all commuting measurements are simultaneously observable, as in quantum mechanics. Clearly, $\mathcal{E}_L \subseteq \mathcal{E}_C$, since two measurements acting on two different subsystems commute.

When we restrict ourselves to $\mathcal{E}_L$ we can express the independence condition from above in the more familiar form of no-signaling:

(3') No-signaling:

$$\forall C, C' \in \mathcal{E}_L \text{ with } C \subseteq C', \;\; p((A_1, \ldots, A_{|C|})|C) = \sum_{A_{|C|+1}, \ldots, A_{|C'|} \in \mathcal{A}^{\times |C'|}} p((A_1, \ldots, A_{|C'|})|C').$$

Intuitively, the no-signaling condition just dictates that the marginal distribution of a particular subset of systems is *independent* of the measurement choices on a disjoint subset of systems. Therefore, we can simplify our description of marginals of no-signaling distributions to just $p(A \in \mathcal{A}^{\times |C|}|C') = p(A|C)$, where the measurement choices on other parties are arbitrary. We will later see that imposing only the special case of the no-signaling condition, versus the full independence condition of (3), makes a crucial difference in the power of the resulting theory with respect to encoding information.

Example 2.2.1. *Consider the set of local experiments for two parties with $\mathcal{A} = \{-1, 1\}, \mathcal{O} = \{X_1, Z_1, X_2, Z_2\}$. Let the probability distribution $p(A|C)$ be described by the following table.*

A					
$(1,1)$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	
$(1,-1)$	0	0	0	$\frac{1}{2}$	
$(-1,1)$	0	0	0	$\frac{1}{2}$	
$(-1,-1)$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	
	$\{X_1, X_2\}$	$\{X_1, Z_2\}$	$\{Z_1, X_2\}$	$\{Z_1, Z_2\}$	C

Clearly, we have positivity, and the sum over each measurement setting (column) is 1. *Finally, note that the marginal probability distribution for either party is constant, $\forall C \in \mathcal{E}_L, \forall A_1 \in \mathcal{A}, \sum_{A_2 \in \mathcal{A}} p((A_1, A_2)|C) = \frac{1}{2}$, therefore this distribution is no-signaling.*

2.2.3 Moments

Any finite, discrete probability distribution has a dual representation in terms of a finite number of moments [48]. We define the product of the outcomes $A = (A_1, \ldots, A_{|C|}) \in \mathcal{A}^{\times |C|}$ of a collection of measurements $C \in \mathcal{E}$ as $A^* = \prod_{i=1}^{|C|} A_i$. The moment for this measurement is defined as

$$m(C) := \sum_{A \in \mathcal{A}^{\times |C|}} p(A|C) A^*. \tag{2.3}$$

Note that for the identity measurement this means $m(\mathbb{I}) = 1$ because of normalization. Also, if you consider the moment for some subset of C, by the independence principle this definition gives a unique value which does not depend on the choice of other measurements made simultaneously.

Since we will only be concerned with measurements with two outcomes $\mathcal{A} = \{\pm 1\}$, we now restrict ourselves to this case for simplicity. For the measurement of a single observable $C = \{M_1\}$ with outcome $A_1 \in \mathcal{A}$, we can easily recover the probabilities from the moments as

$$p((A_1)|\{M_1\}) = \frac{1}{2}\left(1 + A_1\ m(\{M_1\})\right). \tag{2.4}$$

In subsequent notation, we will drop the brackets within parentheses when it increases readability.

Note that we can recover the probability for a specific set of outcomes $\hat{A} \in \mathcal{A}^{\times|C|}$ and measurements $C \in \mathcal{E}$ from these moments. Without loss of generality, let $C = \{M_1, \ldots, M_n\}$.

$$\begin{aligned} & \frac{1}{2^n} \sum_{C' \subseteq C} m(C') \prod_{i, M_i \in C'} \hat{A}_i \\ = \quad & \frac{1}{2^n} \sum_{C' \subseteq C} \left(\sum_{A \in \mathcal{A}^{\times|C'|}} p(A|C') \prod_{i, M_i \in C'} A_i \right) \prod_{i, M_i \in C'} \hat{A}_i \\ = \quad & \frac{1}{2^n} \sum_{A \in \mathcal{A}^{\times|C|}} p(A|C) \sum_{C' \subseteq C} \prod_{i, M_i \in C'} A_i \hat{A}_i \end{aligned}$$

The second line simply uses the definition of $m(C')$ and the third line uses the independence principle to write $p(A|C')$ in terms of $p(A|C)$, allowing us to move the sum over C' inside. Now note that the sum over C' can be broken into n sums over whether or not $M_i \in C'$. For each M_i, if it is in C' we get a factor of $A_i\hat{A}_i$, otherwise a factor of 1.

$$= \quad \frac{1}{2^n} \sum_{A \in \mathcal{A}^{\times|C|}} p(A|C) \prod_{i=1}^{n} (1 + A_i \hat{A}_i)$$

Because the outcomes can only be ±1, the sum can give us only 0 or 2.

$$\begin{aligned} = \quad & \frac{1}{2^n} \sum_{A \in \mathcal{A}^{\times|C|}} p(A|C) \prod_{i=1}^{n} 2\delta_{A_i, \hat{A}_i} \\ = \quad & \frac{1}{2^n} \sum_{A \in \mathcal{A}^{\times|C|}} p(A|C)\ 2^n \delta_{A, \hat{A}} \\ = \quad & p(\hat{A}|C) \end{aligned}$$

2.2.4 Consistency constraints

We are now ready to investigate the constraints that arise due to simultaneous measurement of commuting observables and that will play a crucial role in understanding the differences between quantum theory and p-GNST. Imagine two commuting measurements $[M_i, M_j] = 0$, and their product $M_k = M_i M_j$. In quantum mechanics the outcome of the measurement M_k is the same as the product of the outcomes of M_i and M_j, which can be verified by expanding M_k in terms of M_i and M_j and using the fact that they have a joint eigenbasis. What happens if we take this to be true in any theory? If we are only allowed to make local measurements, then this is a moot point. We can only get $X \otimes X$ by measuring $X \otimes \mathbb{I}$ and $\mathbb{I} \otimes X$ and multiplying the results.

But if we are allowed to make any combination of commuting measurements, this will impose some interesting conditions. For example, in the quantum case we may have $M_1 = X \otimes X$, $M_2 = Z \otimes Z$ and $M_3 = XZ \otimes XZ$. To see that this has consequences in terms of the moments, consider the simple example where $m(M_1) = 1$ and $m(M_2) = 1$, which means that we will deterministically observe outcomes $A(M_1) = A(M_2) = 1$. Hence, $m(M_3) = -1$ should intuitively not be compatible with these two moments for M_1 and M_2.

How can we formalize these conditions? For example, Eq. (2.3) gives us that

$$m(M_1 M_2) = m(M_1, M_2),$$

if we insist that outcomes of products of measurements equal the product of outcomes of individual measurements. For a given set of commuting measurements $C = \{M_1, \ldots, M_m\}$ with $M_j^2 = \mathbb{I}$, let $s(M)$ be the 2^m element vector whose k-th entry is given by

$$s := [s(C)]_k := M_1^{k_1} M_2^{k_2} \ldots M_m^{k_m}, \tag{2.5}$$

with $k \in \{0, 1\}^m$ in lexicographic order. We now define the *moment matrix* K_s by letting the entry in the i-row and j-th column be given by

$$[K_s]_{ij} := m(s_i s_j)/2^m.$$

Claim 2.2.2 (Adapted from Wainwright and Jordan [48]). *Let $C = \{M_1, \ldots, M_m\}$ be a set of commuting measurements. Then $K_s \geq 0$ if and only if p is a probability distribution (satisfying constraints (1) and (2)).*

Proof. In addition to K_s, we define two more $2^m \times 2^m$ matrices, whose components are labeled by vectors $i, j \in \{0, 1\}^m$ in lexicographic order as

$$[P]_{ij} = \delta_{ij} p(A = ((-1)^{i_1}, \ldots, (-1)^{i_m})|C).$$
$$[B]_{ij} = \frac{1}{2^{m/2}}(-1)^{i \cdot j},$$

It is easily verified that B is a unitary matrix. Note that B is an example of a Hadamard matrix.

Now we will show that $K_s = BPB^\top$.

$$\begin{aligned}
\left[BPB^\top\right]_{ij} &= \frac{1}{2^m} \sum_{k,l\in\{0,1\}^m} (-1)^{i\cdot k}\ \delta_{kl}\ p(((-1)^{k_1},\ldots,(-1)^{k_m})|C)(-1)^{l\cdot j} \\
&= \frac{1}{2^m} \sum_{k\in\{0,1\}^m} (-1)^{k\cdot(i\oplus j)} p(((-1)^{k_1},\ldots,(-1)^{k_m})|C) \\
&= \frac{1}{2^m} \sum_{k\in\{0,1\}^m} \prod_{t=1}^{m} ((-1)^{k_t})^{(i_t\oplus j_t)} p(((-1)^{k_1},\ldots,(-1)^{k_m})|C) \\
&= \frac{1}{2^m} \sum_{A\in\mathcal{A}^{\times|C|}} \prod_{t=1}^{m} A_t^{i_t} A_t^{j_t} p(A|C) \\
&= \frac{1}{2^m} m(s_i s_j) = [K_s]_{ij}.
\end{aligned}$$

Clearly, if the probabilities $p(A|C)$ are non-negative (2), then $P \geq 0$ if and only if $K \geq 0$ since B is unitary. Similarly, the fact that $m(\mathbb{I}) = 1$, B is unitary and the trace is cyclic ensures that p satisfies condition (1). □

Example 2.2.3. *As an example, consider the case of two commuting measurement M_1 and M_2 with $M_3 = M_1M_2$. We have $s = (\mathbb{I}, M_1, M_2, M_3)$ and*

$$\mathbf{K}_s = \begin{pmatrix} m(\mathbb{I}) & m(M_1) & m(M_2) & m(M_1M_2) \\ m(M_1) & m(\mathbb{I}) & m(M_3) & m(M_2) \\ m(M_2) & m(M_3) & m(\mathbb{I}) & m(M_1) \\ m(M_3) & m(M_3) & m(M_1) & m(\mathbb{I}) \end{pmatrix} \equiv \begin{pmatrix} 1 & a & b & c \\ a & 1 & c & b \\ b & c & 1 & a \\ c & b & a & 1 \end{pmatrix}$$

Demanding that the eigenvalues of this matrix, $\lambda = ((1+a-b-c), (-1+a+b-c), (-1+a-b+c), (1+a+b+c))$, be non-negative is enough to ensure that $\mathbf{K}_s \succeq 0$. Using the Sylvester criteria, we get the alternate constraints that each moment $|a, b, c| \leq 1$ and $1 - a^2 - b^2 - c^2 + 2abc \geq 0$, and $\lambda_1\lambda_2\lambda_3\lambda_4 \geq 0$.

Our examples are reminiscent of the examples considered in the setting of contextuality [49]. Note that our constraints are related, but nevertheless of a different flavor since we only consider such constraints for measurements which all commute. It may be interesting to consider such a moment matrix in order to determine how "non-contextual" quantum theory is. In Sections 2.4.1 and 2.3 we will develop classes of states which are restricted by imposing specific relationships among various moments. In particular, it will be of crucial importance whether we merely impose such constraints for measurements acting on different systems, or include such constraints for all commuting measurements.

2.3 p-nonlocal theories and their properties

We now define a series of so-called p-nonlocal "theories", each one more constrained than the previous. Our definition is thereby motivated by the uncertainty relations of [22] stated above. We later relate our definitions to Barrett's GNST [18] and what are commonly known as nonlocal boxes. Our aim by constructing this series of simple theories is thereby merely to gain a more intuitive understanding of superstrong nonlocal correlations due to nonlocal boxes.

2.3.1 A theory without consistency constraints

We start with the simplest of all p-theories, which forms the basis of all subsequent definitions. In essence, we will simply allow states violating the uncertainty relation in 2.1 without worrying about anything else. In the spirit of Barrett [18] we start by defining the states which are allowed in our theory, and then allow all linear transformations preserving the set of allowed states. For simplicity, we will only consider the case of $d = 2^n$.

Definition 2.3.1. *A d-dimensional* p-bin state *is a $d \times d$ complex Hermitian matrix*

$$\rho = \frac{1}{d}\left(\mathbb{I} + \sum_{a,b} s_{ab} S_{ab}\right)$$

satisfying

1. *for all a, b, $-1 \leq s_{ab} \leq 1$.*
2. *for any set of mutually anti-commuting strings of Paulis $A_1, \ldots, A_m \in \mathbb{C}^{d \times d}$*

$$\sum_j |\operatorname{Tr}(A_j \rho)|^p \leq 1.$$

It remains to be specified what operations and measurements we are allowed to perform on p-bin states. We define

Definition 2.3.2. *A d-dimensional* p-bin theory *consists of*

1. *states $\rho \in \mathcal{S}_p^d$ where $\mathcal{S}_p^d$ is the set of d-dimensional p-bin states,*
2. *linear operations $T : \mathcal{S}_p^d \rightarrow \mathcal{S}_p^d$,*
3. *measurements described by observables $S_{ab} = S_{ab}^0 - S_{ab}^1$ where S_{ab}^0 and S_{ab}^1 are projectors onto the positive and negative eigenspace of S_{ab} respectively. As in the quantum case we let*

$$p_0 = \operatorname{Tr}(\rho S_{ab}^0) \text{ and } p_1 = \operatorname{Tr}(\rho S_{ab}^1).$$

Starting from a state, we may apply any set of operations T followed by a single measurement.

Note that by virtue of Eq. (2.1) any quantum state is a p-bin state. Note that the converse however does not hold, since the conditions given above do not imply that a p-bin state ρ is positive semi-definite. It seems very restrictive to limit ourselves to a single measurement at the end. The reason for this is that for some p, there exist p-bin states to start with, valid operations and measurements, followed by another operation that give us a states that are no longer a p-bin states [50]. We return to this question, when we consider the set of allowed operations below.

Note that the above definition is well-defined. First, we want that for any measurement S_{ab}, $\{p_0, p_1\}$ forms a valid probability distribution. A small calculation gives us that any p-nonlocal state ρ we have

$$p_v = \mathrm{Tr}(\rho S_{ab}^v) = \frac{1}{2}\left(1 + (-1)^v s_{ab}\right),$$

and thus $0 \leq p_b \leq 1$ and $p_0 + p_1 = 1$. Second, we want the non-signaling conditions to hold. When measuring $S_{ab} \otimes S_{a'b'}$ on a bipartite state

$$\rho_{AB} = \frac{1}{d}\left(\mathbb{I} + \sum_{\ell,m,\ell',m'} S_{\ell,m} \otimes S_{\ell',m'}\right)$$

we have that the probability to obtain outcome u for the measurement on the first system is given by

$$\Pr[u|ab, a'b'] = \sum_{v \in 0,1} \mathrm{Tr}\left(\rho_{AB}(S_{ab}^u \otimes S_{a'b'}^v)\right) = \frac{1}{2}(\mathbb{I} + (-1)^u s_{a,b,0,0}),$$

and hence $\Pr[u|ab, a'b'] = \Pr[u|ab, a''b'']$ for all a', b', a'', b'' as desired. A similar argument can be made to show that the more general independence condition is satisfied.

2.3.1.1 Basic Properties

We now state some basic properties of this theory, which will also hold for a more restricted p-nonlocal theory as outlined below.

Claim 2.3.3. *If ρ is a p-bin state, then ρ is also a q-bin state for $p, q \in \mathbb{Z}$ with $q \geq p$.*

Proof. This follows immediately from the fact that for any $r \in [0, 1]$ we have $r^q \leq r^p$. □

Below, we will apply circuits consisting of the Clifford gates $\{CNOT, X, Z, Y, H\}$ and $\mathbb{I}$. It is easy to see that such unitary operations are allowed transformations taking p-bin states to p-bin states.

Claim 2.3.4. *Let $\rho \in \mathcal{S}_p^d$. Then for any circuit U consisting solely of the gates $\{CNOT, X, Z, Y, H, \mathbb{I}\}$ we have $U\rho U^\dagger \in \mathcal{S}_p^d$.*

Proof. Note that U is composed of single unitaries $U_j = \mathbb{I}^{j-1} \otimes V \otimes \mathbb{I}^{n-j}$ with $V \in \{X, Z, Y, H\}$ and unitaries $U'_j = \mathbb{I}^{j-1} \otimes \text{CNOT} \otimes \mathbb{I}^{n-j-1}$. First, it is straightforward to verify that for any $a, b \in \{0,1\}^n$, there exist $a', b' \in \{0,1\}^n$ such that $U_j S_{ab} U_j^\dagger = S_{a'b'}$, and similarly for U'_j. Second, applying a unitary to any set of anti-commuting operators again gives us anti-commuting operators. Hence, since we have $\sum_j |\operatorname{Tr}(A_j\rho)|^p \leq 1$ for *any* set of anti-commuting strings of Paulis, the resulting state will also have this property. □

It will also be useful to know that

Claim 2.3.5. *Let $\rho_1, \ldots, \rho_n \in \mathcal{S}_p^2$. Then $\bigotimes_{i=1}^n \rho_i \in \mathcal{S}_p^{2^n}$.*

Proof. We proceed by induction. By assumption, $\rho_1 \in \mathcal{S}_p^2$. We will show that for any states $\rho \in \mathcal{S}_p^{2^n}, \sigma \in \mathcal{S}_p^2$, the state $\rho \otimes \sigma \in \mathcal{S}_p^{2^{n+1}}$.

We need to prove that for any set of mutually anti-commuting Pauli's $A_j \in \mathbb{C}^{2^{n+1} \times 2^{n+1}}$ $\sum_j |\operatorname{Tr}(A_j \rho \otimes \sigma)|^p \leq 1$. Each A_j can always be written in terms of a Pauli, B_j acting on ρ, plus a Pauli $\{\mathbb{I}, X, Y, Z\}$ on σ. We separate the A_j into groups according to which Pauli is appended to B_j. Then we can rewrite this as

$$\begin{aligned}
&\sum_{j_\mathbb{I}} |\operatorname{Tr}((B_{j_\mathbb{I}} \otimes \mathbb{I})(\rho \otimes \sigma))|^p + \sum_{j_X} |\operatorname{Tr}((B_{j_X} \otimes X)(\rho \otimes \sigma))|^p \\
+&\sum_{j_Y} |\operatorname{Tr}((B_{j_Y} \otimes Y)(\rho \otimes \sigma))|^p + \sum_{j_Z} |\operatorname{Tr}((B_{j_Z} \otimes Z)(\rho \otimes \sigma))|^p \\
&= \sum_{j_\mathbb{I}} |\operatorname{Tr}(B_{j_\mathbb{I}}\rho)|^p + \sum_{j_X} |\operatorname{Tr}(B_{j_X}\rho)|^p |\operatorname{Tr}(X\sigma)|^p \\
+&\sum_{j_Y} |\operatorname{Tr}(B_{j_Y}\rho)|^p |\operatorname{Tr}(Y\sigma)|^p + \sum_{j_Z} |\operatorname{Tr}(B_{j_Z}\rho)|^p |\operatorname{Tr}(Z\sigma)|^p \leq 1.
\end{aligned}$$

Since all the A_j mutually anti-commute, then for different j, j', $\{B_j \otimes X, B_{j'} \otimes X\} = 0$ implies $\{B_j, B_{j'}\} = 0$, while $\{B_j \otimes X, B_{j'} \otimes Y\} = 0$ implies $[B_j, B_{j'}] = 0$. Then because $\rho \in \mathcal{S}_p^{2^n}$ and $\{B_{j_X}, B_{j'_X}\} = 0$, and, for similar reasons $\{B_{j_X}, B_{j_\mathbb{I}}\} = \{B_{j'_\mathbb{I}}, B_{j_\mathbb{I}}\} = 0$, we know

$$\sum_{j_\mathbb{I}} |\operatorname{Tr}(B_{j_\mathbb{I}}\rho)|^p + \sum_{j_X} |\operatorname{Tr}(B_{j_X})|^p \leq 1.$$

Now we will shorten our notation by writing

$$\begin{aligned}
a_X &= |\operatorname{Tr}(X\sigma)|^p & b_X &= \textstyle\sum_{j_X} |\operatorname{Tr}(B_{j_X}\rho)|^p \\
a_Y &= |\operatorname{Tr}(Y\sigma)|^p & b_Y &= \textstyle\sum_{j_Y} |\operatorname{Tr}(B_{j_Y}\rho)|^p \\
a_Z &= |\operatorname{Tr}(Z\sigma)|^p & b_Z &= \textstyle\sum_{j_Z} |\operatorname{Tr}(B_{j_Z}\rho)|^p \\
& & b_\mathbb{I} &= \textstyle\sum_{j_\mathbb{I}} |\operatorname{Tr}(B_{j_\mathbb{I}}\rho)|^p.
\end{aligned}$$

This allows us to write inequalities implied by the uncertainty relation like:

$$\begin{aligned} a_X + a_Y + a_Z &\leq 1 \\ b_X + b_{\mathbb{I}} &\leq 1 \\ b_Y + b_{\mathbb{I}} &\leq 1 \\ b_Z + b_{\mathbb{I}} &\leq 1. \end{aligned}$$

We can also see that $a_X, a_Y, a_Z, b_X, b_Y, b_Z, b_{\mathbb{I}} \geq 0$. The task at hand is to show that these inequalities imply the one required of a state in $\mathcal{S}_p^{2^{n+1}}$, which we can now rewrite as

$$a_X b_X + a_Y b_Y + a_Z b_Z + b_{\mathbb{I}} \leq 1.$$

We do this by writing down a sum of products of non-negative quantities like $1 - a_X - a_Y - a_Z$ and noting that the result is non-negative.

$$a_X(1 - b_X - b_{\mathbb{I}}) + a_Y(1 - b_Y - b_{\mathbb{I}}) + a_Z(1 - b_Z - b_{\mathbb{I}}) + (1 - b_{\mathbb{I}})(1 - a_X - a_Y - a_Z) \geq 0.$$

That equation can be rewritten as $1 - (a_X b_X + a_Y b_Y + a_Z b_Z + b_{\mathbb{I}}) \geq 0$, which is what we set out to show. Therefore, $\rho \otimes \sigma$ is a valid state, and, by induction, so is $\bigotimes_{i=1}^n \rho_i \in \mathcal{S}_p^{2^n}$ for any n. □

2.3.2 An analogue to box-world

Note that in the above definition we have not placed any constraints at all on the expectation values of commuting measurements. This was not necessary, as we had allowed a single measurement only, where by the above definition $\mathbb{I} \otimes X$ formed such a single measurement. Now consider a two-qubit system, i.e., $d = 4$. Suppose that we have for a particular ρ that

$$\text{Tr}\left((X \otimes \mathbb{I})\rho\right) = \text{Tr}\left((\mathbb{I} \otimes X)\rho\right) = \text{Tr}\left((X \otimes X)\rho\right) = -1.$$

Note that ρ can be a perfectly valid state with respect to the definition given above, but yet we would not consider this to be consistent behavior, if we were allowed to perform subsequent measurements. We now introduce additional constraints that eliminate this inconsistency. It should be clear from Section 2.2.3 that that to achieve full consistency we would have to introduce certain constraints for commuting observables in general. Yet, we will first restrict ourselves to observables on different systems in analogy to "box-world". We will show in Section 2.4.1 that Barrett's GNST and nonlocal boxes essentially correspond to this definition. We will also see in Sections 2.6 and 2.7.1 that these additional constraints play a crucial role in the power of our model with respect to information processing tasks.

Definition 2.3.6. *A* p-box state *is a p-bin state ρ, where in addition we require that for any set $C \in \mathcal{E}_L$ of measurements acting on different systems and $s(C)$ as defined in Eq. (2.5) we have that the corresponding moment matrix K_s defined in Section 2.2.4 satisfies*

$$K_s \geq 0.$$

Note that Claims 2.3.3 and 2.3.5 hold analogously for p-box states. It is important to note though that Claim 2.3.4 does not hold in this case, since for example the CNOT operation can lead to states violating the definition.

2.3.3 A theory with consistency constraints

Finally, we will impose all constraints required from our consistency considerations of Section 2.2.3.

Definition 2.3.7. *A* p-nonlocal state *is a p-box state ρ, where in addition we require that for any set of commuting measurements $C \in \mathcal{E}_C$ and $s(C)$ as defined in Eq. (2.5) we have that the corresponding moment matrix K_s as defined in Section 2.2.4 satisfies*

$$K_s \geq 0.$$

Again Claims 2.3.3 and 2.3.5 hold analogous to the above. When we include all consistency considerations, it is also easy to see that Claim 2.3.4 holds for p-nonlocal states, since for any allowed unitary U we already have by the above that ρ satisfies the constraints given by the set $C' = \{U^\dagger M_1 U, \ldots, U^\dagger M_m U\}$ and hence $U \rho U^\dagger$ remains a valid p-nonlocal state.

2.4 Generalized nonlocal theories

To create a closer analogy between our "theories" derived from relaxed uncertainty relations and nonlocal boxes, we now consider a related class of theories called *generalized no-signaling theories* (GNST) [18], for which we will consider similar relaxations. As already sketched in the introduction, states in a GNST are defined operationally. Consider a laboratory setup where we have a device which prepares a specific state. We then use a measuring device which has a choice of settings allowing us to measure different properties of the system. The measuring device gives us a reading specifying the outcome of the measurement. A particular state in GNST is described completely by means of the probabilities of obtaining each outcome when performing a fixed set of *fiducial* measurements. For example, for a set of fiducial measurements $\mathcal{O} = \{X, Z, Y\}$ with outcomes $\mathcal{A} = \{\pm 1\}$, the probabilities $p(A|C)$ for all $A \in \mathcal{A}$ and $C \in \mathcal{O}$ form a description of the state.

Hence, we will simply use p to refer to a state given by said conditional probabilities. The idea behind considering fiducial measurements stems from the idea that there exists a set of measurement choices that suffice to fully describe the system. In classical mechanics, for instance, we can always in principle make a single measurement which outputs all the information necessary to describe a state. For a qubit, on the other hand, we would need results from at least three different incompatible measurement settings, e.g., spin in three orthogonal directions. We refer to [18] for a definition of GNST and its allowed operations. For us it will only be important to note that similar to the setting of nonlocal boxes, we can make only one measurement on each system, and there is no real notion of post-measurement states defined.

In the following, we will be interested in the special case of multi-partite systems where on each system we can perform one of three fiducial measurements with outcomes ± 1. Using our notation from Section 2.2.2 we write the set of realizable experiments for GNST as

$$\mathcal{E}_G = \{\forall k \in \{1,2,3\}^n : \{W_{1,k_1}, \ldots, W_{n,k_n}\}\},$$

with W_{i,k_i} denoting a choice of the k_ith measurement on the ith system. Later we will connect these measurement choices with Pauli measurements via the relation $W_{i,1} = X_i, W_{i,2} = Z_i, W_{i,3} = X_i Z_i$. A key point of this definition will be that the partitioning of measurements into n systems will be fixed. We also demand that probability distributions should satisfy an independence principle. As we pointed out, when restricted to partitions over disjoint parties, this just reduces to the no-signaling principle. That is, the choice of measurement on one subset of particles can not be used to send a signal to a disjoint subset.

In analogy to the quantum setting [18], we let one gbit refer to a single system on which we can perform our set of fiducial measurements given above. Our definition of a gbit thereby slightly differs from the definition given in [18], which only allows two fiducial measurements X and Z on a single gbit. Yet, in order to compare the hierarchy of GNST-like theories we will construct below to the p-box states from above we adopt this slightly more general definition in analogy to a single qubit in the quantum case. Note that for the set of measurements $C \in \mathcal{E}_G$ specified above, an n-gbit state, specified by $p : \mathcal{A}^{\times n} \times C \to [0,1]$, is in GNST if p satisfies constraints (1), (2), and (3') in Section 2.2.2.

Example 2.4.1. *Consider the following state of one particle in GNST (or one gbit):*

$$\begin{aligned}
p(A = +1|M = X) &= s_x = 1 - p(A = -1|M = X) \\
p(A = +1|M = Z) &= s_y = 1 - p(A = -1|M = Z) \\
p(A = +1|M = XZ) &= s_z = 1 - p(A = -1|M = XZ).
\end{aligned}$$

This state is normalized, and positivity requires $s_x, s_y, s_z \in [0,1]$. The state would be equivalent to *the state of an arbitrary qubit if and only if $s_x^2 + s_y^2 + s_z^2 \leq 1$, that is, if we are constrained to the Bloch sphere.*

For multi-partite states the difference between constraints on qubits and gbits becomes more complicated. We now turn to describing a hierarchy of constraints on GNST theories which will be analogous to uncertainty conditions in p-nonlocal theories and quantum mechanics.

2.4.1 p-GNST

Even though states in GNST are defined without any particular structure to their measurements embedded, we will now impose a physically motivated structure. In particular, we will simply *imagine* in analogy to the quantum setting that measurements X, Z and Y obey the same anti-commutation relations as the Pauli matrices $\{X, Z\} = \{Z, Y\} = \{X, Y\} = 0$. In our definition below, we will for simplicity write $\{\cdot, \cdot\}$ to indicate that we imagine such an anti-commutation constraint to hold exactly when the string of Paulis $\prod_i W_{i,k_i}$ associated with each C would anti-commute.

First of all, this will allow us to artificially impose an uncertainty relation just like Eq. (2.1).

Definition 2.4.2. *A state is in p-GNST if it is in GNST and for any set of measurements $S = \{C \in \mathcal{E}_G\}$ satisfying that for all $C, C' \in S$, $\{C, C'\} = 0$ we have*

$$\sum_{C \in S} |m(C)|^p \leq 1. \tag{2.6}$$

Note that for $p \to \infty$ this condition no longer restricts the states, because we get $\max_{C \in \mathcal{S}} |m(C)| \leq 1$, which is true for the original GNST, and nonlocal boxes. If we would actually add such commutation and anti-commutation constraints we could now again distinguish between adding the consistency constraints of Section 2.2.3 only for measurements acting on different systems, or for all commuting measurements in analogy to the p-box and p-nonlocal theories. In analogy to GNST, where commutation relations were only defined for measurements acting on different systems however, we will stick to this setting, even when considering $p < \infty$. A p-GNST state is thus essentially analogous to a p-box state, except we are allowed to make simultaneous measurements of locally disjoint systems.

2.5 Superstrong nonlocality

Before we show that relaxing the uncertainty equation of Eq. (2.1) leads to superstrong nonlocal correlations, let's take a look at what effect this uncertainty relation actually has on quantum strategies for the CHSH inequality. For this purpose, we will rewrite Tsirelson's bound for the

CHSH inequality in its more common form as

$$|\langle A_0 \otimes B_0\rangle + \langle A_0 \otimes B_1\rangle + \langle A_1 \otimes B_0\rangle - \langle A_1 \otimes B_1\rangle| \leq 2\sqrt{2},$$

where we use A_0, A_1 and B_0, B_1 to denote Alice's and Bob's observables respectively where $A_0^2 = A_1^2 = B_0^2 = B_1^2 = \mathbb{I}$. We will use the fact that in order to achieve the maximum possible quantum violation we must have $\{A_0, A_1\} = 0$ and $\{B_0, B_1\} = 0$ [13, 36, 37]. For $M_1 = A_0 \otimes B_0$, $M_2 = A_0 \otimes B_1$, $M_3 = A_1 \otimes B_0$ and $M_4 = A_1 \otimes B_1$ this means that we have $\{M_1, M_2\} = \{M_1, M_3\} = \{M_2, M_4\} = \{M_3, M_4\} = 0$. Using the uncertainty relation of Eq. (2.1) proving Tsirelson's bound is equivalent to solving the following optimization problem

$$\begin{array}{ll} \text{maximize} & \langle M_1\rangle + \langle M_2\rangle + \langle M_3\rangle - \langle M_4\rangle \\ \text{subject to} & \langle M_1\rangle^2 + \langle M_2\rangle^2 \leq 1 \\ & \langle M_1\rangle^2 + \langle M_3\rangle^2 \leq 1 \\ & \langle M_2\rangle^2 + \langle M_4\rangle^2 \leq 1 \\ & \langle M_3\rangle^2 + \langle M_4\rangle^2 \leq 1. \end{array}$$

By using Lagrange multipliers, it is easy to see that for the optimum solution we have $\langle M_1\rangle^2 = \langle M_4\rangle^2$ and $\langle M_2\rangle^2 = \langle M_3\rangle^2$. By considering all different possibilities, we obtain that with $x = \langle M_1\rangle = -\langle M_4\rangle$ and $y = \langle M_2\rangle = \langle M_3\rangle$ our optimization problem becomes

$$\begin{array}{ll} \text{maximize} & 2(x+y) \\ \text{subject to} & x^2 + y^2 \leq 1. \end{array}$$

Again using Lagrange multipliers, we now have that the maximum is attained at $x = y = 1/\sqrt{2}$ giving us Tsirelson's bound.

Tsirelson's bound can hence be understood as a consequence of the uncertainty relation of [22]. Thus, we intuitively expect that relaxing this relation affects the strength of nonlocal correlations. In a similar way, one can view monogamy of nonlocal correlations as a consequence of Eq. (2.1) [51].

2.5.1 CHSH inequality

2.5.1.1 In p-theories

To see what is possible in p-theories, we first construct the equivalent of a maximally entangled state. Let

$$\rho_p = \frac{1}{2}\left[\mathbb{I} + \left(\frac{1}{2}\right)^{\frac{1}{p}} (X + Y)\right].$$

Note that for $p \to \infty$ this gives us

$$\rho_\infty = \frac{1}{2}\left[\mathbb{I} + X + Y\right].$$

We now proceed analogously to the quantum case to construct

$$\eta_1 = \mathrm{CNOT}(\rho_p \otimes |0\rangle\langle 0|)\mathrm{CNOT}^\dagger,$$

which by Claim 2.3.4 is a valid p-bin and p-nonlocal state. It can also be verified that η_1 forms a valid p-box state.

Claim 2.5.1. *Let $A_1 = X$, $A_2 = Y$, $B_1 = X$ and $B_2 = Y$ be Alice and Bob's observables respectively. Then*

$$\langle CHSH_p\rangle = \mathrm{Tr}(\eta_1(A_1 \otimes B_1 + A_1 \otimes B_2 + A_2 \otimes B_1 - A_2 \otimes B_2)) = 4\frac{1}{2^{1/p}},$$

for all p-theories.

Proof. This follows immediately by noting that

$$\eta_1 = \frac{1}{4}\left(\mathbb{I} + \frac{1}{2^{1/p}}\left(X \otimes X + X \otimes Y + Y \otimes X - Y \otimes Y\right) + Z \otimes Z\right).$$

□

We can also phrase this statement in terms of probabilities as stated in the introduction, by noting that the maximum probability that Alice and Bob win the CHSH game is given by

$$\frac{1}{2} + \frac{\langle CHSH_p\rangle}{8} = \frac{1}{2} + \frac{1}{2 \cdot 2^{1/p}}.$$

It is important to note that this violation can be obtained even when imposing the additional consistency constraints from Section 2.2.3.

2.5.1.2 In p-GNST

We already saw in the introduction that GNST admits states analogous to a nonlocal box, allowing for a maximal violation of the CHSH inequality. We now show that similar states exist for p-GNST theories analogous to p-box states. We first phrase the CHSH inequality in terms of probabilities. In particular, consider the GNST state specified by $p((A_1, A_2)|\{M_1, M_2\}) = \frac{1}{4}(1 + (-1)^{\delta_{M_1,Z_1}\delta_{M_2,Z_2}} A_1 A_2 \lambda)$ for some λ to be chosen below. If each party measures X or Z on their state and outputs the result ±1, the probability that Alice and Bob win the CHSH game is given by

$$\frac{1}{4}(p(1,1|X_1,X_2) + p(-1,-1|X_1,X_2) + p(1,1|X_1,Z_2) + p(-1,-1|X_1,Z_2)$$
$$+p(1,1|Z_1,X_2) + p(-1,-1|Z_1,X_2) + p(1,-1|Z_1,Z_2) + p(-1,1|Z_1,Z_2)) = \frac{1+\lambda}{2}.$$

In terms of the moments, $m(X_1,X_2) = m(X_1,Z_2) = m(Z_1,X_2) = -m(Z_1,Z_2) = \lambda$, and this becomes

$$\frac{1}{4}(2+\frac{1}{2}(m(X_1,X_2)+m(X_1,Z_2)+m(Z_1,X_2)-m(Z_1,Z_2))) = \frac{1+\lambda}{2}.$$

Now we can consider the maximum value of λ that is a valid state in p-GNST. The requirements listed in example 2.2.3 only restrict $|\lambda| \leq 1$. Eq. (2.6) requires $|m(X_1,X_2)|^p + |m(X_1,Z_2)|^p = |m(Z_1,X_2)|^p + |m(Z_1,Z_2)|^p = 2|\lambda|^p \leq 1 \rightarrow \lambda = (\frac{1}{2})^{\frac{1}{p}}$. Therefore in a p-GNST it is possible to win the CHSH game with probability $1/2 + 1/(2 \cdot 2^{1/p})$.

2.5.2 XOR games

We now investigate the case of general 2-player XOR-games for $p \rightarrow \infty$. In such a game we have an arbitrary (but finite) set of questions S and T from which Alice's and Bob's questions $s \in S$ and $t \in T$ are chosen according to a fixed probability distribution $\pi : S \times T \rightarrow [0,1]$. Yet, the set of possible answers remain $A = B = \{0,1\}$ for Alice and Bob respectively. The game furthermore specifies a predicate $V : A \times B \times S \times T \rightarrow \{0,1\}$ that determines the winning answers for Alice and Bob. In an XOR game, this predicate depends only on the XOR $c = a \oplus b$ of Alice's answer a and Bob's answer b. We thus write $V(c|s,t) = 1$ if and only if answers $a \oplus b$ satisfying $a \oplus b = c$ are winning answers for questions s and t. We will also restrict ourselves to unique games, which have the property that for any s,t,b, there exists exactly one winning answer a for Alice (and similarly for Bob).

First of all, note that in the quantum case we may write the probability that Alice and Bob return answers a and b with $a \oplus b = c$ as

$$p(c|s,t) = \frac{1}{2}(1 + (-1)^c \langle\Psi| A_s \otimes B_t |\Psi\rangle),$$

where we again use A_s and B_t to denote Alice's and Bob's observable corresponding to questions s and t respectively and $|\Psi\rangle$ denotes the maximally entangled state. Note that we again have $(A_s)^2 = (B_t)^2 = \mathbb{I}$ from the fact that both measurements have only two outcomes. The probability that Alice and Bob win the game can then be written as

$$\sum_{s,t} \pi(s,t) \sum_c V(c|s,t) p(c|s,t).$$

Let $v_{st} = \langle\Psi| A_s \otimes B_t |\Psi\rangle$. First of all note that for $p \rightarrow \infty$

$$\frac{1}{d}\left(\mathbb{I} + \sum_{st} v_{st} \Gamma_s \otimes \Gamma_t\right) \tag{2.7}$$

with $d = 2^{\max|S|,|T|}$ and Γ_s, Γ_t anti-commuting observables as defined in Section 2.2 is a valid state for any $|v_{st}| \leq 1$. Hence, we can immediately see that

Corollary 2.5.2. *In any ∞-theory, there exists a strategy for Alice and Bob to win a unique XOR game with certainty.*

Proof. Consider the state given in Eq. (2.7) with $v_{st} = \pm 1$ such that $p(c|s,t) = 1$ whenever $V(c|s,t) = 1$. Let Alice and Bob's measurements be given by Γ_s and Γ_t for questions s and t respectively, which are valid measurements for all p-theories with Γ_s, Γ_t constructed as in Section 2.2. □

We leave it as an open question to examine the case of $p < \infty$ for XOR games, since our aim was merely to show that superstrong correlations can exist, if we allow for relaxed uncertainty relations. We can see that letting $v_{st} = \pm 1/(\max|S|,|T|)^{1/p}$ makes Eq. (2.7) a valid state for any choice of p, but this may not generally be the optimal choice. The case of GNST is similar, and it has been shown that any nonlocal correlations can (approximately) be simulated by such boxes [28]. Optimal bounds for p-GNST with $p < \infty$ can be obtained using techniques analogous to [21].

2.6 Superstrong random access encodings

The existence of superstrong nonlocal correlations is by no means the only difference we can observe when moving from quantum theory to p-GNST or p-nonlocal theories. In particular, we now show that we can obtain so-called random access encodings which, depending on the theory, can be exponentially better than those realized by quantum mechanics. We then investigate how uncertainty relations and the restrictions imposed by simultaneous measurements affect this encoding. The existence of such random access encodings will play a crucial role when considering the power of p-GNST theories for communication complexity in Section 2.7.1. In Section 2.7.2 we also use this random access code to prove a lower bound on the sample complexity of learning states in GNST.

2.6.1 In p-GNST

Intuitively, a random access code [52, 19] allows us to encode N bits into a physical system of size n such that we can decode any one bit of the original string with probability at least q. More formally,

Definition 2.6.1. *A $[N, n, q]$-random access code (RAC) is an encoding of a string $x \in \{0,1\}^N$ into an n-gbit state p_x, such that there exist measurements $C \in \mathcal{E}_G$ with outcomes $A \in \mathcal{A}^{\times n}$, and a decoding algorithm $D : \mathcal{A}^{\times n} \to \{0,1\}$ satisfying*

$$Pr(D(A) = x_k) = \sum_{A \in \mathcal{A}^{\times n}} \delta_{D(A),x_k} p_x(A|C) \geq q,$$

where $p_x(A|C)$ is the probability of obtaining outcome A when performing the measurement C.

It has been shown [52, 19] that in the quantum case, we must have $n \geq (1 - h(q))N$, where h denotes the binary entropy function. There also exist classical encodings for which $n = (1 - h(q))N + O(\log N)$ [52]. Hence, quantum states offer at most a modest advantage over classical mechanics and, for $q = 1$, no advantage at all. We now proceed to the surprising result that general no-signaling states lead to extremely powerful random access codes.

Claim 2.6.2. *In GNST, there exists a* $[3^n, n, 1]$*-random access code.*

Proof. An n gbit state in GNST is completely characterized by the probabilities of outcomes for a fixed set of measurements. Recall that a single gbit is a two-level system on which we allow three possible measurements with two possible outcomes each. Also recall that each $C \in \mathcal{E}_G$ can be represented as $\mathcal{E}_G = \{\forall k \in \{1,2,3\}^n : \{W_{1,k_1}, \ldots, W_{n,k_n}\}\}$, with $W_{i,1} = X_i, W_{i,2} = Z_i, W_{i,3} = X_i Z_i$. Note that each measurement C is associated with one of $N = 3^n$ vectors $k = (k_1, \ldots, k_n)$. Let $f : C \to \{1, \ldots, N\}$ be a one-to-one function. For each of the $N = 3^n$ bits we wish to encode, we must specify one measurement C that we can use to extract the jth-bit. Let that measurement be denoted by $f^{-1}(j)$.

We are now ready to define our encoding of the string $x \in \{0,1,2\}^N$ into an n-gbit GNST state p_x via the probabilities

$$p_x(A|C) := \frac{1}{2^n}(1 + A^*(-1)^{x_{f(C)}}),$$

where we use the previously defined notation $A^* = \prod_{i=1}^{|C|} A_i$. It is straightforward to verify that the state is normalized, positive, and satisfies the no-signaling condition.

We now show that any bit of the original string can be decoded perfectly. If we choose to retrieve bit j, we measure $C = f^{-1}(j)$. That means that we get result A with probability $\frac{1}{2^n}(1+A^*(-1)^{x_j}) = \frac{1}{2^n}2\delta_{A^*,(-1)^{x_j}}$. And we get the result $A^* = (-1)^{x_j}$ with probability:

$$\sum_{A^*=(-1)^{x_j}} p_x(A|C) = \sum_{A^*=(-1)^{x_j}} \frac{1}{2^n}2\delta_{A^*,(-1)^{x_j}} = 1.$$

where the last equality follows from the fact that we sum over exactly half the 2^n possible outcomes $A_1, \ldots, A_n$. Hence the decoder $D(A) = \frac{1}{2}(1 - A^*)$ will return x_j with perfect probability. □

What happens if we impose the uncertainty relation in p-GNST? For convenience sake, note that we could rewrite the encoding above in terms of moments, where we let an encoding of a string x be determined by the moment representation of p_x as

$$m_x(C = f^{-1}(k)) := (-1)^{x_k}$$

with all other moments set to 0.

To construct an encoding for p-GNST, we consider

$$m_x(C = f^{-1}(k)) := (-1)^{x_k}\lambda.$$

What's the largest λ that satisfies the uncertainty relation? As we noted earlier the maximum number of anti-commuting Pauli operators is $2n+1$, so the most restrictive condition we could get from the uncertainty relation is $(2n+1)|\lambda|^p \leq 1$. We thus obtain

Claim 2.6.3. *In p-GNST, there exists a $[3^n, n, \frac{1}{2} + \frac{1}{2}\left(\frac{1}{2n+1}\right)^{1/p}]$-random access code.*

Proof. Let $\lambda = (2n+1)^{1/p}$, and note that this satisfies the uncertainty relation. Our encoding is now

$$p_x(A|C) = \frac{1}{2^n}(1 + (-1)^{x_{f(C)}}\ \lambda\ A^*).$$

And our probability of getting the correct sign from our measurement goes down to

$$\Pr(D(A) = x_k) = \frac{1+|\lambda|}{2} = \frac{1}{2} + \frac{1}{2}\left(\frac{1}{2n+1}\right)^{1/p}.$$

□

If $p < \infty$ we get an encoding that gets asymptotically worse for large n. This should be compared to the bound on the number of qubits for a *quantum* random access encoding of $N = 3^n$ bits into k qubits with recovery probability $q = 1/2 + 1/2(1/(2n+1))^{1/p}$. From the bound of [52, 19], we have that the encoding uses exponentially fewer physical bits than what can be obtained in the quantum setting and hence even p-GNST has a powerful coding advantage over quantum mechanics. Note that we are always free to split the N bits into smaller pieces first, and encode each piece independently to keep the recovery probability q constant. This is analogous to the quantum setting where we can encode each 3 bits into one qubit to obtain a random access code with $n = N/3$. Alternatively, we can form a simple repetition code, where we have k copies of the random access codes constructed above. We then have

Claim 2.6.4. *In p-GNST, there exists a $[3^n, (2n+1)^{3/p}n, 1-\varepsilon]$-random access code with $\varepsilon = 2\exp(-(2n+1)^{1/p}/2)$.*

Proof. We take k copies of the RAC defined in Claim 2.6.3, and decode by taking the majority of the individual encodings. Let $Y_j = 1$ if the decoding was successful for the j-th copy, and $Y_j = 0$ otherwise. From the Hoeffding inequality we immediately obtain that for $Y = \sum_{j=1}^{k} Y_j$ and q as defined above

$$\Pr\left[|Y - qk| \geq t\ k\right] \leq 2e^{-2t^2k}.$$

If we set $t = q - 1/2 = 1/2(1/(2n+1))^{1/p}$, that gives us $\Pr\left[Y \leq k/2\right] \leq 2e^{-\frac{1}{2}\left(\frac{1}{2n+1}\right)^{2/p}k}$. Now if we set $k = (2n+1)^{3/p}$, we have used a total of $(2n+1)^{3/p}n$ gbits and will succeed with probability $1 - 2e^{-(2n+1)^{1/p}/2}$ as promised. □

Whereas $(2n+1)^{3/p}n$ is still quite large, note that it is nevertheless only polynomial in n. The length of the RAC is hence still poly-logarithmic in our original input size, where we achieve (near) perfect recovery for large n. Finally, we will need to use one more related result.

Claim 2.6.5. *In p-GNST, for $\gamma \in (0, 1/2)$ and $\hat{n} \geq 2^{2/p}\ln(4/(1/2-\gamma)^2)$, there exists a $[3^{n(\hat{n},p,\gamma)}, \hat{n}, \frac{1}{2}+\gamma]$-random access code with $n(\hat{n}, p, \gamma) = \lfloor\left(\frac{\hat{n}\ 2^{-2/p}}{\ln(4/(1/2-\gamma)^2)}\right)^{\frac{1}{2/p+1}}\rfloor$.*

Proof. Again we take k copies of the RAC defined in Claim 2.6.3, and decode by taking the majority of the individual encodings. The probability to decode correctly in that case was $1 - 2e^{-\frac{1}{2}(\frac{1}{2n+1})^{2/p}k}$. Now we want to adjust k and n to get a code with a fixed success rate and that uses no more than $\hat{n}$ gbits. We need that (i) $kn \leq \hat{n}$, that is, our encoding uses at most $\hat{n}$ physical bits and (ii) $1 - 2e^{-\frac{1}{2}(\frac{1}{2n+1})^{2/p}k} \geq 1/2 + \gamma$, which forces our probability of success to be at least $1/2 + \gamma$. We can satisfy (ii) if we set $k = \ln(4/(1/2-\gamma)^2)(2n+1)^{2/p}$, then (i) tells us that $kn = \ln(4/(1/2-\gamma)^2)(2n+1)^{2/p}n$, from which we have $\ln(4/(1/2-\gamma)^2)2^{2/p}n^{2/p+1} \leq kn \leq \hat{n}$ and thus

$$n \leq \left(\frac{\hat{n}\ 2^{-2/p}}{\ln(4/(1/2-\gamma)^2)}\right)^{\frac{1}{2/p+1}}.$$

Since the smallest system we can encode into is $n = 1$, this tells us that $\hat{n}$ must be at least $2^{2/p}\ln(4/(1/2-\gamma)^2)$. □

Note that although this may not be the best encoding, it suffices to give us the asymptotic behavior for $\hat{n}$.

2.6.2 In p-nonlocal theories

It is instructive to consider such superstrong encodings in the language of p-nonlocal theories to see how such superstrong encodings would look like in terms of Pauli matrices. This will also allow us to compare the consequences of restrictions due to the consistencies of moments from Section 2.2.3 to random access encodings. For the least restrictive p-theory, the p-bin theory, we can construct the following very simple encoding.

Claim 2.6.6. *In p-bin theories, there exists a $[2^{2n}-1, n, \frac{1}{2}+\frac{1}{2}\left(\frac{1}{2n+1}\right)^{1/p}]$-random access code.*

Proof. Consider the encoding of a string $x \in \{0,1\}^N$ with $N = 2^{2n}-1$ into an n p-bit state given by

$$\rho_x := \frac{1}{d}\left(\mathbb{I} + \frac{1}{(2n+1)^{1/p}}\sum_{k=1}^{2^{2n}-1}(-1)_k^x S_k\right),$$

where $S_k = S_{ab}$ is a string of Pauli matrices, where we simply relabeled the indices ab. To decode the kth-bit, we measure S_k. A straightforward calculation shows that the probability to obtain outcome x_k is given by

$$\Pr[x_k] = \frac{1}{2}\operatorname{Tr}\left[(\mathbb{I} + S_k)\,\rho_x\right] = \frac{1}{2} + \frac{1}{2(2n+1)^{1/p}},$$

as promised. Clearly, the uncertainty relation is satisfied. □

Similarly, we obtain the following encoding for p-box theories, which is in one-to-one correspondence with the encodings in p-GNST above.

Claim 2.6.7. *In p-box theories, there exists a* $[3^n, n, \frac{1}{2} + \frac{1}{2}\left(\frac{1}{2n+1}\right)^{1/p}]$*-random access code.*

Proof. Our encoding is analogous to the one above, but we restrict ourselves to including only such strings of Pauli matrices formed by taking tensor products of $\{X, Y, Z\}$, excluding the identity. □

Clearly, we can again obtain an encoding that is poly-logarithmic in the length of the original input analogous to Claim 2.6.4 that has perfect recovery for large n.

2.6.3 The effect of consistency

When viewing such encodings in terms of density matrices, it becomes clear why such encodings do not exist in a quantum setting: all such encodings are in gross violation of the consistency conditions of Section 2.2.3. Even when we restrict ourselves to $p = 2$, we can obtain such encodings whereas in the quantum case we cannot. It is interesting to note that for $p = 2$, the violation we can obtain for, e.g., the CHSH game is exactly the same as in the quantum setting. Thus it is perfectly possible to have such superstrong encodings, while simultaneously being restricted to Tsirelson's bound in the CHSH game for a 2 qubit state. This clearly shows how limited our p-bin, p-nonlocal, but also p-GNST theories really are. Since GNST is equivalent to a theory based on nonlocal boxes, this also shows that considering such boxes is somewhat limiting, and possibly ignores some aspect present in quantum theory that are of importance for information processing.

2.7 Implications for information processing

We now turn to a number of interesting implications of p-GNST and p-theories to information processing. In particular, we will see that both allow us to save significantly on the amount of data we need to transmit to solve certain communication problems. In fact, we will see that there exists a task for which there exists an *exponential* gap between the amount of communication required when compared with quantum theory. Other information tasks on the other hand become more difficult. We will see that when trying to learn states approximately we need to perform exponentially more measurements in the case of GNST.

2.7.1 Communication complexity

Imagine two (or more) parties, Alice and Bob, who each have an input $x \in \{0,1\}^n$ and $y \in \{0,1\}^n$ respectively, unknown to the other party. Their goal is to compute a fixed function $f : \{0,1\}^{2n} \to \{0,1\}^m$ by communicating over a channel. The central question of communication complexity is how many bits they need to transmit in order to compute f. Typically, we thereby only require one party (Bob) to learn the result $f(x,y)$. To help them reduce the amount of communication, Alice and Bob may possess additional resources such as shared randomness, entanglement, nonlocal boxes, or communicate over a quantum channel, and may impose different measures of success. For example, they could be interested in computing f only with a certain probability instead of computing it exactly. It is well-known that if Alice and Bob can share nonlocal boxes, they can compute any Boolean function $f : \{0,1\}^{2n} \to \{0,1\}$ perfectly by communicating only a single bit [5], which is even true when the nonlocal boxes have slight imperfections [27]. Here, we consider the case where Alice and Bob have *no* a priori resources, however, we they are able to exchange p-GNST or p-nonlocal states over a suitable channel.

2.7.1.1 One-way communication

We first of all make a very modest statement and show that in *any* one-way communication protocol, where Alice sends a single message to Bob, we are able to save a constant number of bits, when computing a Boolean function f. These savings are an immediate consequence of the existence of superstrong random access codes that we discussed in Section 2.6. To communicate with Bob, Alice constructs the string

$$m = f(x,0), \ldots, f(x, 2^n - 1)$$

and encodes $m \in \{0,1\}^{2^n}$ into a random access code ρ_m. To retrieve the correct answer, Bob simply retrieves bit $x_y = f(x,y)$ from ρ_m. Evidently, this type of saving is particularly interesting in the case where Alice and Bob would need to communicate n bits to compute f, which is the case classically and quantumly if $f = IP$ is the inner product [53]. By Claims 2.6.2, 2.6.3, 2.6.7, and 2.6.6 we immediately obtain that

Claim 2.7.1. *Let $p \to \infty$. Then in to compute the inner product Alice needs to transmit at most k bits to Bob, where*

$$k = \begin{cases} (1/\log 3)n & \text{for } p\text{-GNST and } p\text{-nonlocal theories} \\ n/2 & \text{for } p\text{-bin theory} \end{cases}$$

2.7.1.2 Private information retrieval

More striking though are the possibilities of p-GNST or p-theories for the task of private information retrieval: Here, one (or more) database servers each hold a copy of the database string $x \in \{0,1\}^n$. A database user should be able to retrieve any bit x_i of his choosing, while the servers should not learn the desired index i. A protocol that satisfies these parameters is the trivial one, where the server simply sends the entire string x to the user. The question is thus, whether it is possible to perform this task by communicating less than n bits. If only a single server is used, it is known that the trivial protocol is optimal and we need to communicate $\Theta(n)$ bits, even if we are allowed quantum communication [54]. It is clear that the superstrong encodings from above, allow us to beat this bound trivially, by asking the server to encode x into a superstrong random access code. Hence we have as an immediate consequence of Claims 2.6.2, 2.6.4, 2.6.6, and 2.6.7 we have

Claim 2.7.2. *In any p-GNST, p-bin, and p-box theory, there exists a single server private information retrieval scheme requiring $O(polylog(n))$ bits of communication for large n.*

2.7.2 Learnability

We consider a scenario in which there is an unknown state for which we are trying to learn an approximate description. In particular, imagine some arbitrary probability distribution over possible two-outcome measurements. We are given the expectation value for each measurement in a finite set picked according to this distribution. We then construct an approximate description of the state which agrees with all the expectation values we have observed so far. This description is considered to be good if it predicts the correct results for most future measurements drawn from the same distribution. The central question is how many measurement results we need to be able construct a good description.

The existence of strong random access codes has implications for state learning. Aaronson [20] used an upper bound on the number of bits that can be encoded into an n qubit RAC to upper bound the number of measurements needed to learn an approximate description of an n qubit state. He took solace in the fact that, despite the exponential number of parameters describing a quantum state, a linear (in the number of qubits) number of measurements suffice to learn an approximate description of the state. If an exponential number of measurements were really required, we could never hope to do enough measurements to verify the identity of quantum states of a few hundred particles.

We show the converse for states in p-GNST. We use our constructions of random access codes to lower bound the number of measurements needed to learn an approximate description of the state. We find that an exponential number of measurements is required to find such a description and therefore one could never hope to do enough measurements to learn a description of a state with a modest number of particles, even approximately. This holds even for theories where $p = 2$ and the violation of the CHSH inequality is the same as for quantum mechanics. This demonstrates an unusually powerful theory which starkly contrasts with quantum mechanics and the p-nonlocal theory.

We begin with a section defining the relevant tools: a definition of the learning scenario, and a measure of state complexity known as the “fat shattering dimension.” We then restate a known lower bound on the number of samples needed for learning in terms of the fat shattering dimension. In the next section, we derive lower bounds on learnability for p-GNST theories. First, we use our random access codes to lower bound the fat shattering dimension for p-GNST states. Then we can use this result to lower bound the number of samples needed to learn p-GNST states.

2.7.3 Tools

We begin by introducing some terminology from statistical learning theory. Let the set $\mathcal{S}$ denote the sample space, which will correspond to the space of possible measurements in our case. A probabilistic concept over $\mathcal{S}$ is just a function $F : \mathcal{S} \to [0, 1]$, and is equivalent to a state which maps measurement choices to expectation values. A set of such concepts is referred to as the concept class $\mathcal{C}$ over $\mathcal{S}$ and corresponds to the set of all states. We consider the learning situation in which you are given the value of the target concept (state) over some samples drawn independently according to an arbitrary distribution. The goal is to output a hypothesis concept that will give values close to the target concept for most samples drawn from the same distribution. A sample size that is large enough to allow this to be accomplished with high probability is said to be sufficient. To restate the connection, in GNSTs we will say that a state corresponds to a concept, and a measurement on the state to a sample. We will make these notions precise before we demonstrate the connection between RACs and fat-shattering dimension in 2.7.5.

We adopt our definition of probabilistic concept learning from Anthony and Bartlett[55].

Definition 2.7.3 (Anthony and Bartlett [55])**.** *Let $\mathcal{S}$ be a sample space, let $\mathcal{C}$ be a probabilistic concept class over $\mathcal{S}$, and let $\mathcal{D}$ be a probability measure over $\mathcal{S}$. Fix an element $\rho \in \mathcal{C}$, as well as error parameters $\varepsilon, \eta, \gamma > 0$ with $\gamma > \eta$. Let $k_0(\eta, \gamma, \epsilon, \delta)$ be some function of the error parameters. Suppose we draw a training set of k samples $\mathcal{T} = (s_1, \ldots, s_k)$ independently according to $\mathcal{D}$, and then choose any hypothesis $\sigma_{\mathcal{T}} \in \mathcal{C}$ such that $|\sigma_{\mathcal{T}}(s_i) - \rho(s_i)| \leq \eta$ for all $s_i \in \mathcal{S}$. Then if for $k \geq k_0(\eta, \gamma, \epsilon, \delta)$*

$$\Pr_{s \in \mathcal{S}} \left[|\sigma_{\mathcal{T}}(s) - \rho(s)| > \gamma \right] \leq \varepsilon$$

with probability at least $1 - \delta$ over $\mathcal{T}$, we say that k_0 is a sufficient sample size to learn $\mathcal{C}$.

This says that if the size of the training set, k, is bigger than k_0, then with probability $1 - \delta$, the training set $\mathcal{T}$, that we pick according to $\mathcal{D}$ will be a good training set. That is, a hypothesis concept σ which matches the target state on the training set will only be different from the target state on some other sample with small probability, ϵ.

To define a lower bound on k_0, we will need a measure of complexity called the *fat-shattering dimension.*

Definition 2.7.4 (Aaronson [20]). *Let $\mathcal{S}$ be a sample space, let $\mathcal{C}$ be a probabilistic-concept class over $\mathcal{S}$, and let $\gamma > 0$ be a real number. We say a set $\{s_1, \ldots, s_k\} \subseteq \mathcal{S}$ is γ-fat-shattered by $\mathcal{C}$ if there exist real numbers $\alpha_1, \ldots, \alpha_k$ such that for all $B \subseteq \{1, \ldots, k\}$, there exists a probabilistic concept $\rho \in \mathcal{C}$ such that for all $i \in \{1, \ldots, k\}$,*

(i) if $i \notin B$ then $\rho(s_i) \leq \alpha_i - \gamma$, and

(ii) if $i \in B$ then $\rho(s_i) \geq \alpha_i + \gamma$.

Then the γ-fat-shattering dimension of $\mathcal{C}$, or $\mathrm{fat}_{\mathcal{C}}(\gamma)$, is the maximum k such that some $\{s_1, \ldots, s_k\} \subseteq \mathcal{S}$ is γ-fat-shattered by $\mathcal{C}$. (If there is no finite such maximum, then $\mathrm{fat}_{\mathcal{C}}(\gamma) = \infty$.)

The fat-shattering dimension lower bounds the number of samples needed to learn a probabilistic concept.

Lemma 2.7.5 (Anthony and Bartlett [55]). *Suppose $\mathcal{C}$ is a probabilistic concept class over $\mathcal{S}$ and set $0 < \gamma < \eta < 1, \epsilon, \delta \in (0,1)$. Then if $\mathrm{fat}_{\mathcal{C}}(\gamma) \geq d \geq 1$ and $\gamma^2 \geq 4d2^{-\sqrt{d/6}}$, any sample size m_0 sufficient to learn $\mathcal{C}$ satisfies*

$$m_0(\eta, \gamma, \epsilon, \delta) \geq max\left(\frac{1}{32\epsilon}\left(\frac{d}{2\ln^2(4d/\gamma^2)} - 1\right), \frac{1}{\epsilon}\ln\frac{1}{\delta}\right)$$

This concludes the results we will need from statistical learning theory.

2.7.4 Lower bounds on sample complexity

Our next step is to show that the existence of random access codes lower bounds the fat-shattering dimension. First we have to carefully define what "concept" we will be learning and what constitutes our sample space. For the purposes of learning in GNSTs, the sample space is just the set of possible measurements, where we allow general measurements by first making some fiducial measurement on the state, and then post-processing the result using some decoding function. So we can define $\mathcal{S}_{\mathrm{GNST}} := \{(C, D) | C \in \mathcal{E}_G, D : \mathcal{A}^{\times n} \to \{0,1\}\}$. For some sample $(C, D) \in \mathcal{S}_{GNST}$, a concept is specified by the state ρ_x in a GNST via the the probability $\rho_x(C, D) := \sum_{A \in \mathcal{A}^{\times n}} D(A) p_x(A|C)$, where p_x is an n-partite state in some GNST. Then the concept class $\mathcal{C}_{GNST}$ is the set of concepts specified by all the states in GNST.

Note that a "sample" is stronger than a typical notion of measurement. Usually we say that the measurement gives a result with some probability, but given some sample, the concept ρ actually returns the probability of that outcome occurring. This stronger notion of sampling is all we consider here since we are only lower bounding the number of samples needed.

Claim 2.7.6. *Let the concept class $\mathcal{C}_{GNST}$ over $\mathcal{S}_{GNST}$ consist of all $\rho_x(C,D) = \sum_{A\in\mathcal{A}} D(A)p_x(A|C)$, where p_x describes any n-partite states in a GNST, over the sample space $\{(C,D)|C \in \mathcal{E}_G, D : \mathcal{A}^n \rightarrow \{0,1\}\}$. For integers $n, N(p,n)$ and $\gamma \in (0,1)$, if there exists an $[N(p,n), n, \frac{1}{2}+\gamma]$-RAC then* $\mathrm{fat}_{\mathcal{C}_{GNST}}(\gamma) \geq N$.

Proof. By the RAC definition, there exist a set of measurements $\{(C,D),\ldots,(C^{(N)},D^{(N)})\}$ and states specified by (the concepts) ρ_x for $x \in \{0,1\}^N$ so that

(i) if $x_i = 0$ then $\rho_x(C^{(i)},D^{(i)}) \leq \frac{1}{2} - \gamma$

(ii) if $x_i = 1$ then $\rho_x(C^{(i)},D^{(i)}) \geq \frac{1}{2} + \gamma$.

Therefore, this set of samples is γ fat-shattered by $\mathcal{C}_{\mathrm{G}NST}$. Since $\mathrm{fat}_{\mathcal{C}_{GNST}}$ is the size of the largest sample set shattered, $\mathrm{fat}_{\mathcal{C}_{GNST}} \geq N(p,n)$. □

Combining Claim 2.6.5 with 2.7.6 and 2.7.5, we get the following result.

Corollary 2.7.7. *For $\hat{n}$-partite concepts in $\mathcal{C}_{p-GNST}$ and error parameters $\varepsilon, \eta, \gamma, \delta > 0$ with $\gamma > \eta$, if $\hat{n} \geq 2^{2/p}\ln(4/(1-\gamma)^2)$ and*

$$k < max\left(\frac{1}{32\epsilon}\left(\frac{3^{n(\hat{n},p,\gamma)}}{2\ln^2(4\cdot 3^{n(\hat{n},p,\gamma)}/\gamma^2)} - 1\right), \frac{1}{\epsilon}\ln\frac{1}{\delta}\right)$$

for $n(\hat{n},p,\gamma) = \lfloor\left(\frac{\hat{n}\ 2^{-2/p}}{\ln(4/(1-\gamma)^2)}\right)^{\frac{1}{2/p+1}}\rfloor$, then k is not a sufficient sample size to learn states in $\mathcal{C}_{p-GNST}$.

That is, we need $\mathrm{O}(3^{\hat{n}^{\frac{1}{2/p+1}}}/\hat{n}^{\frac{2}{2/p+1}})$ samples to learn an $\hat{n}$-partite state in p-GNST to great accuracy. For $p = 2$ we have an uncertainty relation analogous to quantum mechanics that rules out super-quantum violations of the CHSH bound. Nevertheless it still takes $\mathrm{O}(3^{\sqrt{\hat{n}}}/\hat{n})$ samples to learn these states, as compared to $\mathrm{O}(n)$ in the quantum case.

2.8 Conclusion and open questions

We have shown that relaxing uncertainty relations can lead to superstrong nonlocal correlations. This is quite intuitive when considering Tsirelson's bound as a consequence of such an uncertainty relation in the quantum setting. We then constructed a range of theories inspired by such relaxations, and investigated their power with respect to a number of information processing problems. In particular, we obtained superstrong random access encodings and savings for communication complexity. At the same time, however, it turned out to become harder to learn a state in such a theory. We then

	p-bin	p-GNST/p-box	p-nonlocal	Quantum	Classical
Non-signaling	yes	yes	yes	yes	yes
Satisfies p-uncertainty	yes	yes	yes	p=2	n/a
Simultaneous measurements	no	local	commuting	commuting	all
CHSH violation	$\frac{1}{2}+\frac{1}{2^{1/p+1}}$	$\frac{1}{2}+\frac{1}{2^{1/p+1}}$	$\frac{1}{2}+\frac{1}{2^{1/p+1}}$	$\frac{1}{2}+\frac{1}{2^{1/2+1}}$	$\frac{3}{4}$
RAC bits to encode N bits	O(polylog(N))	O(polylog(N))	?	$\Omega(N)$	$\Omega(N)$
PIR from N bits	O(polylog(N))	O(polylog(N))	?	$\Omega(N)$	$\Omega(N)$
"Learning" states	hard	hard	?	easy	easy

Table 2.1: Summary of properties and results for various theories.

discussed what makes such superstrong encodings possible in our p-theories, but also in GNST. We identified a number of simple constraints that prevent us from constructing a similar encoding in the quantum setting. Our work may indicate that using "box-world" to understand any other problems within quantum information beyond nonlocal correlations may be difficult, as "box-world" differs from the quantum setting with respect to such constraints, at least when drawing a one-to-one analogy from a gbit to a qubit as in GNST [18]. It is important to note that these constraints did not prevent us from observing superstrong nonlocal correlations, but merely forbid our encodings in Section 2.6. If one would like to use "box-world" to understand other aspects one could either impose such consistency constraints, or look for a different approach to defining such theories. GNST was defined by first specifying states and then allowing all operations that take valid states to valid states. If one would have specified the theory in terms of allowed transformations, instead of states, such encodings could also have been ruled out. For example, in the quantum setting one can transform operators $X \otimes X$, $Z \otimes Z$, and $XZ \otimes XZ$ into a bipartite form via a unitary operation. When looking at a density matrix expressed in terms of strings of Pauli matrices, its coefficients (which directly determine the moments for measurements of strings of Paulis) must obey similar constraints to the coefficients belonging to bipartite operators of the form $\mathbb{I} \otimes X, X \otimes \mathbb{I}, X \otimes X$ for example.

Finally, it is clear that both the uncertainty relation and the consistency constraints are obeyed in the quantum setting, since we demand that for any ρ we have $\mathrm{Tr}(\rho) = 1$ and $\rho \geq 0$ to be a valid quantum state. Not surprisingly, both forms of constraints are thus necessary (but in higher dimensions not always sufficient) conditions for $\rho \geq 0$. Such characterizations are not easy for $d > 2$ [43, 44, 45, 46], and it remains an interesting open problem to find an intuitive interpretation for such conditions in higher dimensions, and their consequence for information processing tasks.

2.A Appendix: Monogamy from uncertainty relations

In this appendix we provide a sketch of a proof that monogamy constraints can be derived directly from generalized uncertainty relations.

Consider a setting in which we have three parties, A,B, and C. They have access to part of some shared state ρ, and each party has a choice of two measurements, e.g. A_1, A_2, with outcome $\{\pm 1\}$. For two parties we can define the CHSH operator $\beta_{AB} = A_1 \otimes B_1 \otimes \mathbb{I} + A_2 \otimes B_1 \otimes \mathbb{I} + A_1 \otimes B_2 \otimes \mathbb{I} - A_2 \otimes B_2 \otimes \mathbb{I}$ and its expectation value $\langle \beta_{AB} \rangle = \mathrm{Tr}(\rho \beta_{AB})$. Monogamy relations quantify the trade-off between the maximum CHSH value attained by A and B, versus the the maximum value that can be simultaneously attained by B and C. The best known result quantifying this trade-off is[36]

$$\langle \beta_{AB} \rangle^2 + \langle \beta_{BC} \rangle^2 \leq 8 \tag{2.8}$$

We start by assuming that there exists locally anticommuting observables that achieve the maximal value of the monogamy relation. Although this is true in all known cases, we are not aware of a proof of this fact. Given this assumption, we show that for locally anti-commuting observables, the generalized uncertainty relations directly imply Eq. (2.8).

Now, we assume that each party's local measurements anti-commute, i.e. $\{A_1, A_2\} = 0$. The only constraint we will impose is the generalized uncertainty relation. That is, for measurements M_i, so that $\{M_i, M_j\} = \delta_{ij}$,

$$\sum_i \langle M_i \rangle^2 \leq 1$$

For instance, $\langle \mathbb{I} \otimes B_1 \otimes C_1 \rangle^2 + \langle \mathbb{I} \otimes B_1 \otimes C_2 \rangle^2 + \langle A_1 \otimes B_2 \otimes \mathbb{I} \rangle^2 + \langle A_2 \otimes B_2 \otimes \mathbb{I} \rangle^2 \leq 1$.

Now we construct the vector of variables

$$\begin{aligned} x &= (x_1, \ldots, x_{12}) \\ &:= (\langle \mathbb{I} \otimes B_1 \otimes C_1 \rangle, \langle \mathbb{I} \otimes B_1 \otimes C_2 \rangle, \langle \mathbb{I} \otimes B_2 \otimes C_1 \rangle, \langle \mathbb{I} \otimes B_2 \otimes C_2 \rangle, \langle A_1 \otimes \mathbb{I} \otimes C_1 \rangle, \langle A_1 \otimes \mathbb{I} \otimes C_2 \rangle, \\ &\quad \langle A_1 \otimes B_1 \otimes \mathbb{I} \rangle, \langle A_1 \otimes B_2 \otimes \mathbb{I} \rangle, \langle A_2 \otimes \mathbb{I} \otimes C_1 \rangle, \langle A_2 \otimes \mathbb{I} \otimes C_2 \rangle, \langle A_2 \otimes B_1 \otimes \mathbb{I} \rangle, \langle A_2 \otimes B_2 \otimes \mathbb{I} \rangle) \end{aligned}$$

We rewrite the monogamy relation in terms of these variables.

$$q(x) = 8 - (\langle \beta_{AB} \rangle^2 + \langle \beta_{BC} \rangle^2) = 8 - (x_7 + x_8 + x_{11} - x_{12})^2 - (x_1 + x_2 + x_3 - x_4)^2 \geq 0$$

Under the constraints implied by the uncertainty relation, which we label as $g_i(x) \geq 0$, we want to show $q(x)$ is always non-negative.

We begin by noting that the following two inequalities are a result of the generalized uncertainty relation.

$$\begin{aligned} g_1(x) &= 1 - x_1^2 - x_2^2 - x_8^2 - x_{12}^2 \geq 0 \\ g_2(x) &= 1 - x_3^2 - x_4^2 - x_7^2 - x_{11}^2 \geq 0 \end{aligned}$$

We will also need the following six polynomials which can be obtained by using semi-definite programming in combination with the real Positivstellensatz.[56]

$$\begin{aligned} f_1(x) &= \sqrt{2}(x_{11} + x_{12}) \\ f_2(x) &= \sqrt{2}(x_2 - x_1) \\ f_3(x) &= \sqrt{8/3}x_3 - \sqrt{2/3}(x_1 + x_2) \\ f_4(x) &= \sqrt{1/3}(x_1 + x_2 + x_3) + \sqrt{3}x_4 \\ f_5(x) &= \sqrt{8/3}x_7 - \sqrt{2/3}(x_{11} - x_{12}) \\ f_6(x) &= -\sqrt{1/3}(x_{11} - x_{12}) - \sqrt{1/3}x_7 + \sqrt{3}x_8 \end{aligned}$$

Now one can verify that $8 - \langle \beta_{BC} \rangle^2 + \langle \beta_{AB} \rangle^2 = q(x)$ can be written as the sum of positive terms $g_1(x), g_2(x)$ and squares of polynomials like $f_i(x)$.

$$q(x) = 4(g_1(x) + g_2(x)) + \sum_{i=1}^{6} f_i(x)^2 \geq 0$$

Therefore, $\langle \beta_{BC} \rangle^2 + \langle \beta_{AB} \rangle^2 \leq 8$. To show that this bound is tight, one needs only produce a point that achieves equality. For this case,

$$x_1 = x_2 = x_3 = -x_4 = \frac{1}{\sqrt{2}} \quad x_{i>4} = 0.$$

Chapter 3

Tests of nonlocality

> *...In that Empire, the craft of Cartography attained such Perfection that the Map of a Single province covered the space of an entire City, and the Map of the Empire itself an entire Province. In the course of Time, these Extensive maps were found somehow wanting, and so the College of Cartographers evolved a Map of the Empire that was of the same Scale as the Empire and that coincided with it point for point. Less attentive to the Study of Cartography, succeeding Generations came to judge a map of such Magnitude cumbersome, and, not without Irreverence, they abandoned it to the Rigours of sun and Rain. In the western Deserts, tattered Fragments of the Map are still to be found, Sheltering an occasional Beast or beggar; in the whole Nation, no other relic is left of the Discipline of Geography.*
>
> Jorge Luis Borges

In this section we argue that an exact description of the boundary between local and nonlocal correlations would be cumbersome. Instead we propose an approximate but exponentially smaller description of this boundary. We will first cast the problem of testing for nonlocal correlations in terms of the problem of finding the convex hull of a set of vertices. We demonstrate that methods involving linear matrix inequalities produce nonlinear bounds that efficiently approximate the convex hull, yielding sufficient tests for violations of locality. After comparing the efficacy of these methods for detecting nonlocal correlations, we discuss their application to Bayesian graphs and general machine learning problems.

3.1 The Bell polytope

We begin by recalling the experimental setup described in Chapter 1. There we had p parties where each i-th party had a choice of $X_i = 1, \ldots, s$ measurement settings with the possibility of finding one of $A_i = 1, \ldots, r$ outcomes or results. The results of repeating these experiments many times can be described by the $d = (rs)^p$ probabilities $p(A_1 \ldots A_p | X_1 \ldots X_p)$. We will depict the vector of these probabilities with the vector $x \in \mathbb{R}^d$. To say that our experimental setup is "local" or more precisely, described by a local hidden variable theory, is to say that each outcome depends only on the choice of measurement and some discrete, possibly infinite, hidden variable R[1].

$$p(A_1 \dots A_p|X_1 \dots X_p) = \sum_R p(A_1|X_1, R) \dots p(A_p|X_p, R)p(R) \tag{3.1}$$

First we want to show that if we call the vector of probabilities x, then x can be written as a convex mixture of a finite number of vectors that represent *local, deterministic* behaviors. If $A \in \{1, \dots, r\}$, then we can write some probability distribution $p(A)$ as a mixture of deterministic behaviors: $p(A) = \sum_{j=1}^{r} \lambda_j \delta_{A,j},\ \lambda_j \geq 0, \sum_j \lambda_j = 1$. Here δ is a discrete delta function and $\delta_{A,j}$ can be considered a deterministic behavior for A because if $p_{det}(A) = \delta_{A,j}$, then $A = j$ with probability 1.

For a conditional probability distribution $p(A_i|X_i)$, a deterministic behavior means, for each X_i, A_i takes some some value with probability 1: $p_{det}(A_i|X_i) = \delta_{A_i, f_{m_i}(X_i)}$. Clearly, there are r^s functions $f_m : \{1, \dots, s\} \to \{1, \dots, r\}$ enumerated(for the i-th party) by $m_i = 1, \dots, r^s$ and therefore r^s deterministic behaviors. If we dictate that the $p(A_i|X_i, R)$ on the right hand side of Eq. (3.1) must be local deterministic behaviors.

$$p_{det}(A_1 \dots A_p|X_1 \dots X_p) = \prod_{i=1}^{p} \delta_{A_i, f_{m_i}(X_i)} \tag{3.2}$$

This gives us r^{sp} different local deterministic behaviors (one for each $(m_1, \dots, m_p)$) for the full probability distribution. Now we call each local deterministic behavior $v^{(m)} \in \mathbb{R}^d, m = 1, \dots, r^{sp}$ and then add in some shared randomness by making a mixture of these deterministic behaviors depending on some variable i which occurs with probability λ_i.

$$x = \sum_{i=1}^{n_v} \lambda_i v^{(i)}, \lambda_i \geq 0, \sum_i \lambda_i = 1 \tag{3.3}$$

Clearly, the x that can be written this way forms a convex polytope in d dimensions of the $n_v = r^{sp}$ vertices. Also, we have seen that it is clearly a probability distribution of the form in Eq. (3.1). A more complicated argument that we have not shown is that every probability distribution of the form in Eq. (3.1) can be written in this way [57]. This argument basically proceeds by showing that the local randomness in $p(A_i|X_i, R)$ can be converted to shared randomness which only the i-th party acts on.

Definition 3.1.1. *For an experimental setup with p parties, making choices of s measurement settings with r possible results, we will refer to the convex polytope formed as the convex hull of $n_v = r^{sp}$ vertices in $d = (rs)^p$ dimensions and described in Eq. (3.3) as an (r, s, p) Bell polytope.*

Finding the convex hull of a set of vertices is not always a difficult problem. For instance, Avis

[1] In general this could be continuous. The same conclusion holds [49], but we choose this formulation to highlight other connections later on.

and Fukuda [58] demonstrate an algorithm for finding the n_f facets of a polytope described by n_v vertices in a d dimensional space in time $\mathrm{O}(n_v n_f d)$ and space $\mathrm{O}(n_v d)$. The catch is that this method is only guaranteed to be efficient for simple polytopes: that is, polytopes for which each vertex lies at the intersection of exactly d facets, and no more. For non-simple polytopes the number of facets may be exponential in the number of vertices [59]. Unfortunately, the Bell polytope is not simple and determining membership in a class of polytopes that includes some Bell polytopes has been shown to be NP-complete [60].

To demonstrate the tedium of attempting to enumerate all the facets of the Bell polytope one has only to review the few works that attempt to do so. In the case of the $(2, 2, 2)$ polytope there is, after symmetry reduction, only one nontrivial facet referred to as the CHSH inequality [11]. For even slightly larger polytopes, however, the number of facets becomes unwieldy. Pitowsky and Svozil [61] used brute force computational techniques to find all the facets of the $(2, 3, 2)$ polytope and the $(2, 2, 3)$ polytope with total numbers of facets 684 and 53856, respectively. Werner and Wolf [59] analytically derive an expression to find all the facets of the $(2, 2, p)$ polytope and find that there are 2^{2^p} of them. This is why we choose to look for a smaller number of nonlinear inequalities that provide simple sufficient tests for a point to lie outside the Bell polytope. Next, we discuss some properties of linear matrix inequalities and why they provide a natural candidate for such a test.

3.1.1 Symmetries in the Bell polytope

Treating the Bell polytope as generic clearly forfeits the advantage we gain from knowing its many symmetries. For instance, we know that any valid probability distribution should still be so, even after relabeling of measurements, outcomes, or parties. Another source of symmetries is even easier to implement for the purpose of dimension reduction: normalization and no-signaling. Both normalization and no-signaling give us linear equations that we know every x must satisfy. By noting that for each of s^p settings the probability for the results add to unity, we can reduce the dimension by s^p. For no-signaling it is a bit more difficult, because many of the no-signaling equalities (see 1.3) are linearly dependent. For the $(2, 2, 2)$ Bell polytope, for instance, we reduce the dimension from $d = 16 \rightarrow 12$(normalization)$\rightarrow 8$ (no-signaling). In the analysis to follow, we have always first reduced the dimensionality of the polytope in this way.

3.2 Linear matrix inequalities

A linear matrix inequality (LMI) has the form [62]

$$F_0 + \sum_i F_i x_i \succeq 0 \tag{3.4}$$

where the F_i are symmetric $m \times m$ matrices and the operator $A \succeq B$ is used to indicate that $A - B$ is positive semidefinite or, equivalently, has only nonnegative eigenvalues. LMIs have several properties we will require. First, the set of x satisfying Eq. (3.4) is a convex set. This set can alternately be represented as a semi-algebraic set $\{x : g_1(x) \geq 0 \ldots g_n(x) \geq 0\}$ for some polynomials with max $degree(g_i(x)) \leq m$. This can be seen, for instance, by taking the determinant of all the principal minors to be non-negative, a necessary and sufficient condition for positive semidefiniteness.

Given some F_i, optimizing a linear function $b \cdot x$ over an LMI is referred to as a semidefinite program and can be solved efficiently using interior point methods [63]. Alternatively, we can specify some points that lie inside our convex set; then optimizing over the parameters in F_i is a semidefinite program, as we now show. The convex hull of some vertices $v^{(1)} \ldots v^{(n_v)} \in \mathbb{R}^d$ is the minimal convex set containing these vertices. Therefore, any convex set containing the vertices is guaranteed to be an outer bound for the polytope. We can cast the condition that our LMI in Eq. (3.4) defines a set including the vertices as another LMI.

$$\text{diag}_k(F_0 + \sum_i F_i v_i^{(k)}) \succeq 0 \tag{3.5}$$

Here diag_k indicates that we construct the block diagonal matrix where each block corresponds to the condition that vertex k is inside our convex set.

Clearly, once we have found a set of F_i satisfying the constraint 3.5, violating that constraint tells us we are outside of the outer bound of the Bell polytope, and therefore the Bell polytope itself. Therefore, this solution provides a simple sufficient test for nonlocality. The difficulty is in searching for F_i that provide a "good" test, or rather, an outer bound that is as close to the Bell polytope as possible.

3.3 Methods for approximating convex hulls

Now that we have cast the problem of finding tests for nonlocality in terms of finding the convex hull of a set of vertices, we proceed to describe several methods for finding outer bounds on this polytope.

3.3.1 Minimum volume ellipsoid

We begin by considering the most common approach to approximating the convex hull of a set of points with the minimum volume ellipsoid (MVE) containing the points. Searching for such an ellipsoid can even be cast as a convex optimization problem and therefore solved efficiently using interior point methods [64]. See [65], for example, for a discussion of the complexity of finding the minimum volume ellipsoid.

In the following definition A is a positive semidefinite, symmetric, $d \times d$ matrix that, therefore, defines a convex set via $\{x : (x-c)^\dagger A(x-c) \leq 1\}$, where c represents the center of the ellipsoid. Maximizing the log of the determinant is equivalent to minimizing the volume [64], and the last line dictates that all the sample data which belong to this class should be inside the ellipsoid.

$$\begin{aligned} \max_{A \in \mathcal{S}^{d\times d}, c \in \mathbb{R}^d} & \quad \log\det A \\ A \succeq 0 & \\ (v^{(k)} - c)^\dagger A (v^{(k)} - c) & \quad \leq 1, k = 1 \ldots n_v \end{aligned} \tag{3.6}$$

Applying this technique to the Bell polytope with $(r = 2, s = 2, p = 2)$ gives an outer bound that is far outside even the no-signalling polytope. Comparison of various methods for approximating the $(2, 2, 2)$ Bell polytope are summarized in Figure 3.1. We might hope to extend this volume minimization technique to more complicated volumes. Unfortunately, there is no easy way to determine the volume of convex sets with higher dimensional boundaries.

3.3.2 Schur improvement over MVE

In this section we present a general method for improving on the minimum volume ellipsoid. First we need to review the properties of the Schur complement, which we will use to rewrite the MVE in LMI form. Then we will use the Schur complement again to find an improved description of the polytope.

For a block diagonal matrix,

$$\begin{pmatrix} C & D \\ D^\dagger & E \end{pmatrix} \succeq 0$$

is equivalent via the Schur complement[66] to the statement,

$$C \succeq 0, C - D^\dagger E^{-1} D \succeq 0. \tag{3.7}$$

To see that the equation for an ellipse $x^\dagger A x \leq 1$ can be written in this form, note that we require $A \succeq 0$ which implies the existence of a Cholesky factorization $A = B^\dagger B$, where B is a $d \times d$ matrix and Bx is a $d \times 1$ vector. Therefore we can rewrite the ellipse equation as $(Bx)^\dagger (Bx) \leq 1$.

$$\begin{pmatrix} 1 & (Bx)^\dagger \\ Bx & \mathbb{I} \end{pmatrix} \succeq 0 \tag{3.8}$$

Note that we can also use the Schur complement to write this as

$$\mathbb{I} \succeq (Bx)(Bx)^\dagger. \tag{3.9}$$

Clearly, if we wanted to define a smaller set we could consider $\mathbb{I} \succeq \mathbb{I} + \sum M_i x_i \succeq (Bx)(Bx)^\dagger$, noting that $\mathbb{I} + \sum M_i x_i \succeq (Bx)(Bx)^\dagger$ can still be cast as an LMI. Of course, we still want the vertices to be inside our convex set, which leads to a condition linear in variables M_i,

$$\mathbb{I} + \sum M_i v_i^{(k)} \succeq (Bv^{(k)})(Bv^{(k)})^\dagger.$$

Now, the question remaining is to how enforce the condition that $\mathbb{I} + \sum M_i x_i$ is smaller than $\mathbb{I}$. Various possibilities can be cast as semidefinite programs. For instance, minimizing the maximum eigenvalue of a matrix can be cast this way.

$$\begin{aligned} &\min_{y,t} t \\ &t\mathbb{I} \succeq G_0 + G_i y_i \\ &H_0 + H_i y_i \succeq 0 \end{aligned} \tag{3.10}$$

This minimizes the maximum eigenvalue of the matrix $G_0 + G_i y_i$ over y, given some LMI constraint. An even simpler condition is to minimize the trace of a matrix, which is the same as minimizing the sum of the eigenvalues.

Given $b_i = \text{column}_i(B)$, vectors formed from the i-th columns of B defined in Section 3.3.1, we define the semidefinite program,

$$\begin{aligned} &\min_{M \in \mathbb{S}^d} \sum_{k,i} \operatorname{tr} M_i v_i^{(k)} \\ &F_i = \begin{pmatrix} 0 & (b_i)^\dagger \\ b_i & M_i \end{pmatrix} \\ &\text{diag}_k(\mathbb{I} + \sum_i F_i v_i^{(k)}) \succeq 0 \end{aligned} \tag{3.11}$$

From the Schur complement we can see that a solution to this problem corresponds to the set of x defined by the LMI

$$\mathbb{I} + \sum_i \begin{pmatrix} 0 & b_i^\dagger \\ b_i & M_i \end{pmatrix} x_i \succeq 0. \tag{3.12}$$

This defines a convex set which, from the third line of Eq. (3.11) is guaranteed to contain the vertices. Furthermore, our minimization had the effect of making this set smaller than the MVE, which we recover by simply setting $M = 0$. The result of using this technique on the $(2, 2, 2)$ polytope is presented in Table 3.1. We used a popular front-end for SDP solvers called YALMIP[67]. The solutions shown here used the specific solver SeDuMi [68].

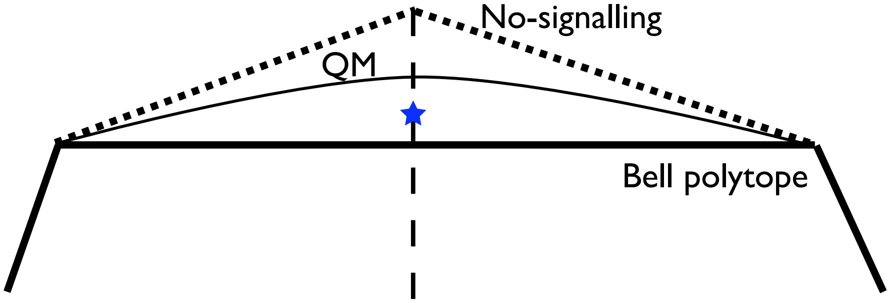


Figure 3.1: A schematic representation of the geometry of the (2,2,2) Bell polytope [69]. The solid line represents a facet corresponding to the CHSH inequality. The dotted line represents the larger polytope consisting of all nonlocal correlations consistent with the no-signalling principle. The region accessible by quantum mechanics is represented by a curved line between the two. The blue star represents the closest achievable outer bound obtained as detailed in Table 3.1.

	Max η
Local (Bell polytope)	0.5
Achievable by quantum states	~ 0.70
Nonlocal (no-signalling polytope)	1
Minimum volume ellipsoid	1.42
Schur improvement to MVE	0.64
Maximum curvature (SOS)	1.01

Table 3.1: The dashed line in Fig 3.1 parametrizes the states between the center of the Bell polytope ($\eta = 0$) and the maximally nonlocal state ($\eta = 1$). The table summarizes the maximum value of η for which a state is inside the corresponding set.

3.3.3 Maximum curvature

In this section we would like to highlight a recent promising approach to the problem of approximating convex functions. In [70], the authors consider a convex set constructed as the sub-level set of a convex function $g(x) \leq 1$. They impose convexity by searching over the set of $g(x)$ that are constructed as linear combinations of monomials plus a condition that ensures the Hessian of $g(x)$ is positive semidefinite. They use a stricter condition than positivity by imposing the Hessian be a sum of squares because this can be cast as an LMI [56]. Note in our approach we impose convexity directly by only considering LMI sets.

After constructing a function so that convexity can be cast as an LMI, they give a heuristic condition for minimizing the volume of sets defined by high degree polynomials. In particular, they suggest maximizing the curvature subject to the conditions of convexity, and inclusion of points in the convex hull. This can be cast as a determinant maximization problem just as in 3.3.1.

This approach seems promising and could be easily adapted to polytope fitting. Instead of maximizing the curvature in all directions, we maximize the curvature only at the vertices. We have implemented this approach for the $(2, 2, 2)$ Bell polytope with limited success. In particular, we used monomials of a maximum degree of 4 in the problem variables. Then we attempted to maximize the curvature at each vertex. This led to an outer bound close to the no-signaling polytope. Unfortunately, using higher degree monomials quickly increases the computational complexity, unless one can restrict to a specific family of monomials that is known to better represent the set.

The idea of maximizing curvature could be adapted to the LMI approach. On the boundary of our LMI, $g(x) = \det(F_0 + \sum_i F_i x_i) = 0$. The Hessian matrix H of $g(x)$ would have components $H_{i,j} = \partial_{x_j} \partial_{x_i} g(x)$. Even under the simplifying conditions that $F_0 = \mathbb{I}$ and F_i are traceless, $H_{i,j} \sim \operatorname{tr} F_i F_j$, which is not linear. Therefore, it is not clear that curvature maximizing approach can be adapted to search for small LMI representations of convex polytopes.

3.4 Complexity of tests

We will put all complexities in terms of the original problem variables (r, s, p), remembering that $d = (rs)^p$ and $n_v = r^{sp}$. First of all, we consider the complexity of finding an outer approximation of the (r, s, p) Bell polytope in Section 3.3.2. In general, the complexity of semidefinite programs is polynomial in the size of the problem, though current bounds are not very tight and depend on the problem structure. The leading complexity in our case comes from solving the determinant maximization problem for finding the MVE. This problem has been studied in depth [65] and has been found to have complexity $\mathrm{O}(n_v^{3.5} \ln n_v)$.

Given a description of the Bell polytope and of our relaxation of it, we compare the complexity of determining whether a given probability distribution is outside. For the Bell polytope, we simply check the linear inequalities for the facets. Unfortunately there could be $O(2^{2^p})$ of them [59]. To check our LMI from Section 3.3.2, we have to check whether a $(d+1) \times (d+1)$ matrix is positive semidefinite. Using a Cholesky decomposition,e.g., would require $O(d^3 = ((rs)^p)^3)$ operations. This is exponentially easier to check, and doubly so in terms of s. On the other hand, the construction was still exponential in s; the LMI construction and testing are only jointly polynomial in r.

3.5 Application to statistical learning

We now consider applying the methods of the previous section to computational problems unrelated to quantum mechanics. First we discuss the direct connection between hidden variable theories and Bayesian graphs. Second, we consider the use of improved approximations for the convex hull of points for classifying clusters of data.

3.5.1 Hidden nodes in Bayesian graphs

Bayesian statistics are an important tool for dealing with uncertainty in data and missing information[71]. Unfortunately, it can be difficult to determine the correct relationship between data sources, particularly if some unknown factors, or hidden variables, are affecting the data. We would like a test that tells us when correlations in some probability distribution over observed data can be explained solely in terms of some unknown hidden variable.

In this section, we will illustrate the connection to Bayesian graphs by exploring a particular probability model depicted by the graph in Figure 3.2. We will identify the set of probability distributions over observed data that are consistent with this model. Then we will devise a test to distinguish whether a given distribution is in this set. From a physics perspective, if we view A and B as causally disconnected detectors, and X and Y as measurement choices for those detectors, than correlations stronger than those predicted by a hidden variable model indicate the violation of locality. On the other hand, we can consider these variables in an abstract way that could apply to any data set using the language of Bayesian graphs.

We will speak of probability distributions over observed variables that, for simplicity, take on some finite set of values. For instance $p(A=i|X=j)$ denotes the probability that the variable A takes on the specific value i, given that we know that X has taken value j. The Bayesian graph in Figure 3.2 has the following meaning in terms of probability distributions [71].

$$p(ABXY) \quad = \quad \sum_R p(A|XR)p(B|YR)p(R)p(X)p(Y)$$

If we consider the conditional probabilities of $p(AB|XY)$, then we have an identical probability structure to the one considered in Eq. (3.1). Let us assume that A, B, X, Y take values in $\{0, 1\}$,

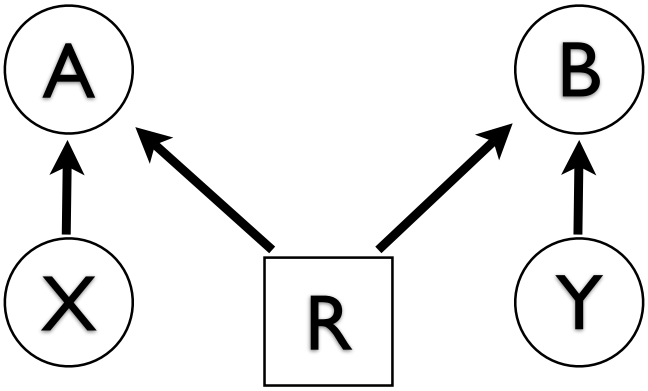


Figure 3.2: Bayesian graph with one hidden node R

while our unknown hidden variable R is unrestricted in the values it can take.Therefore, if we consider this space of probabilities geometrically, we have a structure identical to the Bell polytope.

In the analysis of Bayesian graphs, one usually considers "independence relations". These are relationships among the variables that can be read directly from the graph structure using a set of intuitive rules. Mathematically an independence relation has the following meaning:

$$U \perp V|W \Longleftrightarrow p(UVW)p(W) = p(UW)p(VW) \ \ \forall U, V, W \tag{3.13}$$

One usually says "U is independent of V given W" which has the rough meaning that if you know the value of W, then information about V tells you nothing about U. Typically, one may use independence relations in the data to infer the structure of a Bayesian graph representing the relations between variables. Unfortunately, if there is a hidden variable, then one cannot use independence relations involving that variable to deduce structure. Three independence relations that can be read off the graph in Figure 3.2 are the following [71].

$$\begin{aligned} AX &\perp Y \\ BY &\perp A \\ X &\perp Y \end{aligned} \tag{3.14}$$

Other relations (that do not involve R) can be derived from these. We can take the first of these and rewrite it using Eq. (3.13)

$$p(AXY) = p(AX)p(Y) \ \ \forall A, X, Y. \tag{3.15}$$

Unfortunately, this approach has the same limitations as we have previously discussed for the Bell polytope. Recall that no-signalling relations for two parties have the form

$$\sum_B p(AB|XY) = \sum_B p(AB|XY') \quad \forall B, X, Y, Y'.$$

Now we rewrite the sums as marginal probabilities and multiply both sides with $p(X)$ (also noting that $X \perp Y$ in this case).

$$p(AX|Y) = p(AX|Y') = P(AX). \tag{3.16}$$

Rearranging Eq. (3.15) gives the same result. We see that these independence relations give identical constraints to the no-signalling polytope. Thus, using independence relations to identify whether a given probability distribution is inconsistent with the graph in Figure 3.2 does no better than using the no-signalling polytope to identify whether a probability distribution is inconsistent with the Bell polytope, as depicted in Figure 3.1.

To conclude, we have shown that, at least for some Bayesian graphs, identifying whether probability distributions are consistent with some model can be at least as difficult as identifying facets of the Bell polytope. Since the given example is identical to the Bell polytope, the approximations given in Section 3.3 could work equally well for this problem. It remains an open question whether other Bayesian graphs share this difficulty and, if so, whether convex approximations can be used as efficient tests for deciding between them.

3.5.2 Convex bounds for classifying data

We propose a new method for representing and classifying data and demonstrate its efficacy for learning from certain hypothesis classes.Specifically, we show that a piecewise quadratic boundary containing these classes can be found efficiently using semidefinite programming. This method sidesteps many problems of similar approaches such as the need for high dimensional feature spaces, tuning of parameters, convergence problems, and local minima. By using a number of low degree curves, we improve on piecewise linear techniques while avoiding the danger faced by, e.g., kernel methods of overfitting in some highly nonlinear feature space. A description based on this Quadratic Boundary Method (QBM) also generates a piecewise quadratic decision boundary between classes. A particular hypothesis class for which we demonstrate usefulness of this method consist of data points generated as arbitrary mixtures of some particular finite set of data points. These classes are geometrically described by convex polytopes. Unfortunately, polytopes in higher dimensions may have a huge number of facets, so checking a point against the linear inequalities implied by each facet is NP-complete in the number of vertices [60]. On the other hand, known techniques which solve the problem more efficiently ignore the fine structure of the polytope. We show that QBM replaces the exponential number of linear surfaces with a small number of quadratic ones, while preserving more of the structure than Gaussian based cluster techniques, and avoiding problems typical of neural networks.

The rich variety of neural network and kernel machine methods contain many candidates to solve almost any learning problem. On the other hand, this wealth of techniques can become a burden if one is unsure which methods are appropriate. One can view neural networks as iteratively searching through a high dimensional space for a non-unique local minima that depends partly on the problem and partly on the number of nodes and variety of threshold function used. So although neural networks can learn linear, quadratic, or hyperquadratic boundaries between classes [72], it is not always clear what the meaning of these boundaries is. Support vector machines, on the other hand, have an obvious meaning in terms of finding an optimal unique linear separator. Unfortunately, doing so often requires a transformation to some ad hoc feature space.

By basing our technique on the construction of semidefinite programs, we avoid many of these pitfalls. Semidefinite programs are a class of well-studied convex optimization problems for which unique optimal solutions can be found efficiently and which have a clear geometric meaning [64]. Although we are limited by convexity requirements, for some classes of problems this is a natural constraint. We demonstrate the applicability of our method to one such class of convex problems and discuss possible extensions to non-convex regimes. Another interesting difference from other classifiers is that we learn a one dimensional convex function for each class separately, which we interpret as a measure of the distance of any point from that class. Our decision boundary is only constructed at the end based on which class is closest in terms of our distance function to a given point. This implies that, unlike SVM, or neural networks, we can add new classes independently, and re-form our decision boundaries without doing any new calculation involving points from our previously described classes.

We consider a set of labeled training data of the form (q_i, x_i) where $q_i = \{1, 2, \ldots\}$ labels which class the sample $x_i \in \mathbb{R}^d$ is a member of. The set of all training samples in the class g is called $\mathcal{T}_g$ and is a subset of the set of all possible samples with label g called $\mathcal{S}_g$.

One motivation for our proposal is to provide a better understanding of data that arise from a mixture of canonical processes. To be precise, imagine some set of samples $\mathcal{S}_1 = \{x : x = \sum_{j=1}^{m} \lambda_j v_j, \lambda_j \geq 0, \sum_j \lambda_j = 1, v_j \in \mathbb{R}^d\}$. The v_j represent specific data outcomes that are mixed according to a hidden variable λ.

We will demonstrate such a model, and the deficiency of current methods for handling it, with a physically motivated example. Because it is difficult to visualize high dimensional data, we will purposely choose a less realistic, low dimensional one for illustration purposes. We would like to reiterate that although in low dimensions, with few vertices, it may seem trivial to determine all the

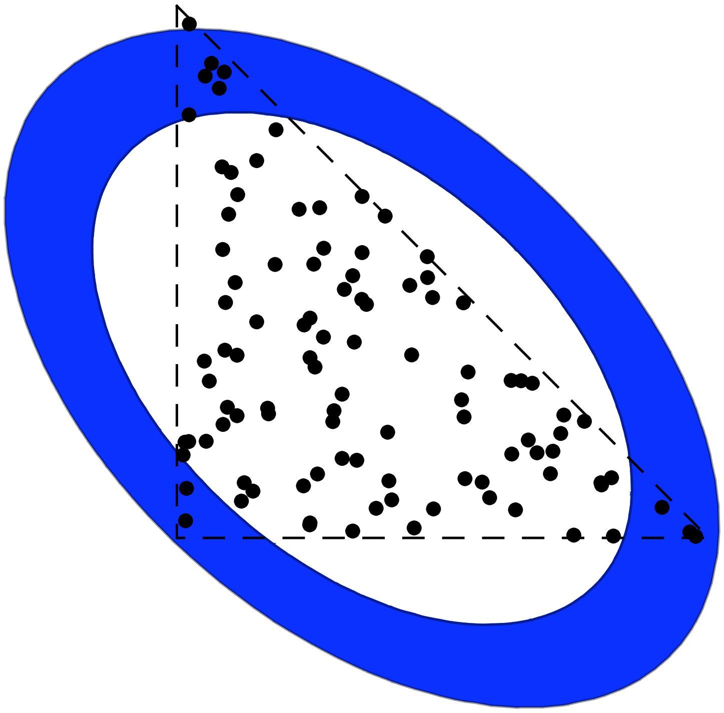


Figure 3.3: The vertices of the triangle (dotted lines) represent outcomes which are mixed according to some hidden variable model. Points represent sample data generated by an example hidden variable model. The shaded region represents a contour of a Gaussian describing the data. According to the Gaussian model, points land with equal probability in all sections of the blue shaded region with a total probability of $\sim 10\%$.

facets of some polytope, in high dimensions this approach can become exponentially harder in the number of vertices [59].

Given some labeled training data we would like to learn a description that can classify future data. In our simplified example, we suppose that our data consist of points $x \in \mathbb{R}^2$, and that for a specific class that we are learning the boundaries of, $x = \sum_i \lambda_i v_i$. That is, x is a mixture of particular vertices. The points in Figure 3.3 represent sample training data generated according to some hidden variable model.

We would like a classification that includes some of the structure of the polytope. A typical way to represent this cluster of points would be to assume a Gaussian distribution. The contours in Figure 3.3 come from assuming a Gaussian distribution with a mean and covariance that match the training data. For the shaded region in Figure 3.3, slices of equal area are (approximately) equiprobable. Clearly, large portions of data outside the polytope are considered just as probable as regions inside. One solution that seems easy is to try to learn linear descriptions that approach each of the facets of the polytope. Indeed, in two dimensions, an algorithm which learns the boundary of a polygon from some points runs in time $O(m \log m)$, if we are given m sample points. In higher dimensions, however, not only does the algorithm run polynomially more slowly in m, but the number of facets may be exponential in the number of vertices [59]. That means that even if we could learn all the facets (linear inequalities) which describe the class, it would be inefficient to use them to check whether a point is inside the class.

The next section proposes a middle ground, where we approximate the polytope with a small number of nonlinear, specifically quadratic, inequalities. This preserves some of the structure of our class while reducing the computational load necessary to determine membership.

3.5.2.1 Method

As in Section 3.3.1 we use the minimum volume ellipsoid as our starting point, and then consider improvements to it. Note that our earlier requirement that all training points are inside the ellipsoid can be vastly simplified when considering clusters of data. Just as SVMs rely only on a small number of support vectors near the the linear surface being learned, so the ellipsoid method will rely only on a small number of points near the vertices and these points can be found efficiently [65] as we discuss in the section on complexity.

First, we repeat this method to learn a description of each class. Then for each class, we can view the convex function $f(x) = (x - c)^\dagger A(x - c)$ as a distance, and the points inside the class are just those within a distance of 1. This gives us a measure for how far arbitrary points are from each class. This is depicted graphically in Figure 3.4.

After we have defined an ellipsoid and associated convex function for each class, we would like a method to decide to which class an unlabeled data point is most likely to belong. A natural way to decide is to use our distance measure to put the point in the class that it is the smallest distance away from. This leads to piecewise quadratic decision boundaries as depicted in Figure 3.5.

We already have some desirable properties: a quadratic class membership test, a built-in notion of distance from some class, and a natural way to form decision surfaces between classes. We still have not achieved our goal of capturing any of the structure of a polytope, though. One path for improvement would be to use higher degree polynomials. In that case, we can no longer cast the problem as a semidefinite program, so it may be hard to solve efficiently.

The new idea that we introduce here is to instead use the intersection of a small number of quadratic inequalities (which can be found efficiently) to describe our set more precisely. The intersection of convex sets is a convex set, and, by our prior reasoning, if this set includes the vertices, it must also include the polytope in this case. We will consider the intersection of k sets $\{x : \forall i \in \{1, \ldots, k\}, (x - c_i)^\dagger A_i(x - c_i) \leq 1\}$ For each i, c_i, and A_i are solutions to the semidefinite

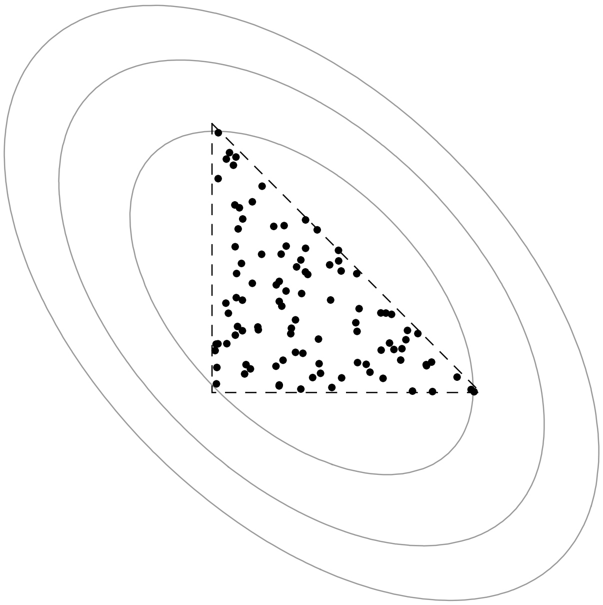


Figure 3.4: A contour plot of $(x-c)^\dagger A(x-c) \leq \{1, 2, 3\}$. The inner contour represents the minimum volume ellipsoid containing the points.

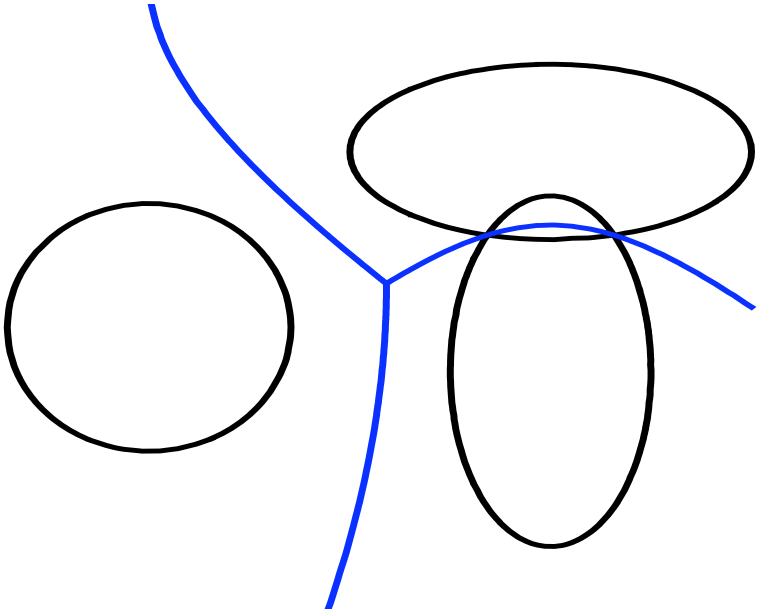


Figure 3.5: Black ellipsoids depict learned boundaries for three different classes of training data. Blue lines depict decision surface for determining how to classify an unlabeled sample. Note that boundaries are piecewise quadratic and that non-separable classes still generate a decision boundary.

program:

$$\begin{aligned} \max_{c_i, A_i \in \mathcal{S}^{d\times d}} \quad & \text{Condition}_i(A_i, c_i) \\ A_i \succeq 0 \quad & \qquad\qquad\qquad\qquad\qquad\qquad (3.17) \\ (x-c)^{\dagger} A (x-c) \leq 1, \quad & \qquad \forall x \in \mathcal{T} \end{aligned}$$

All that remains is to set some limit on the number of intersecting sets we would like to consider, along with conditions for each set. One choice is to individually maximize the diagonal elements of A, leading to a number of optimizations linear in the dimension. Geometrically, this can be interpreted as finding the narrowest ellipsoid around a specific axis. Although this simple method achieves an improvement over the single minimum volume ellipsoid, the existence of an optimal choice in tradeoff between adding more conditions and a better description of the polytope is a question for further research.

An example is provided in Figure 3.6, where we chose $k = 2$ and maximized two different elements of the matrix A. Note that the area between parallel lines is a special case of an ellipsoid. Clearly, the intersection of these ellipsoids provide a better description of the polytope than the minimum volume ellipsoid. For an intersection of ellipsoid we can define our distance function as $f(x) = \max_i((x - c_i)^{\dagger} A_i(x - c_i))$, then all the points inside the class are still within a distance of 1.

Alternately one could consider directly applying the method of 3.3.2 as depicted in Figure 3.1. Although this gives a good improvement to MVE by solving only one semidefinite program, by incorporating higher dimensional polynomials, we lose some notion of a distance, and therefore our ability to construct decision surfaces between classes.

3.5.2.2 Complexity

If we use k intersecting ellipsoids, then the complexity of our algorithm should be at most k times the complexity of solving the well-studied minimum volume ellipsoid problem in Eq. (3.6). In fact, we may do better by making our k optimizations over some function that is easier to compute than the log of the determinant. Nevertheless, current algorithms depend on the number of sample points m with $\mathcal{O}(m^{3.5} \log m)$. Results were calculated in [65] for m as high as $30,000$ and dimensions as high as $d = 500$, in times under 30 seconds on a regular PC. Clearly, large problems can be feasibly handled with this technique.

3.5.2.3 Non-convex sets

For linear classifiers like SVM, it is easy to construct examples for which the method fails. For instance, if class a consists of a ring of points surrounding another class b which is in the middle of the ring, then there is no linear classifier which will distinguish the two classes. Similarly, because

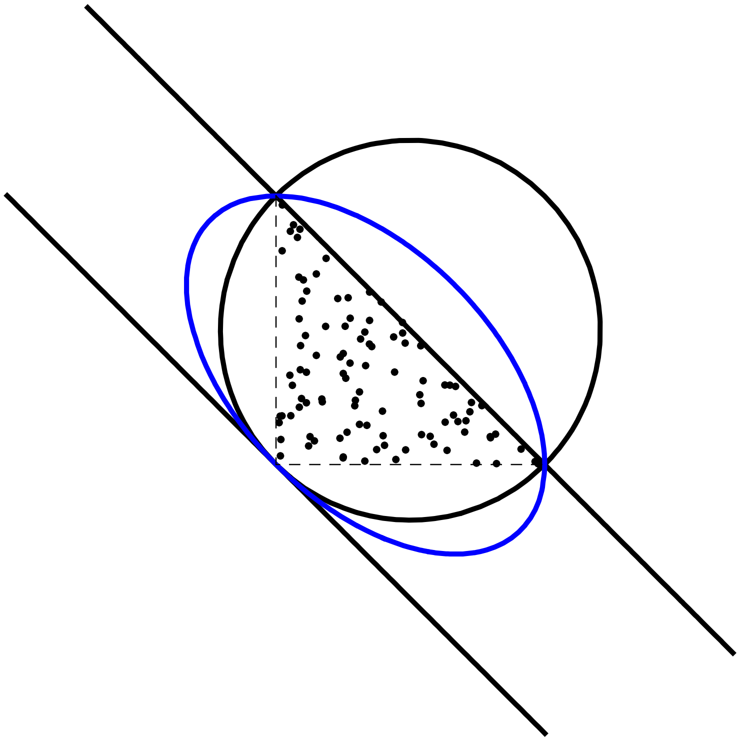


Figure 3.6: The intersection of the ellipsoids represented by black lines forms a convex set which more closely represents the polytope than the minimum volume ellipsoid (blue line)

QBM approximates any set as convex, then a non-convex class like a will clearly be mischaracterized. The solution is the same in both cases, we consider a feature space in which our classifier is adequate. What is not known is what feature space will be adequate, and constructing feature spaces in very high dimensions may make a problem intractable. We have already seen in Figure 3.5, a situation in which QBM produces a decision boundary that can not be formed by a linear classifier. QBM is appropriate for data which are convex, and clearly this is a more general class than data which are linearly separable. It remains unknown whether, in cases where both methods fail, there exist simpler feature spaces in which QBM succeeds while SVM fails.

3.5.3 Discussion

We have shown the applicability of approximate hidden variable tests to the problem of learning the structure of Bayesian graphs. We have also introduced a new learning technique which specifically addresses the problem of learning in hidden variable models and demonstrated the technique on example data.

Many open questions remain. Are there other Bayesian graphs with hidden variables which suffer from a gap between probability distributions described by the graph versus those that satisfy the independence relations? If so, can they be addressed by convex methods? Are there other choices of conditions in Equation 3.17 that provide a better description of the model? Are there other models (besides hidden variable models) for which the QBM provides a better description than other techniques? Although semidefinite programs used in formulating QBM are known to be easily solved in most practical applications, few theoretical bounds are known. Therefore, numerical tests seem to be the best way to estimate performance in real world applications. Can QBM be combined with other techniques to provide nonlinear bounds more general or more efficiently than current techniques?

3.6 Other techniques

We now turn to a few alternate techniques for constructing hidden variable tests. Despite the fact that these techniques are computationally infeasible, we present them here in the hope that they can still be edifying.

3.6.1 Using algebraic geometry to construct tests of nonlocality

A *semi-algebraic set* can be defined as a collection of polynomial equalities and inequalities defined over $\mathbb{R}^n$. We consider A, B, X, Y taking values in $\{0, 1\}$. For any conditional probability distribution we have:

$$\forall X, Y, \sum_{AB} P(AB|XY) = 1 \quad \text{Normalization} \tag{3.18}$$

$$\forall A, B, X, Y, P(AB|XY) \geq 0 \quad \text{Positivity.} \tag{3.19}$$

If we take our variables to be the components of our probability vector, this can be considered a semi-algebraic set. We can specify particular theories by imposing further conditions for example, those described previously: locality in definition 1.3.1 and no-signaling in definition 1.3.3. Adding these conditions still leaves us with a semi-algebraic set.

Consider the truth statement for whether a hidden variable theory can produce a certain observed probability distribution $P(AB|XY)$.

$$\begin{gathered} \exists P(R), P(A|X,R), P(B|Y,R) \text{ such that} \\ P(AB|XY) = \sum_{R} P(R)P(A|X,R)P(B|Y,R) \end{gathered} \tag{3.20}$$

We do not know and cannot measure the value of the hidden variables. However, this statement can be converted via *cylindrical algebraic decomposition* into a system of equalities and inequalities on the observed variables only [73]. This would yield an automatic way of generating Bell inequalities. Unfortunately, this technique is doubly exponential, too slow for even the simplest cases.

Implicitization is a technique that uses Gröbner bases to find the smallest *algebraic variety* (set of equality statements) that contains a semi-algebraic set [74]. Before we can apply this method to Eq. (3.20), we have to confront the fact that our hidden variable R could, in principle, be infinite. To make the problem easier we will suppose that R is in fact finite and observe the results.

We apply this method to a local theory with a hidden variable of one bit. That is, we consider states that can be written in terms of a hidden variable as in Eq. (3.20) except that we restrict $R = \{0, 1\}$. This yields three conditions on the observed distribution: normalization, the no-signalling condition, and this:

"Bell equality" for a 1-bit HV theory

$$\begin{array}{ccc} \mathrm{P(01|11)P(10|11)P(11|00)+} & & \mathrm{P(01|11)P(10|01)P(11|10)+} \\ \mathrm{P(00|11)P(11|01)P(11|10)+} & = & \mathrm{P(00|11)P(11|00)P(11|11)+} \\ \mathrm{P(01|10)P(10|01)P(11|11)} & & \mathrm{P(01|10)P(10|11)P(11|01).} \end{array}$$

Interestingly, for a hidden variable of more than one bit, this condition ceases to be true. Although we get interesting relations for finite hidden variable theories, implicitization cannot be used to obtain generic Bell inequalities.

Implicitization has two limitations: it can only find equality constraints, and, although it can solve small examples, it scales poorly. Therefore, we need a technique that allows us to relax our original problem enough to find useful bounds. One such method is to search for bounded degree certificates using the *positivstellensatz*.

3.6.2 Bounded degree certificates

The *real positivstellensatz* [56], for simplicity stated here with only one equality and one inequality, states that the set

$$\{x \in \mathbb{R}^n | f(x) = 0, g(x) \geq 0\} \text{ is empty}$$

if and only if there exists polynomial $t(x)$ and sum-of-squares (and therefore positive) polynomials $s_0(x)$ and $s_1(x)$ so that

$$t(x)f(x) + s_0(x) + s_1(x)g(x) = -1.$$

Finding polynomials that achieve this condition provides a *certificate* that the original set is empty. If we search through the space of coefficients for polynomials of a bounded degree, the constraints are linear and can be cast in the form of a *semi-definite program.*

We can also use Smüdgen's theorem to find upper or lower bounds for some polynomial, given those constraints [75]. To lower bound the polynomial $h(x)$ subject to constraints above is basically equivalent to searching for a certificate of infeasibility for the the function $h(x) - \lambda$. As we search for polynomials of higher degrees, we get bounds closer to the actual global maximum.

3.6.3 CHSH example

As a simple example and proof of principle, we'll consider optimizing the value of the CHSH game (also called the XOR game) over the constraints in Eq. (3.20). We want to optimize the probability, ω, to "win" the game, which happens when $A \oplus B = X \wedge Y$ (with $A, B, X, Y \in \{0, 1\}$).We already know what the answer should be: $\omega_{local} \leq 0.75$, $\omega_{quantum} \leq 0.85355$, $\omega_{nonlocal} \leq 1$.

For a local model with 1 bit of shared randomness, the positivstellensatz returns the following bounds on the value of the CHSH game.

System	Max degree	Bound on CHSH
1 bit HV	4	0.8536
1 bit HV	6	0.75 (tight)
Nonlocal	0	1 (tight)

So we see that a search over low degree polynomials, done using an SDP, returns the correct bound. Interestingly, the degree 4 relaxation returns the quantum bound and it takes a degree of 6 to find the true bound. Note that the nonlocal constraints actually comprise a linear program and don't require the positivstellensatz construction[36].

Algebraic geometry and the positivstellensatz in particular provide powerful tools to answer a very general class of questions about GPTs. The general form of questions we can answer in this framework is: “Do there exists parameters in theory X, that allow state or transformation Y?” and, “What is the optimal value of a polynomial function of the observed variables for some theory X?” Unfortunately, answers to these questions cannot usually be efficiently found using this technique.

Chapter 4

Nonlocality in Curved Space

I saw a man pursuing the horizon;
Round and round they sped.
I was disturbed at this;
I accosted the man.
"It is futile" I said,
"You can never – "
"You lie," he cried,
And ran on.
Stephen Crane

In this chapter we explore the relationship between nonlocal correlations and spacetime curvature. In particular, we demonstrate a scenario in which presence or absence of nonlocal correlations can be used to determine the structure of spacetime.

Quantum fields in the Minkowski vacuum are entangled with respect to local field modes. This entanglement can be swapped to spatially separated quantum systems using standard local couplings. A single, inertial field detector in the exponentially expanding (de Sitter) vacuum responds as if it were bathed in thermal radiation in a Minkowski universe. We show that using two inertial detectors, interactions with the field in the thermal case will entangle certain detector pairs that would not become entangled in the corresponding de Sitter case. The two universes can thus be distinguished by their entangling power.

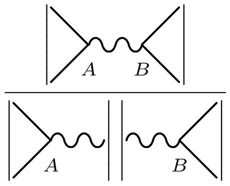
4.1 Background

Information in curved spacetime has played a prominent role in the attempt to understand the interface between quantum physics and gravity [76, 77, 78, 79]. While abstract properties of curved-space quantum fields (including their entanglement) can be studied directly [80, 81, 82, 83, 84], an operational approach involving observers with detectors historically has been a critical component of theoretical progress in this area [78, 85]. With the birth of quantum information theory [86], quantum systems could now be analyzed in terms of their use for information-theoretic tasks like quantum computation [86], quantum teleportation [87], and quantum cryptography [88]. Entanglement is a phenomenon that is uniquely quantum mechanical in nature [89] and can be considered both an information-theoretic and a physical resource [90]. It is known that the Minkowski vacuum possesses long-range entanglement [83] that can be swapped to local inertial systems using standard quantum coupling mechanisms [91]. Variations on this theme can be considered, including accelerating detectors [92], thermal states [93], and curved spacetime. Our focus will be on curvature. For this, we choose an exponentially expanding (de Sitter) universe [82, 94] for its simplicity and because of its importance to cosmology [95].

4.1.1 Nonlocality in flat space

The prototype for our results is a perturbative calculation of entanglement in flat space by Reznik[91]. Reznik considers two causally separated detectors that interact with the Minkowski vacuum. Schematically, we can view Reznik's result as relating the entanglement between Alice and Bob's detectors to the ratio of processes depicted by the following Feynman diagrams.

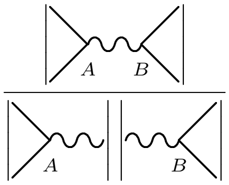
Entanglement $\sim$



(4.1)

Essentially, we're comparing the rate for the process in the numerator, which is entangling, to the process in the denominator which is described by the typical thermal spectrum any spacetime localized detector sees. When this term is larger than 1, Reznik showed that the detectors can be used to demonstrate violation of a Bell inequality. Although both terms fall off as one probes higher energy field modes, the thermal part in the denominator falls off faster, as depicted in Figure 4.1, allowing a small amount of entanglement to be distilled.

Another exploration of entanglement in flat space-time that has been more widely studied is the case of accelerating observers. In this case, we consider Alice and Bob's detectors to be accelerating away from each other at some rate proportional to κ. If they remain on these trajectories, they will be causally disconnected in the sense that signals sent from one never reach the other and vice versa [85]. Because they remain causally disconnected, any entanglement present in their detectors must come from the vacuum itself. This scenario was considered again perturbatively by Reznik, but was also solved exactly analytically [92], with similar results verifying the effectiveness of Reznik's calculation. The results for an analysis of Eq. (4.1) with detectors with an energy gap Ω coupled to the vacuum for a time T take the following form.



$$\sim \frac{\Omega T e^{-\Omega/2\kappa}\frac{1}{e^{\Omega/\kappa}-1}}{\Omega T e^{-\Omega/\kappa}\frac{1}{e^{\Omega/\kappa}-1}} \sim e^{\Omega/2\kappa}$$

This shows that no matter how weak the acceleration, Alice and Bob can distill some entanglement if their detectors probe high enough energy wavelengths. We now proceed to consider the effect of spacetime curvature on entanglement.

4.2 Setup

We wish to demonstrate a connection between a physical property of spacetime (curvature) and an information-theoretic resource (entanglement). While it is possible to directly study the entanglement present in a quantum field in de Sitter spacetime, this sometimes leads to difficulties [81] that are not present in a more operational approach. Still, it is known that entanglement between field modes can directly encode a spacetime's curvature parameters [84]. Motivated by a desire to be as operational as possible, we examine how curvature affects a field's usefulness as an *entangling resource*—i.e., its ability to entangle distant quantum systems ("detectors") using purely local interactions. We begin by reviewing the response of a single, inertial detector interacting with a massless, conformally coupled scalar field. The result in the vacuum de Sitter case is identical to that in the case of a thermal ensemble of field particles in flat spacetime [85, 78]. Next, we ask the question, *can entanglement be used to distinguish de Sitter vacuum expansion from Minkowski-space heating?* We show that with two detectors on comoving trajectories, there exists a parameter regime in which the local systems that couple to the field will become entangled despite the presence of extra thermal noise in each individual detector. Interestingly, this region of parameter-space in the expanding case is a *proper subset* of the same region in the locally equivalent thermal case. Thus, while both universes affect a local inertial detector in exactly the same way, entanglement between two detectors can be used to distinguish them.

4.2.1 Interaction Hamiltonian

We start with the following experimental setup, which is nearly identical to that used by Reznik et al. [91], using units where $\hbar = c = k_B = 1$. We pose our problem completely in operational terms, but our goal is to show proof of principle—not necessarily practicality of the method. We suppose that the inhabitants of a particular planet launch a satellite into space to measure the temperature of the universe they inhabit. On board this satellite is a qubit (a two-level quantum system), initially in the ground state $|0\rangle$, that gets coupled locally and for a limited time to a scalar field using a simple De Witt monopole coupling [96]. The time-dependent interaction Hamiltonian for this detector is, in the interaction picture,

$$H_I(\tau) = \eta(\tau)\phi\big(x(\tau)\big)\left(e^{+i\Omega\tau}\sigma^+ + e^{-i\Omega\tau}\sigma^-\right) , \tag{4.2}$$

where τ is the proper time of the satellite, $\eta(\tau)$ is a weak time-dependent coupling parameter (which we'll call the detector's "window function"), $x(\tau)$ is the worldline of the satellite, $\phi(x)$ is the field operator at the spacetime location x, and the rest represents the interaction-picture Pauli operator $\sigma_x(\tau)$ for the local qubit with (tunable) energy gap Ω. Roughly speaking, the detector works by inducing oscillations between the two levels at a strength governed by the local value of the field.

From now on, we refer to this qubit as a "detector", although the process of "detection" includes only the field interaction (before projective measurement). We wish to examine when two such detectors become entangled through their local interactions with the field, so we delay classical readout to allow for general quantum postprocessing, which may be necessary to show violation of a Bell inequality [97].

The window function $\eta(\tau)$ is used to turn the detector on and off, but the transitions must be sufficiently smooth so as not to excite the field too much in the process [98]. Beyond this requirement, on physical grounds, our results should not depend on the details of the window function as long as it is approximately time bounded, so we will always choose $\eta(\tau)$ to be proportional to a Gaussian, $\eta(\tau) = \eta_0 e^{-(\tau-\tau_0)^2/2\sigma^2}$, where $\eta_0 = \eta(\tau_0) \ll 1$ is a small unitless constant that enforces the weak-coupling limit and allows us to use perturbation theory. This window function approximates the detector being "on" when $|\tau - \tau_0| \lesssim \sigma$ and "off" the rest of the time and also has a nice analytic form.

4.2.2 First order perturbation

Without loss of generality, we can set $\tau_0 = 0$. To lowest nontrivial order in η_0, the qubit after the interaction (but before readout) will be found in the state $\rho = A\,|1\rangle\langle 1| + (1 - A)\,|0\rangle\langle 0|$, where

$$A = \int_{-\infty}^{\infty} d\tau \int_{-\infty}^{\infty} d\tau'\, \eta(\tau)\eta(\tau')e^{-i\Omega(\tau-\tau')}D^+\big(x(\tau); x(\tau')\big) , \tag{4.3}$$

where $D^+(x; x') = \langle\phi(x)\phi(x')\rangle$ is the Wightman function for the field, with expectation taken with respect to the state of the field (assumed to be a zero-mean Gaussian state, but not necessarily the vacuum). Repeated measurement in the $\{|0\rangle, |1\rangle\}$ basis for a variety of values of Ω allows for determination of the state of the detector as a function of Ω [1]. As is clear from Eq. (4.3), the state is completely determined by the detector response function $D^+\big(x(\tau);x(\tau')\big)$, which is the Wightman function taken at two different proper times along the worldline of the detector [85].

4.2.3 Spacetime structure

We consider two possible universes. The first is Minkowski, $ds^2 = dt^2 - \sum_{i=1}^3 dx_i^2$, with the field in a thermal state with temperature T with respect to the inertial trajectory $\{x_i\} =$ (constant). The second is a de Sitter universe, $ds^2 = dt^2 - e^{2\kappa t}\sum_{i=1}^3 dx_i^2$, where κ is the expansion rate, in the conformal vacuum. The conformal vacuum is the natural choice in this case because it is the unique, coordinate-independent vacuum state dictated by the symmetries of the spacetime. Furthermore, it can be justified on physical grounds because the conformal vacuum coincides with the massless limit of the adiabatic vacuum for de Sitter space [85]. Thus, we can think of this analysis as applying to the following two ways of adiabatically modifying the Minkowski vacuum: (1) very slowly heating the universe to a temperature T, and (2) very slowly ramping up the de Sitter expansion rate (from zero) to a final value of κ.

The variables $\{x_i\}$ are comoving coordinates, and t is cosmic time. (Since the Minkowski metric is the special case $\kappa = 0$, this terminology carries over to it, as well.) In both universes, worldlines of constant $\{x_i\}$ are inertial trajectories (geodesics), and intervals of proper time equal those of cosmic time ($\Delta\tau = \Delta t$). In both cases, the scalar field $\phi(x)$ is massless and conformally coupled [85], satisfying $[\Box_x + \frac{1}{6}R(x)]\phi(x) = 0$, where the Ricci scalar $R(x) = 12\kappa^2$ is a constant proportional to the expansion rate κ.

Gibbons and Hawking [78] showed that the detector response function for any inertial observer in the de Sitter case is *exactly the same* as that of a detector at rest in a thermal bath of field particles with temperature $T = \kappa/2\pi$ in flat spacetime. Thus, a single detector alone cannot distinguish between the two cases if it forever remains on a given inertial trajectory. In both cases considered above, the detector is at rest in the comoving frame and thus,

$$D_T^+\big(x(\tau);x(\tau')\big) = -\frac{T^2}{4}\,\mathrm{csch}^2[\pi T(t - t' - i\epsilon)]\ , \tag{4.4}$$

where the subscript T indicates that this is a detector response function for a thermal state at temperature T. When the satellite begins sending back measurement data, the reconstructed $A(\Omega)$ is found to be consistent with the detector being at rest in a thermal bath of field particles at a small but nonzero temperature T. If the inhabitants wish to know whether this perceived thermality is a result of heating or expansion, though, they must be more creative.

Obviously, they could use astrophysical clues (like we have done on Earth) and/or Doppler-

[1] More general measurements will be required to demonstrate entanglement between two such detectors, though.

shift measurements [2] to determine whether their universe is expanding or not, but we are going to restrict them to using only satellite-mounted detectors of the sort described above on fixed inertial trajectories. If the detectors are to be useful, then, they will need more than one.

We propose the following alternative that makes use of entanglement to distinguish the two universes. We imagine two satellites, each having many qubits that interact locally with the scalar field. (Having many detectors allows access to many copies of the same state.) We assume that the satellites have no initial entanglement with each other and that the qubits each begin in the ground state. After interacting with the field, measurement is delayed to allow for general quantum operations (local to each satellite) on the multitude of qubits on board. In the end, however, the only data that can be transmitted back to the home planet are measurement results, plus information about the postprocessing and the particular measurements performed.

In an attempt to be as simple as possible, we analyze the case of two inertial detectors, a and b, on the comoving trajectories $x_1 = \pm L/2$ (with $x_2 = x_3 = 0$). Due to the homogeneity and isotropy of space in both scenarios, this case is remarkably general—but not entirely so since one could imagine the detectors in motion with respect to each other (beyond the relative motion generated by any expansion). For simplicity, we'll also require that the two detectors have synchronized local clocks with $\tau_{a,b} = t$, equal resonant frequencies $\Omega_{a,b} = \Omega$, and identical window functions $\eta_{a,b}(\tau) = \eta_0 e^{-\tau^2/2\sigma^2}$. Finally, we desire that $L \gg \sigma$ so that the detector-field interactions can be considered noncausal events [3]. As we shall see, these restrictions will still allow the inhabitants, located at $x_i = 0$, to distinguish expansion from heating.

By spatial symmetry, each detector alone must respond using the detector response function from Eq. (4.4) and thus provides no useful information. The only hope, then, is in the correlations between the detectors. We will focus on those correlations that signal the presence of *entanglement* of the detectors after interaction with the field. For a pair of qubits, the negativity [99] of a state is nonzero if and only if the systems are entangled [100]. Since we have access to (by assumption) multiple copies of an entangled state of pairs of qubits, a local measurement protocol (on the many copies of the state) always exists to verify entanglement by showing a violation of a Bell inequality [97, 101]. This can be verified by a third party using classical data received from both satellites.

We will focus on finding the regimes in which entanglement is nonzero, rather than on the magnitude of the entanglement for two reasons. First, the amount of extractable entanglement is small enough to be impractical as a resource and will depend on the details of the detector coupling. Second, we are primarily interested in understanding a qualitative difference between the quantum behavior of curved and flat spacetime; examining entanglement ensures that this is a genuinely

[2] The thermal Minkowski case exhibits Doppler shifting for detectors at different velocities [85]; the vacuum de Sitter case does not [78].

[3] Although the Gaussian window functions technically have tails that extend forever, none of the results change if we assume a smooth cutoff of the Gaussian (to zero) around, say, 10σ as long as both L and T^{-1} are still much larger than this.

quantum mechanical effect [89].

4.3 Calculation of entanglement

An analogous calculation to Reznik's [91] shows that the negativity of the joint state of the qubits is $N = \max\big(|X| - A, 0\big)$, where A is the individual detector response from Eq. (4.3), while X is defined as

$$\begin{aligned} X &= -\int_{-\infty}^{\infty} dt \int_{-\infty}^{t} dt'\, \eta(t)\eta(t') e^{i\Omega(t+t')} \\ &\quad \times \left[D^+\big(x_a(t); x_b(t')\big) + D^+\big(x_b(t); x_a(t')\big)\right] \\ &= -2 \int_{t'<t} dt\, dt'\, \eta(t)\eta(t') e^{i\Omega(t+t')} D^+\big(x_a(t); x_b(t')\big) \ . \end{aligned} \tag{4.5}$$

The limits of integration enforce time ordering [102], so we can use the Wightman function as shown. This is useful because symmetry of the two detectors means that $D^+\big(x_a(t); x_b(t')\big) = D^+\big(x_b(t); x_a(t')\big)$, a fact used to obtain the second line. This integral measures the *amplitude* that the detectors will exchange a virtual particle, while A measures the *probability* that each detector becomes excited either by absorbing or emitting a particle.

4.3.1 Zero curvature

We begin by considering when the qubits become entangled when $T = 0$. (We also *define* $\kappa \equiv 2\pi T$ from now on so we can talk about expansion rates in terms of the associated Gibbons-Hawking temperature.) This case corresponds to the one considered by Reznik [91] using different window functions. In the $T = 0$ case, the Wightman function used in X is

$$D_0^+\big(x_a(t); x_b(t')\big) = \frac{-1}{4\pi^2\left[(t - t' - i\epsilon)^2 - L^2\right]} \ , \tag{4.6}$$

and the detector response function (used in A) is obtained by letting $L \to 0$ and is also obtainable as the limit of Eq. (4.4) as $T \to 0$. Both X and A can be evaluated analytically:

$$X_0 = -\frac{e^{-\frac{L^2}{4\sigma^2} - \sigma^2\Omega^2} \sigma \operatorname{erfi}\left(\frac{L}{2\sigma}\right)}{4L\sqrt{\pi}} \ , \tag{4.7}$$

$$A_0 = \frac{e^{-\sigma^2\Omega^2} - \sqrt{\pi}\sigma\Omega \operatorname{erfc}(\sigma\Omega)}{4\pi} \ , \tag{4.8}$$

where σ is the width of the window function (the time for which the detector is turned on), and the subscripts indicate that these are the Minkowski vacuum results, with $\operatorname{erfi}(z) = -i \operatorname{erf}(iz)$ and $\operatorname{erfc}(z) = 1 - \operatorname{erf}(z)$, where $\operatorname{erf}(z)$ is the error function. In the Minkowski vacuum case, the detectors become entangled if and only if $|X_0| > A_0$. A sketch of the behavior of these terms for a fixed L is shown in Figure 4.1. In the Minkowski vacuum case, the detectors become entangled iff $|X_0| > A_0$ [100]. This region in the L-Ω plane is above the slanted black line in Figure 4.2.

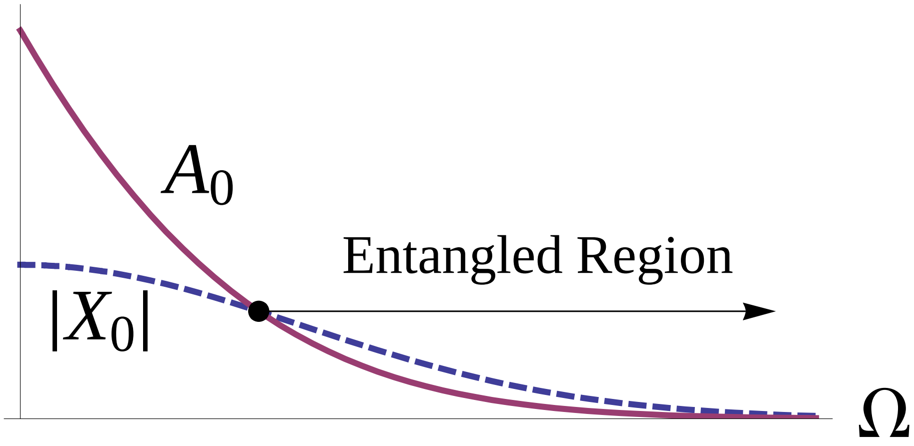


Figure 4.1: Entanglement from the Minkowski vacuum. The time-energy uncertainty relation $\Delta t \Delta E \gtrsim \frac{1}{2}$ implies that a field detector (with resonant frequency Ω) operating for a finite time has a nonzero probability A_0 of becoming excited even when the field is in the vacuum state. When the magnitude of the correlation amplitude $|X_0|$ for two detectors exceeds this value, the detectors become entangled [91, 100].

4.3.2 Nonzero curvature

Let's see what happens with a nonzero temperature. Since we are interested in the possibility that the perceived thermality is due to de Sitter expansion, we have a restriction on the temperature, which sets the scale for the cosmic horizon $L_H = \kappa^{-1} = (2\pi T)^{-1}$. If observers are to exist at all, this horizon must be much larger than their typical scale of experience, which can't be much smaller than σ if the detector is to be useful to them. (Consider how useful a "detector" that operates on the scale of the Hubble time would be for humans.) Thus, for de Sitter expansion even to be a possibility, we require that $T \ll \sigma^{-1}$.

In both cases, the detector response function is given by Eq. (4.4), while the Wightman function to be used in X in the thermal case is [103]

$$D^+_{\text{th}}\big(x_a(t); x_b(t')\big) = \frac{T}{8\pi L} \times \Big\{\coth\big[\pi T(L - y)\big] + \coth\big[\pi T(L + y)\big]\Big\} \tag{4.9}$$

and in the de Sitter case is [85]

$$D^+_{\text{dS}}\big(x_a(t); x_b(t')\big) = \left(\frac{-1}{4\pi^2}\right)\left[\frac{\sinh^2(\pi T y)}{\pi^2 T^2} - e^{2\pi T x} L^2\right]^{-1} , \tag{4.10}$$

where $x = t + t'$, and $y = t - t' - i\epsilon$ in both. One can verify that in both cases, taking $L \to 0$ gives Eq. (4.4), and taking $T \to 0$ gives Eq. (4.6).

In both the thermal and de Sitter cases, the integral in Eq. (4.5) can be well approximated by an asymptotic series in T (as $T \to 0$), generated from the Taylor expansion of D^+_{th} and D^+_{dS}, respectively, about $x = y = 0$. Although the radius of convergence of the Taylor series is finite, for any reasonable detector setup, we are requiring that $L \gg \sigma$. Since the nearest pole is either $O(L)$ or $O(T^{-1})$ away, the Gaussian window function, whose width is much smaller than either L or T^{-1}, will regularize, within the integral, any reasonably truncated Taylor approximation to the Wightman function. This results in a valid asymptotic series for X in either case, as $T \to 0$. The integral in Eq. (4.3) can be done similarly by writing $D^+_T = D^+_0 + \Delta D^+_T$ (noting that the pole at $y = 0$ has been eliminated in ΔD^+_T) and calculating the temperature-dependent correction to Eq. (4.8). Numerical checks of particular cases verify that these approximations are valid. The results are presented in Figure 4.2.

4.4 Discussion of results

Several points are in order here. First, detectors see anything at all in the Minkowski vacuum case because the time-energy uncertainty relation, $\Delta t \Delta E \gtrsim \frac{1}{2}$, implies that a detector operating for a finite time has a nonzero probability A_0 of becoming excited, even when the field is in the vacuum state. Entanglement exists when virtual particle exchange dominates over local noise. When the magnitude of the exchange amplitude $|X_0|$ exceeds A_0, the detectors become entangled [91, 100]. Because of how both functions scale with Ω and L, in the vacuum case one can always reduce the local noise below $|X_0|$ by sufficiently increasing Ω. In the thermal and de Sitter cases, the local noise profile A fails to decrease fast enough for large Ω, resulting in a maximum entangling frequency for a given L, as well as a maximum separation beyond which entanglement is impossible, regardless of Ω.

What does this mean for our curious planetary inhabitants? Let's assume they have two satellites, with detectors of the sort we've been using, located on comoving trajectories as described above, with $\kappa^{-1} < L < 2\kappa^{-1}$ so that in the de Sitter case they would be outside of each other's cosmic horizon but within that of the home planet (so they can still send messages to it, as described in Figure 4.3). The satellites are programmed to interact the field locally with qubits having a resonant frequency that will lead to entanglement in the thermal case and to a separable state in the de Sitter case (e.g., the red star in Figure 4.2). After the interactions, they each run a local measurement protocol that implements one side of a test of Bell inequality violation, after which they send data back to the home planet for analysis. If thermality is a result of expansion, there will be no entanglement, but if it is a result of heating in flat spacetime, then the entanglement can be verified upon receipt of the transmissions from both satellites. Because this effect only manifests when the detectors pass

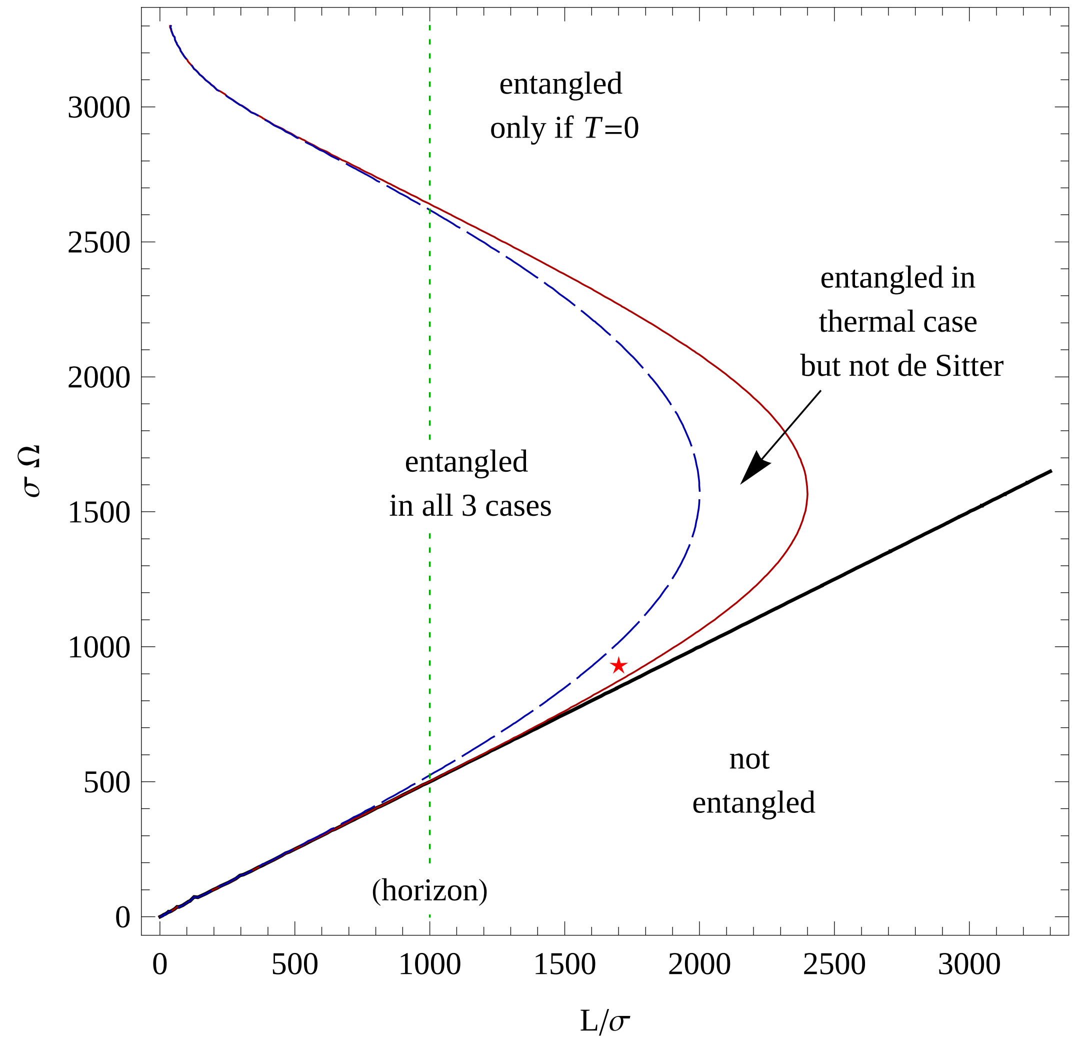


Figure 4.2: Entanglement profile for detector pairs in several universes—σ is detection time, Ω is detector resonance frequency, L is detector separation. The slanted black line is the entanglement cutoff in the Minkowski vacuum case (entangled above, separable below). The solid red curve is the thermal Minkowski cutoff, and the dashed blue curve is the de Sitter vacuum cutoff, both with perceived local temperatures satisfying $2\pi T = 10^{-3}\sigma^{-1}$. The de Sitter horizon distance ($10^3\sigma$) is given by the dotted green line. The red star indicates one particular detector setup that could be used to distinguish expansion from heating.

beyond each others' cosmic horizons (in the de Sitter case), a third party is required to make the determination.

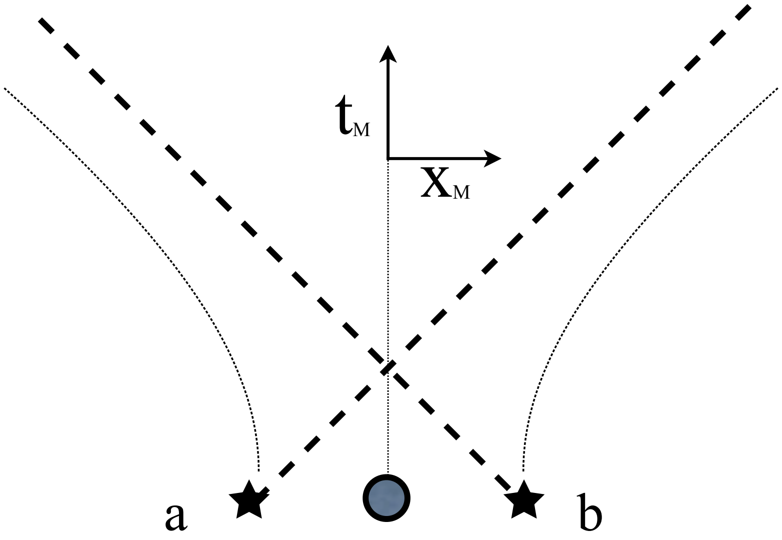


Figure 4.3: Spacetime diagram in Minkowski coordinates in the rest frame of the home planet (circle). Null rays travel at 45 degrees and light dotted lines represent geodesics in de Sitter space. Messages sent from detectors a or b ($\star$) never reach the other detector because of the Hubble expansion of the universe. However, the home planet can receive and analyze the messages, differentiating the entanglement scenarios depicted in Figure 4.2 .

We have demonstrated that while expansion and heating give rise to the same (thermal) signature in a single inertial particle detector, for certain choices of detector parameters, a heated field in flat spacetime is able to entangle detector pairs that the conformal vacuum in the associated de Sitter universe cannot. Thus, the universes can be distinguished by their *entangling power*. Two detectors are required and must be beyond each others' cosmic horizons (in the de Sitter case) to see the effect. Although, if present, the entanglement is exceedingly small, in principle its presence can always be determined by classical communication of local measurement data to a third party, as long as the verifier is able to receive messages from both detectors. These results are contrary to the intuition that "curvature generates entanglement" between field modes [84], since from it one would expect a *larger* entangled region in the de Sitter case. The ability of the field to swap its entanglement to local detectors is an operational question, though, and for this setup, the vacuum in a curved spacetime has less entangling power than a corresponding heated field in flat spacetime, even though both produce the same local detector response.

Bibliography

[1] Greg Ver Steeg and Nicolas C. Menicucci. Entangling power of an expanding universe. *Physical Review D (Particles, Fields, Gravitation, and Cosmology)*, 79(4):044027, 2009.

[2] Greg Ver Steeg and Stephanie Wehner. Relaxed uncertainty relations and information processing. *ArXiv Quantum Physics e-prints*, arXiv:0811.3771, November 2008. To appear in Quantum Information and Computation.

[3] S. Aaronson. Is Quantum Mechanics An Island In Theoryspace? *ArXiv Quantum Physics e-prints*, January 2004.

[4] T. Banks, M.E. Peskin, and L. Susskind. Difficulties for the evolution of pure states into mixed states. *Nuclear Physics B*, 244:125–134, 1984. quant-ph/0412028.

[5] W. van Dam. Impossible consequences of superstrong nonlocality. quant-ph/0501159, 2005.

[6] S. Popescu and D. Rohrlich. Quantum nonlocality as an axiom. *Foundations of Physics*, 24(3):379–385, 1994.

[7] L. Hardy. Quantum Theory From Five Reasonable Axioms. 2001.

[8] J. Preskill. Do Black Holes Destroy Information? In S. Kalara and D. V. Nanopoulos, editors, *Black Holes, Membranes, Wormholes and Superstrings*, 1993.

[9] B. F. Toner. *Quantifying quantum nonlocality.* PhD thesis, California Institute of Technology, 2006.

[10] John S. Bell. On the problem of hidden variables in quantum mechanics. *Rev. Mod. Phys.*, 38(3):447–452, Jul 1966.

[11] J. Clauser, M. Horne, A. Shimony, and R. Holt. Proposed experiment to test local hidden-variable theories. *Physical Review Letters*, 23:880–884, 1969.

[12] S. Popescu and D. Rohrlich. Quantum nonlocality as an axiom. *Foundations of Physics*, 24(3):379–385, 1994.

[13] B. Tsirelson (Cirel'son). Quantum generalizations of Bell's inequality. *Letters in Mathematical Physics*, 4:93–100, 1980.

[14] M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information.* Cambridge University Press, 2000.

[15] N. Gisin. Bell inequalities: many questions, a few answers. *ArXiv Quantum Physics e-prints*, February 2007.

[16] J. Barrett. Nonsequential positive-operator-valued measurements on entangled mixed states do not always violate a Bell inequality. *Physical Review A*, 65(4), April 2002.

[17] Sandu Popescu. Bell's inequalities and density matrices: Revealing "hidden" nonlocality. *Phys. Rev. Lett.*, 74(14):2619–2622, Apr 1995.

[18] J. Barrett. Information processing in generalized probabilistic theories. *Physical Review A*, 75(3), March 2007.

[19] A. Ambainis, A. Nayak, A. Ta-Shma, and U. Vazirani. Quantum dense coding and a lower bound for 1-way quantum finite automata. In *Proceedings of 31st ACM STOC*, pages 376–383, 1999. quant-ph/9804043.

[20] S. Aaaronson. The learnability of quantum states. *Royal Society of London Proceedings Series A*, 463:3089–3114, December 2007.

[21] A. Doherty, Y.-C. Liang, B. Toner, and S. Wehner. The quantum moment problem and bounds on entangled multi-provers games. In *Proceedings of the 23rd IEEE Conference on Computational Complexity*, pages 199–210, 2008.

[22] S. Wehner and A. Winter. Higher entropic uncertainty relations for anti-commuting observables. *Journal of Mathematical Physics*, 49:062105, 2008.

[23] J. S. Bell. On the Einstein-Podolsky-Rosen paradox. *Physics*, 1:195–200, 1965.

[24] B. Tsirelson (Cirel'son). Quantum analogues of Bell inequalities: The case of two spatially separated domains. *Journal of Soviet Mathematics*, 36:557–570, 1987.

[25] S. Popescu and D. Rohrlich. Nonlocality as an axiom for quantum theory. In *The dilemma of Einstein, Podolsky and Rosen, 60 years later: International symposium in honour of Nathan Rosen*, 1996.

[26] S. Popescu and D. Rohrlich. Causality and nonlocality as axioms for quantum mechanics. In *Proceedings of the Symposium of Causality and Locality in Modern Physics and Astronomy: Open Questions and Possible Solutions*, 1997.

[27] G. Brassard, H. Buhrman, N. Linden, A. Methot, A. Tapp, and F. Unger. A limit on nonlocality in any world in which communication complexity is not trivial. *Physical Review Letters*, 96:250401, 2006.

[28] M. Forster and S. Wolf. The universality of non-local boxes. In *Proceedings of QCMC*, 2008.

[29] H. Buhrman, M. Christandl, F. Unger, S. Wehner, and A. Winter. Implications of superstrong nonlocality for cryptography. *Proceedings of the Royal Society A*, 462(2071):1919–1932, 2006. quant-ph/0504133.

[30] G. M. D'Ariano. Probabilistic theories: what is special about quantum mechanics, 2008.

[31] L. Masanes, A. Acin, and N. Gisin. General properties of nonsignaling theories. *Physical Review A*, 73(1), January 2006.

[32] H. Barnum, J. Barrett, M. Leifer, and A. Wilce. Generalized No-Broadcasting Theorem. *Physical Review Letters*, 99(24), December 2007.

[33] H. Barnum, J. Barrett, M. Leifer, and A. Wilce. Teleportation in general probabilistic theories, 2008.

[34] Gen Kimura, Takayuki Miyadera, and Hideki Imai. Optimal state discrimination in generic probability models, 2008.

[35] B. Toner and F. Verstraete. Monogamy of bell correlations and tsirelson's bound, 2006. quant-ph/0611001.

[36] B. Toner and F. Verstraete. Monogamy of bell correlations and tsirelson's bound, 2006. quant-ph/0611001.

[37] M. Seevinck and J. Uffink. Local commutativity versus bell-inequality violation for entangled states and versus non-violation for separable states. *Physical Review A*, 76:042105, 2007.

[38] I. Damgaard, S. Fehr, L. Salvail, and C. Schaffner. Cryptography in the Bounded Quantum-Storage Model. In *Proceedings of 46th IEEE FOCS*, pages 449–458, 2005.

[39] I. Damgård, S. Fehr, R. Renner, L. Salvail, and C. Schaffner. A tight high-order entropic quantum uncertainty relation with applications. In *Advances in Cryptology—CRYPTO '07*, volume 4622 of *Lecture Notes in Computer Science*, pages 360–378. Springer-Verlag, 2007.

[40] S. Wehner and J. Wullschleger. Composable security in the bounded-quantum-storage model. In *ICALP 2008*, pages 604–615, 2008.

[41] S. Wolf. Personal communication. 2008.

[42] H. Barnum, O. Dahlsten, M. Leifer, and B. Toner. Nonclassicality without entanglement enables bit commitment. arXiv:0803.1264, 2008.

[43] G. Kimura. The bloch vector for n-level systems. *Physical Review A*, 315:339, 2003.

[44] R. A. Bertlmann and P. Krammer. Bloch vectors for qudits. *Journal of Physics A: Math. Theor.*, 41:235303, 2008.

[45] K. Dietz. Generalized bloch spheres for m-qubit states. *Journal of Physics A: Math. Gen.*, 36(6):1433–1447, 2006.

[46] S. Wehner. Unpublished note. 2008.

[47] S.J. Summers. On the independence of local algebras in quantum field theory. *Reviews in Mathematical Physics*, 2(2):201–247, 1990.

[48] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical report, Dept. of Statistics, September 2003.

[49] A. Peres. *Quantum Theory: Concepts and Methods.* Kluwer Academic Publishers, 1993.

[50] S. Bravyi and A. Kitaev. Universal quantum computation with ideal clifford gates and noisy ancillas. *Physical Review A*, 71:022316, 2005.

[51] G. Ver Steeg and S. Wehner. Monogamy of non-local correlations from an uncertainty relation. In preparation, 2008.

[52] A. Ambainis, A. Nayak, A. Ta-Shma, and U. Vazirani. Dense quantum coding and a lower bound for 1-way quantum automata. In *Proceedings of STOC '99*, pages 376–383, New York, NY, USA, 1999. ACM.

[53] R. de Wolf. Quantum communication and complexity. *Theoretical computer science*, 287(1):337–353, 2002.

[54] I. Kerenidis and R. de Wolf. Exponential lower bound for 2-query locally decodable codes via a quantum argument. In *Proceedings of 35th ACM STOC*, pages 106–115, 2003. quant-ph/0208062.

[55] M. Anthony and P. L. Bartlett. Function learning from interpolation. *Comb. Probab. Comput.*, 9(3):213–225, 2000.

[56] Pablo A. Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming*, 96, 2003.

[57] E. Kushilevitz and N. Nisan. *Communication Complexity.* Cambridge University Press, 1997.

[58] David Avis and Komei Fukuda. A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra. In *SCG '91: Proceedings of the seventh annual symposium on Computational geometry*, pages 98–104, New York, NY, USA, 1991. ACM.

[59] R.F. Werner and M.M. Wolf. All-multipartite bell-correlation inequalities for two dichotomic observables per site. *Physical Review A*, 64:032112, 2001.

[60] Itamar Pitowsky. Correlation polytopes: their geometry and complexity. *Math. Program.*, 50(3):395–414, 1991.

[61] I. Pitowsky and K. Svozil. Optimal tests of quantum nonlocality. *Physical Review A*, 64(1), July 2001.

[62] Arkadi Nemirovski. Lectures on modern convex optimization. In *Society for Industrial and Applied Mathematics*, 2001.

[63] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[64] Lieven Vandenberghe, Stephen Boyd, and Shao-Po Wu. Determinant maximization with linear matrix inequality constraints. *SIAM J. Matrix Anal. Appl.*, 19(2):499–533, 1998.

[65] Peng Sun and Robert M. Freund. Computation of minimum-volume covering ellipsoids. *Oper. Res.*, 52(5):690–706, 2004.

[66] Fuzhen Zhang. *The Schur complement and its applications.* Springer, 2005.

[67] J. Lfberg. Yalmip: A toolbox for modeling and optimization in MATLAB. In *Proceedings of the CACSD Conference*, Taipei, Taiwan, 2004.

[68] J. Sturm and AdvOL. SeDuMi. http://sedumi.mcmaster.ca/.

[69] Nicolas Brunner, Valerio Scarani, and Nicolas Gisin. Bell-type inequalities for nonlocal resources. *Journal of Mathematical Physics*, 47(11):112101, 2006.

[70] Alessandro Magnani. Tractable fitting with convex polynomials via sum-of-squares.

[71] J. Pearl. *Causality.* Causality, by Judea Pearl, pp. 400. ISBN 0521773628. Cambridge, UK: Cambridge University Press, March 2000., March 2000.

[72] David Casasent and Sanjay Natarajan. A classifier neural net with complex-valued weights and square-law nonlinearities. *Neural Netw.*, 8(6):989–998, 1995.

[73] D. Geiger and C. Meek. Quantifier elimination for statistical problems. *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 1999.

[74] D. Geiger and C. Meek. Graphical models and exponential families. *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, 1998.

[75] Konrad Schmdgen. The k-moment problem for compact semi-algebraic sets. *Mathematische Annalen*, 289.

[76] Jacob D. Bekenstein. Black holes and entropy. *Phys. Rev. D*, 7(8):2333–2346, Apr 1973.

[77] S. W. Hawking. Particle creation by black holes. *Communications in Mathematical Physics*, 43(3):199–220, 1975.

[78] G. W. Gibbons and S. W. Hawking. Cosmological event horizons, thermodynamics, and particle creation. *Phys. Rev. D*, 15(10):2738–2751, May 1977.

[79] Raphael Bousso. The holographic principle. *Rev. Mod. Phys.*, 74(3):825–874, Aug 2002.

[80] Stephen Hawking, Juan Maldacena, and Andrew Strominger. Desitter entropy, quantum entanglement and ads/cft. *JHEP*, 0105:001, 2001.

[81] Naureen Goheer, Matthew Kleban, and Leonard Susskind. The trouble with de sitter space. *JHEP*, 0307:056, 2003.

[82] Raphael Bousso. Adventures in de sitter space. 2002.

[83] Stephen J. Summers and Reinhard Werner. Maximal violation of bell's inequalities is generic in quantum field theory. *Communications in Mathematical Physics*, 110(2):247–259, June 1987.

[84] Jonathan L. Ball, Ivette Fuentes-Schuller, and Frederic P. Schuller. Entanglement in an expanding spacetime. *Phys. Lett. A*, 359(6):550–554, December 2006.

[85] N. D. Birrell and P. C. W. Davies. *Quantum Field Thoery in Curved Space.* Cambridge, 1982.

[86] M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information.* Cambridge University Press, Cambridge, UK, 2000.

[87] Charles H. Bennett, Gilles Brassard, Claude Crépeau, Richard Jozsa, Asher Peres, and William K. Wootters. Teleporting an unknown quantum state via dual classical and einstein-podolsky-rosen channels. *Phys. Rev. Lett.*, 70(13):1895–1899, Mar 1993.

[88] Hoi-Kwong Lo and Norbert Lutkenhaus. Quantum cryptography: from theory to practice. 2007.

[89] Lluis Masanes, Yeong-Cherng Liang, and Andrew C. Doherty. All bipartite entangled states display some hidden nonlocality. 2007.

[90] Robin Blume-Kohout, Carlton Caves, and Ivan Deutsch. Climbing mount scalable: Physical resource requirements for a scalable quantum computer. *Foundations of Physics*, 32(11):1641–1670, November 2002.

[91] Benni Reznik, Alex Retzker, and Jonathan Silman. Violating bell's inequalities in vacuum. *Physical Review A (Atomic, Molecular, and Optical Physics)*, 71(4):042104, 2005.

[92] Serge Massar and Philippe Spindel. Einstein-podolsky-rosen correlations between two uniformly accelerated oscillators. *Physical Review D (Particles, Fields, Gravitation, and Cosmology)*, 74(8):085031, 2006.

[93] Daniel Braun. Entanglement from thermal blackbody radiation. *Physical Review A (Atomic, Molecular, and Optical Physics)*, 72(6):062324, 2005.

[94] Marcus Spradlin, Andrew Strominger, and Anastasia Volovich. Les houches lectures on de sitter space. 2001.

[95] Alan H. Guth. *The Inflationary Universe: The Quest for a New Theory of Cosmic Origins.* Basic Books, 1997.

[96] B. S. DeWitt. In *General Relativity, An Einstein Centenary Survey.* Cambridge, 1979.

[97] Lluís Masanes. Asymptotic violation of bell inequalities and distillability. *Physical Review Letters*, 97(5):050503, 2006.

[98] L Sriramkumar and T Padmanabhan. Finite-time response of inertial and uniformly accelerated unruh - dewitt detectors. *Classical and Quantum Gravity*, 13(8):2061–2079, 1996.

[99] G. Vidal and R. F. Werner. Computable measure of entanglement. *Phys. Rev. A*, 65(3):032314, Feb 2002.

[100] Michal Horodecki, Pawel Horodecki, and Ryszard Horodecki. Separability of mixed states: necessary and sufficient conditions. *Physics Letters A*, 223(1-2):1–8, November 1996.

[101] Michał Horodecki, Paweł Horodecki, and Ryszard Horodecki. Inseparable two spin- $\frac{1}{2}$ density matrices can be distilled to a singlet form. *Phys. Rev. Lett.*, 78(4):574–577, Jan 1997.

[102] Michael E. Peskin and Daniel V. Schroeder. *An Introduction to Quantum Field Theory.* Perseus, Cambridge, Massachusetts, 1995.

[103] H. Arthur Weldon. Thermal green functions in coordinate space for massless particles of any spin. *Phys. Rev. D*, 62(5):056010, Aug 2000.