A Mathematical Foundation for Locality

AN ABSTRACT
SUBMITTED ON THE THIRD DAY OF JUNE, 2014
TO THE DEPARTMENT OF MATHEMATICS
OF THE SCHOOL OF SCIENCE AND ENGINEERING OF
TULANE UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
BY

Peter Bierhorst

Peter L. Bierhorst

APPROVED:

Michael Mislove, ph.d.
chairman

Mahir Can, ph.d.

Gustavo Didier, ph.d.

Lev Kaplan, ph.d.

Michelle Lacey, ph.d.

Abstract

This work is motivated by two non-intuitive predictions of Quantum Mechanics: *non-locality* and *contextuality*. Non-locality is a phenomenon whereby interactions between spatially separated objects appear to be occurring faster than the speed of light. Contextuality is a phenomenon whereby the outcome of a measurement cannot be interpreted as the revelation of an intrinsic fixed property of the system being measured, but instead necessarily depends on the configuration of the measurement apparatus.

Quantum Mechanics predicts non-local behavior in certain types of experiments collectively known as Bell tests. However, ruling out all possible alternative local theories is a subtle and demanding task. In this work, we lay out a mathematically-rigorous framework for analyzing Bell experiments. Using this framework, we derive the famous Clauser-Horne-Shimony-Holt (CHSH) inequality, an important constraint that is obeyed by all local theories and violated by Quantum Mechanics. We further demonstrate how to analyze the data of a CHSH experiment without assuming that successive experimental trials are independent and/or identically distributed.

We also derive the Clauser-Horne (CH74) inequality, an inequality that is more well-suited for realistic Bell experiments using photons. We demonstrate a robust method for statistically analyzing the data of a CH74 experiment, and show how to calculate exact p-values for this analysis, improving on the previously-best-known (loose) upper bounds obtained from Hoeffding-style inequalities.

The work concludes with an exploration of contextuality. The Kochen-Specker theorem – a result demonstrating the contextual nature of Quantum Mechanics – is applied to resolve a conjecture in Domain Theory regarding the spectral order on quantum states.

A Mathematical Foundation for Locality

A DISSERTATION
SUBMITTED ON THE THIRD DAY OF JUNE, 2014
TO THE DEPARTMENT OF MATHEMATICS
OF THE SCHOOL OF SCIENCE AND ENGINEERING OF
TULANE UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
BY

Peter Bierhorst
PETER L. BIERHORST

APPROVED:

MICHAEL MISLOVE, PH.D.
CHAIRMAN

MAHIR CAN, PH.D.

Gustavo Didier
GUSTAVO DIDIER, PH.D.

LEV KAPLAN, PH.D.

MICHELLE LACEY, PH.D.

Acknowledgments

I would like to graciously acknowledge the help of my thesis director, Mike Mislove, whose support, guidance, and enthusiasm have been invaluable in writing this dissertation. I also want to thank Gustavo Didier, Lev Kaplan, and Keye Martin for their important contributions at various stages of the research process, and the Tulane mathematics department for its supportive environment. Finally, I would like to thank my parents, Gene and Susan, and my wife Lauren, for believing in me throughout.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Background

1.1 Locality and Non-Contextuality: An Overview

This work is motivated by non-intuitive predictions of Quantum Mechanics, a well-supported theory of the physical world. Quantum mechanics predicts *non-locality*, a phenomenon in which two spatially separated systems can display correlations that seem to require the systems to be interacting with each other instantaneously. Definitively ruling out the possibility that these correlations are explainable by less exotic means – such as the possibility that the correlations can be attributed to signals travelling between the systems at the speed of light – requires subtle arguments, which we will explore at length in this work. We will also explore another prediction of Quantum Mechanics known as *contextuality*. This phenomenon, in contrast to non-locality, can be exhibited by a single system, so long as the system can be subjected to multiple distinct measurements. The system displays contextuality if the outcome of the measurement of a particular property cannot be interpreted as the revelation of an intrinsic fixed property of the system, but instead necessarily depends on the configuration of the measurement. In this work, the proof of quantum contextuality will be applied to another area of Quantum Mechanics, known as Quantum Information Theory, with interesting results.

Non-locality and contextuality are not simple notions, and a longer exploration of these ideas is merited. We start by examining quantum non-locality, the quantum prediction that certain pairs of spatially separated systems can display correlations that cannot be explained solely by events in their shared history.

It should be pointed out that Quantum Mechanics is not the first physical theory to predict non-local behavior. Indeed, the original theory of gravitation due to Newton was a non-local theory. The non-locality of Newtonian gravity is not difficult to grasp, and so we will examine it as a good starting point towards understanding the more subtle notion of quantum non-locality.

Under Newtonian gravity, two pieces of matter attract each other with a force inversely proportional to the square of the distance between them. Importantly, this force is felt *instantaneously* – so if one piece of matter moves to the side, the direction of the force it exerts changes without delay, no matter how large the distance between the two objects.

In principle, such behavior could be used to send signals instantaneously across unlimited spatial distances. To see how this would work, consider the scheme of Figure 1.1 where an observer is monitoring the gravity exerted by two distant spheres of equal mass that are initially attached to each other. Suppose then that a small explosion breaks the two spheres apart so that one sphere is

now travelling towards the observer and the other is travelling away from the observer:

Figure 1.1: Sending an instantaneous signal via Newton's Law of Gravitation

As soon as the configuration of the spheres changes, the total gravity felt by the observer in the direction of the spheres changes as well. This can be worked out directly from Newton's law of gravitation,

$$F = \frac{Gm_o m_{w_i}}{d_i^2},$$

where F is the force of gravity, G is a numerical constant, m_o is the mass of the observer, m_{w_i} is the mass of sphere i (for simplicity, assume that $m_{w_1} = m_{w_2}$, and d_i is the distance from the observer to the center of sphere i. The total force felt in the direction of the two spheres is the sum of the two individual forces F that each sphere generates. Thus we can calculate that

$$F_{\text{before}} = \frac{Gm_o m_{w_1}}{d_1^2} + \frac{Gm_o m_{w_2}}{d_2^2} \qquad F_{\text{after}} = \frac{Gm_o m_{w_1}}{(d_1 - \epsilon)^2} + \frac{Gm_o m_{w_2}}{(d_2 + \epsilon)^2}, \tag{1.1}$$

where ϵ is a small distance that the spheres traverse immediately after the explosion. It is easy to check that in general, the two forces in (1.1) are not equal. Therefore the observer can tell exactly when the explosion occurs, if he has a sufficiently sensitive measurement device to detect small changes in gravity.

In practice, such an experiment would be hard to perform, as gravity is a weak force and small changes in gravity are hard to detect. In the meantime, Newton's law of gravity has been superseded by General Relativity, which makes more accurate predictions and does not predict nonlocal behavior. According to General Relativity, in a scheme like that of Figure 1.1 the change in gravity is not felt instantaneously, but instead is propagated at the speed of light. If the observer in Figure 1.1 is a light year away from the two spheres, he will not become aware of the explosion until a year has elapsed, no matter how sensitive his gravity-detecting apparatus.

While our best description of gravitation, General Relativity, is local, our best description

of subatomic particles, Quantum Mechanics, is not. The non-locality of Quantum Mechanics (or Quantum Field Theory, if we want to study electromagnetic interactions) is more subtle than the scheme of Figure 1.1, and does not allow for the sending of signals faster than the speed of light. Quantum non-locality predicts only that certain spatially separate random processes will display correlations that exceed what would be possible if the processes were wholly separable.

The fundamental manifestation of quantum non-locality is the violation of Bell inequalities, named after John Bell, who first showed that Quantum Mechanics predicts non-local behavior in 1964 [1]. The most important Bell inequality is arguably the Clauser-Horne-Shimony-Holt (CHSH) inequality, first derived in [2], a constraint on the behavior we would expect to see from two processes that are not communicating with each other.

To introduce the setting of the CHSH inequality, suppose D_1 and D_2 are two random binary-output processes occurring at great distance from each other. It is useful to think of the processes as "black boxes" whose internal workings we do not attempt to describe, and the processes randomly generate one of two outcomes, which we label "+1" or "−1," repeatedly at fixed intervals. Suppose additionally that each black box has an external dial on it with two settings, so the D_1 dial can be set to either a or a', and the D_2 dial can be set to either b or b'. It is possible that the dial can affect the distribution of various outcomes of the process, but the outcome is always still either +1 or −1. A scheme depicting this setup is given in Figure 1.2.

Figure 1.2: The D_1 and D_2 processes. D_1 and D_2 are separated by a large distance.

To formulate the CHSH inequality, we need to use terms of form $E_{a,b}(D_1D_2)$, which we define as the expected value of the product of the outcome of D_1 and D_2, given that the setting at D_1 is a and the setting at D_2 is b. This will necessarily be some number between −1 and +1. The CHSH inequality then states that if the processes D_1 and D_2 are governed by a local theory, the

following must hold:

$$E_{ab}(D_1D_2) - E_{a'b}(D_1D_2) + E_{ab'}(D_1D_2) + E_{a'b'}(D_1D_2) \leq 2. \tag{1.2}$$

Now, it is clear that just from the mathematical definitions, the quantity in (1.2) must be bounded above by four. To understand where the bound two comes from, it will help to examine just one term, say $E_{ab}(D_1D_2)$, so for now take the settings to always be a and b. Suppose first that D_1 outputs $+1$ and -1 with equal probability. Then suppose that D_2 also outputs $+1$ and -1 with equal probability, but furthermore D_1 and D_2 always give the same result. Then D_1D_2 will always be $+1$, so $E_{ab}(D_1D_2) = 1$.

This type of correlation between D_1 and D_2 implies that the two systems are connected in some way. One possibile explanation is that the two detectors are instantaneously communicating superluminally to align the $+1$ and -1 counts, but we need not rely on action at a distance to explain the correlation. Perhaps all that is going on inside the "black boxes" is that both system D_1 and D_2 are detecting some third random process, call it λ, that is spatially located between D_1 and D_2 and is emitting a stream of random $+1$ and -1 signals. D_1 and D_2 merely report these signals as they receive them. So if D_1 and D_2 are separated by one light year and λ is directly between them emitting signals at the speed of light, then D_1 and D_2 are reporting what happened at λ six months earlier, and the D_1/D_2 correlation is explained by a local theory.

In short, we have described a "hidden variable" λ that is sending the following two signals to D_1 and D_2, with equal probability:

If D_1 is set to a, output $+1$. If D_2 is set to b, output $+1$.

If D_1 is set to a, output -1. If D_2 is set to b, output -1.

Now, this "instruction set" only refers to two measurement settings, so it is incomplete. But perhaps λ could be sending a complete set of instructions, such as the following example:

If D_1 is a, output $+1$. If D_1 is a', output -1. If D_2 is b, output $+1$. If D_2 is b', output $+1$. (1.3)

If λ always sends the above signal, we can see that $E_{ab}(D_1D_2) = +1$, $E_{a'b}(D_1D_2) = -1$, $E_{ab'}(D_1D_2) = +1$, and $E_{a'b'}(D_1D_2) = -1$, so the left-hand side of (1.2) is less than or equal

to 2, and the CHSH inequality is obeyed.

There are sixteen possible assignments of $+1$ and -1 like the one in (1.3), and it is straightforward to check that the CHSH inequality holds for each of the sixteen assignments. This means if λ were to set the values of D_1 and D_2 by probabilistically sampling from among these sixteen assignments, the CHSH inequality would have to be satisfied, no matter the distribution of λ. So if the CHSH inequality is *violated*, we could rule out a theory whereby a process λ is governing the outcomes of D_1 and D_2 by randomly sampling from the set of instructions like (1.3).

Indeed, Quantum Mechanics predicts a violation of the CHSH inequality in certain experimental configurations. However, many issues remain to be examined in this work. For instance, what is the precise meaning of a *local theory* – is there a more general definition than "a random process λ that sends deterministic instruction sets to both detectors?" The first four sections of Chapter 2 introduce and discuss a mathematically rigorous framework for formulating a more general definition and analyzing CHSH locality experiments. The rest of Chapter 2 demonstrates how to draw rigorous conclusions from an experimental violation of the CHSH inequality without ignoring frequently-overlooked possibilities, such as local theories that vary over time with dependence on the outcomes of earlier trials. Chapter 3 generalizes these methods to account for experimental imperfections that are unavoidable in any real-world test of quantum non-locality. Recent experimental results in the field are also discussed, and it turns out that to date no experiment has been able to definitively rule out all possible local theories. If an experiment that is capable of ruling out all possible local theories is successfully performed, the results of Chapter 3 provide a clear roadmap for how to analyze the experimental results and draw statistical conclusions. Currently, multiple experimental teams are attempting to successfully perform such an experiment, and some experts in the field believe that it is only a matter of few years before this will be achieved [3].

This work is also concerned with the notion of quantum contextuality. Our initial description of contextuality – that a system is contextual if measurement outcomes depend on the configuration of the measurement – is quite vague. To get a better understanding of contextuality, we can study a simplified model of a quantum measurement that encodes all the key features of the proof of quantum contextuality.

Suppose that the object being measured is a regular two-dimensional sphere. We are interested in the color of the sphere at various points, which can be either red or green. Suppose furthermore that we are unable to look at the whole sphere at any given time, but can only choose a certain set of points to examine, and such an examination can only be performed once. The set of

points that that we can choose to look at must consist of three points that are mutually orthogonal: that is, if we take the object to be the unit sphere S^2 in $\mathbb{R}^3$, then we can examine the points corresponding to the tips of three unit vectors $\vec{v}_1$, $\vec{v}_2$, $\vec{v}_3$ if and only if $\langle \vec{v}_1, \vec{v}_2 \rangle = \langle \vec{v}_2, \vec{v}_3 \rangle = \langle \vec{v}_3, \vec{v}_1 \rangle = 0$, where $\langle \vec{v}_i, \vec{v}_j \rangle$ is the dot product of $\vec{v}_i$ and $\vec{v}_j$. As a final constraint, suppose that we know in advance, based on our physical theory for measuring the sphere, that we will always see exactly one red point and two green points for any choice of $\vec{v}_1$, $\vec{v}_2$, and $\vec{v}_3$.

Now here is the question: if it is indeed true that we will always see exactly one red point and two green points, is there any fixed way in which the sphere could have been painted red and green prior to the measurement? That is, when we choose three points to examine, are we seeing the intrinsic color of these points, or does the sphere set the color of these points only upon the measurement?

It turns out that it is a geometric fact that there is no painting of S^2 that is compatible with this measurement scheme. If you paint some parts of a sphere red and the rest of it green, then no matter how you go about it, there will always be some orthogonal triple that either has all green points or two or more red points. This notion is made precise by the following theorem:

Theorem 1.1.1. *(Kochen-Specker Theorem) Consider the unit sphere S^2 in $\mathbb{R}^3$, consisting of vectors $\vec{v}$ for which $\langle \vec{v}, \vec{v} \rangle = 1$. Then there is no map $f : S^2 \to \{0, 1\}$ satisfying the following condition: for every orthogonal triple $\{\vec{x}_i\}_{i=1}^3$ of vectors in S^2 (that is, $i \neq j \Rightarrow \langle \vec{x}_i, \vec{x}_j \rangle = 0$), f maps exactly one of the x_i to 1, and the other two x_i to 0.*

Proof. This is a consequence of the original result of Kochen and Specker [4]. A more accessible proof that is easier to visualize can be found in [5]. □

Hence for the measurement scheme we have described, the color that is detected cannot be taken to have been a fixed, intrinsic property of the sphere that existed prior to the measurement.

While the above model of measurement is not quite a faithful description of any valid quantum measurement, there *are* valid quantum measurements that can be modeled by collections of three orthogonal vectors in $\mathbb{C}^3$. For such measurements, the outcome amounts to assigning one of two values x and y to each of the three chosen vectors, and for any choice of three orthogonal vectors, Quantum Mechanics predicts that exactly one of the vectors will be assigned the value x. Due to the way that $\mathbb{R}^3$ sits inside of $\mathbb{C}^3$, any fixed assignment of x and y to all the unit vectors of $\mathbb{C}^3$ compatible with this quantum requirement would imply the existence of a map that contradicts the result of Theorem 1.1.1, so no fixed assignment of x and y exists. Hence Quantum Mechanics

is contextual: measurement outcomes are not intrinsic properties, but rather can depend on the configuration of the measurement.

In Chapter 4, we will relate the Kochen-Specker theorem to the use of Domain Theory in the study of Quantum Information Theory. The Kochen-Specker theorem will be used to resolve a conjecture about the spectral order, a domain-theoretic order on quantum states that was introduced in [6]. The nature of this result implies that in a certain sense, the spectral order witnesses the contextuality of Quantum Mechanics. This raises some new questions that are also explored in Chapter 4.

Chapters 2 and 3 make extensive use of measure-theoretic probability, and Chapter 4 requires concepts from Domain Theory. The remaining sections of Chapter 1 give the necessary background in these fields, as well as Quantum Mechanics.

1.2 Quantum Mechanics

For the reader unfamiliar with Quantum Mechanics, the dissertation can still be read and the main points appreciated if the results of the quantum predictions are just taken as a given. Indeed, actual quantum calculations do not form a large portion of the work. However, only a basic familiarity with Quantum Mechanics is needed to understand the calculations that yield the relevant quantum predictions, and this section provides the necessary background. The presentation here is similar to that given in Chapter 2 of Nielsen and Chuang [7], a good reference for Quantum Mechanics in finite dimensional spaces. Some concepts from linear algebra are necessary, and we assume that the reader is familiar with notions such as vector spaces, linear maps, inner products, norms, and tensor products.

Definition 1.2.1. A quantum *state space* is a complex vector space with an inner product (known as a *Hilbert space*). A quantum *state vector* is a unit vector in the state space.

The state vector describes aspects of a quantum system. For instance, the polarization of a single photon or the "spin" of an electron can be described by a state vector. The state vector is harder to interpret than some notions from classical physics, such as, for instance, the representation of "position" with a vector in $\mathbb{R}^3$. However, the method by which the state vector translates into observable physical predictions via measurements is mathematically straightforward.

Some quantum systems reside in infinite-dimensional state spaces, but for the purposes of this dissertation, we will always take the state space to be finite-dimensional, and thus isomorphic

to $\mathbb{C}^n$ for some positive integer n. In mathematical writing, elements of $\mathbb{C}^n$ are generally denoted using vector notation, such as $\vec{v} \in \mathbb{C}^n$. However, in keeping with standard physics notation, we will use Dirac bra-ket notation, as given by the following definition.

Definition 1.2.2. (Dirac notation) State vectors (unit vectors in $\mathbb{C}^n$) are denoted $|v\rangle$, $|w\rangle$, etc. The inner product of $|v\rangle$ and $|w\rangle$ is denoted $\langle v|w\rangle$. The linear function $l : \mathbb{C}^n \to \mathbb{C}$ induced by $|v\rangle$ via the inner product – ie, $l(|w\rangle) = \langle v|w\rangle$ – is denoted by $\langle v|$. The tensor product of $|v\rangle$ and $|w\rangle$, which is an element of $\mathbb{C}^n \otimes \mathbb{C}^n \sim \mathbb{C}^{n^2}$, can be denoted as $|v\rangle|w\rangle$ or $|vw\rangle$.

A *projection map* is a familiar concept from linear algebra. For our purposes, we define a projection map to be a linear map $P : \mathbb{C}^n \to \mathbb{C}^n$ for which $P^2 = P$. Projection maps allow for the quantum definition of a *measurement*.

Definition 1.2.3. (Projective measurements) For a given state space $\mathbb{C}^n$, a quantum *projective measurement* is a collection of projection maps $\{P_1, ...P_m\}$ for which $\sum_{i=1}^{m} P_i = I_n$, where $I_n : \mathbb{C}^n \to \mathbb{C}^n$ is the identity map.

This definition merits some explanation. Suppose there is a quantum system represented by a state vector $|\phi\rangle \in \mathcal{C}^n$, and one wants to physically measure the system by way of a projection measurement, $\{P_1, ...P_m\}$. According to quantum mechanics, there are m possible outcomes, one for each of the projectors that constitute the measurement. A particular outcome will occur with a certain probability, given by the following formula:

$$P(\text{Outcome } i \text{ occurs}) = ||P_i|\phi\rangle||^2, \tag{1.4}$$

where $P_i|\phi\rangle$ is the application of the map P_i to $|\phi\rangle$, and $||\cdot||$ is the norm induced by the inner product. The experimenter performing the measurement will then see that outcome i has occurred, and after the measurement, the new state vector for the quantum system will be $\frac{P_i|\phi\rangle}{||P_i|\phi\rangle||}$.

This is still somewhat abstract. What happens in a real laboratory setting is that one has some sort of apparatus that performs the projective measurement, and the apparatus will generate a distinct macroscopic response for each of the m possible outcomes. One never observes the state vector $|\phi\rangle$ directly.

Let us consider a simple example. Take the state space to be $\mathbb{C}^2$, and let $|0\rangle$ and $|1\rangle$ represent a pair of orthogonal unit vectors in $\mathbb{C}^2$. Then a general unit vector can be written as $|\phi\rangle = \alpha|0\rangle + \beta|1\rangle$, where α and β are complex numbers for which $\overline{\alpha}\alpha + \overline{\beta}\beta = 1$. Now, consider

a projection measurement $\{P_1, P_2\}$ where P_1 is the projection map onto $|0\rangle$ (that is, the unique projection map for which $P_1|0\rangle = |0\rangle$ and $P_1|\phi\rangle = |\phi\rangle$ iff $|\phi\rangle = c|0\rangle$ for some $c \in \mathbb{C}$), and P_2 is the projection map onto $|1\rangle$. Then we have

$$P_1 : \mathbb{C}^2 \to \mathbb{C}^2, \qquad P_1|\phi\rangle = \alpha|0\rangle$$
$$P_2 : \mathbb{C}^2 \to \mathbb{C}^2, \qquad P_2|\phi\rangle = \beta|1\rangle.$$

By the rules governing projective measurements, outcome "1" will be seen with probability $\overline{\alpha}\alpha$ and outcome "2" will be seen with probability $\overline{\beta}\beta$. Of course, if the experiment is performed just once, only one particular outcome will be observed and the outcome cannot be predicted with certainty (unless $\overline{\alpha}\alpha$ happens to equal either 0 or 1) but is governed instead by a probability distribution.

A nice pictorial scheme for polarization of photons can be used to give some intuition for what might be going on. This scheme should not be taken as a definitive interpretation of the quantum mechanical description of what occurs on the microscopic level, but it is helpful. Referring to Figure 1.3, imagine the "Incoming Photon" to be a small wave packet travelling along in the direction of the arrow. The wave corresponds to the electric field component of electromagnetic radiation, and the polarization of the photon refers to the angle of the plane containing the wave. In the figure, we have labeled two axes $|0\rangle$ and $|1\rangle$, and the plane containing the photon wave appears to be about halfway between these two axes. So if we wanted to draw a unit vector parallel to this plane, the vector would be $|\phi\rangle = \frac{1}{\sqrt{2}}|0\rangle + \frac{1}{\sqrt{2}}|1\rangle$. Let's take this $|\phi\rangle$ to be the quantum state vector that describes the polarization of the photon, where other choices α and β in $\mathbb{R}$ would correspond to state vectors of different polarizations. (Unfortunately the scheme starts to break down if we worry about how to interpret general complex choices for α and β, so we don't consider these for now.)

Now, the figure shows that the photon is travelling towards a "Detector." This apparatus will perform the $\{P_1, P_2\}$ projective measurement on $|\phi\rangle$. In our scheme, we can think of the photon being "forced" to go through either the up-down slot, or the left-right slot. Mathematically, the photon will yield either outcome 1 or 2, and the probability of each is 50%. If the incoming photon is tilted a little more towards one axis – say, for instance, $\frac{1}{\sqrt{10}}|0\rangle + \frac{3}{\sqrt{10}}|1\rangle$ – then the probability distribution would put more than 50% of the weight on one of the outcomes.

Let's assume that outcome 1 occurs. Then according to the rules of projective measurements,

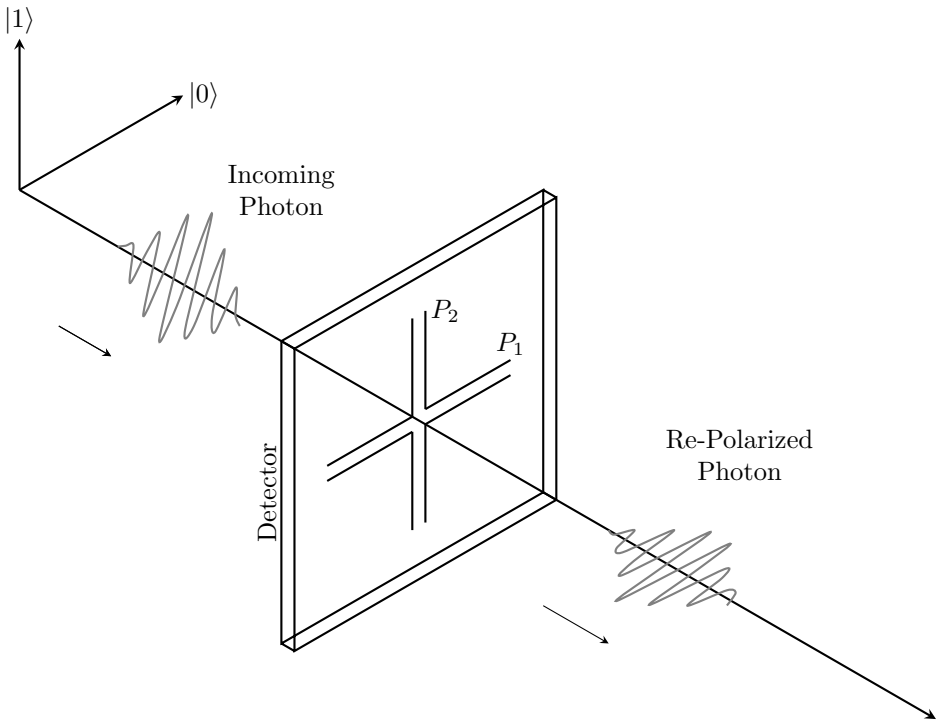

Figure 1.3: Scheme for photon measurement

the new state of the photon will be

$$\frac{P_1|\phi\rangle}{||P_1|\phi\rangle||} = \frac{\frac{1}{\sqrt{2}}|0\rangle}{\left|\left|\frac{1}{\sqrt{2}}|0\rangle\right|\right|} = |0\rangle.$$

This corresponds nicely to what is shown in the figure: the outcoming photon, after traveling through the left-right slot (outcome 1), is now polarized along the $|0\rangle$ axis.

In this experiment, it should be noted that the only thing the experimenter records is whether the photon went through the P_1 or the P_2 gate. One cannot "watch" a photon as it travels through space, but can only interact with it by way of a projection measurements. In a laboratory setting, the detector of Figure 1.3 would display some sort of macroscopic response that indicates which projection outcome occurs; other than that, nothing is directly witnessed by the experimenter.

As we see, the mathematics involved in measuring the polarizations of single photons is not especially difficult. The situation changes when we describe systems of multiple photons. This is because, according to the postulates of Quantum Mechanics, systems of multiple quantum states are described by state vectors that exist in the *tensor product* of the respective state spaces. Formally, if two (separate) copies of $\mathbb{C}^2$ are the individual state spaces for two quantum systems, then the state

space of the joint system will be given by $\mathbb{C}^2 \otimes \mathbb{C}^2 = \mathbb{C}^4$. If $|\phi\rangle$ and $|\psi\rangle$ represent the polarizations of two different photons that are not interacting with each other, then the state of the joint system is $|\phi\rangle \otimes |\psi\rangle = |\phi\psi\rangle \in \mathbb{C}^2 \otimes \mathbb{C}^2$. This is known as a *product state* because the joint state of the two photons is the tensor product of two individual state vectors. However, the state vector of two photons is not limited to product states, and there are other unit vectors in $\mathbb{C}^2 \otimes \mathbb{C}^2$ that are valid state vectors that cannot be decomposed into the product of two state vectors in $\mathbb{C}^2$. Such states are known as *entangled states*, and these states can always be expressed as unit-length linear combinations of product states. One such state in particular that we will be using is the *singlet state* in $\mathbb{C}^2 \otimes \mathbb{C}^2$, given by

$$\frac{|01\rangle - |10\rangle}{\sqrt{2}}. \tag{1.5}$$

This state has interesting properties when measured, as we will see in Chapter 2.

The last quantum notion that we introduce is that of the *density operator*, an alternate characterization of quantum states that will be useful in Chapter 4. If $|\phi\rangle$ is the state vector of a quantum system, then the corresponding density operator representation of this quantum system is defined to be the projection map P_ϕ onto $|\phi\rangle$, where P_ϕ is usually denoted $|\phi\rangle\langle\phi|$. Density operators can also be used to describe a probability distribution over many states. If a quantum state is not known with certainty but is instead thought to be in one of the states in the collection $\{|\phi\rangle\}_{i=1}^n$ with the probability of being in state $|\phi_i\rangle$ given by p_i, then the density operator representation of this "mixed state" is given by

$$\sum_{i=1}^{n} p_i |\phi_i\rangle\langle\phi_i|. \tag{1.6}$$

It can be shown that an operator can be expressed in the form of equation (1.6) if and only if the operator is a self-adjoint positive operator with trace equal to one. Hence the collection of self-adjoint, positive, trace-one operators can be used to represent the collection of probabilistic mixtures of quantum states for a given n-dimensional state space.

1.3 Measure-Theoretic Probability

As is clear from the previous section, Quantum Mechanics is inherently a probabilistic theory. To do work in this field, we need to be familiar with the mathematical theory of probability. We will work in the most general framework for probability theory: the *measure-theoretic* setting. Thus the results in this dissertation will hold as widely as possible. Indeed, on a few occasions we will find ourselves re-proving familiar results, such as the CHSH inequality, as the original proofs

in physics papers tend to tacitly make simplifying assumptions about the nature of the probability spaces being used.

To introduce the concept of a measure-theoretic probability space, we require the notions of a *σ-algebra* and a *measure*.

Definition 1.3.1. Let T be a set. A collection of subsets of T, denoted $\mathcal{T}$, is a *σ-algebra* (over T) if the following three conditions hold:

1. Both $\emptyset$ and T belong to $\mathcal{T}$.
2. For any set A belonging to $\mathcal{T}$, the complement $A^C = T \backslash A$ also belongs to $\mathcal{T}$.
3. If $\{A_i\}_{i=1}^{\infty}$ is a sequence of sets for which each A_i is in $\mathcal{T}$, then $\cup_{i=1}^{\infty} A_i$ also belongs to $\mathcal{T}$.

The intersection of a collection of σ-algebras is again a σ-algebra. Hence we can speak of the *smallest* σ-algebra containing a collection of sets $\mathcal{S}$ by taking the intersection of all σ-algebras that include $\mathcal{S}$. This is known as the σ-algebra *generated by* $\mathcal{S}$. This important notion is used to define the *product σ-algebra*: if $\mathcal{T}$ and $\mathcal{U}$ are σ-algebras over T and U, then the product σ-algebra over $T \times U$, denoted $\mathcal{T} \otimes \mathcal{U}$, is defined to be the σ-algebra generated by the collection $\{A \times B \mid A \in \mathcal{T}, B \in \mathcal{U}\}$. The concept of product σ-algebras is used in Section 2.4.

Two well-known σ-algebras over $\mathbb{R}$ are the *Borel σ-algebra* and the collection of *Lebesgue-measurable sets*. The definitions of these σ-algebras can be found in, for example, Bartle [8].

Definition 1.3.2. Let $\mathcal{T}$ be a σ-algebra over the set T. Then a function $f : \mathcal{T} \to \mathbb{R}$ is a *measure* (on $\mathcal{T}$) if the following conditions hold:

1. $f(\emptyset) = 0$.
2. $f(A) \geq 0$ for all $A \in \mathcal{T}$.
3. If $\{A_i\}_{i=1}^{\infty}$ is a disjoint sequence of sets in $\mathcal{T}$, then $f(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} f(A_i)$.

Furthermore, f is a *probability measure* if, in addition to the conditions above, $f(T) = 1$.

Definition 1.3.3. A *probability space* is a triple $(\Omega, \mathcal{F}, P)$, where Ω is a set, $\mathcal{F}$ is a σ-algebra over Ω, and P is a probability measure on $\mathcal{F}$. Elements of $\mathcal{F}$ are called *events*, and we refer to $P(A)$ as *the probability that event A occurs.*

In this dissertation, we will use many probabilistic notions, such as independence, conditional expectation, etc. Many of these notions rely on the concept of a *measurable function.*

Definition 1.3.4. If f is a function from T to $\mathbb{R}$, and $\mathcal{T}$ is a σ-algebra over T, then f is *measurable with respect to* $\mathcal{T}$, or $\mathcal{T}$*-measurable*, if

$$\forall \alpha \in \mathbb{R}, \qquad f^{-1}(\alpha, \infty) \in \mathcal{T}. \tag{1.7}$$

More generally, if $f : T \to U$ maps to a set U with σ-algebra $\mathcal{U}$, we say that f is measurable if

$$\forall B \in \mathcal{U}, \qquad f^{-1}(B) \in \mathcal{T}. \tag{1.8}$$

If $\mathcal{U}$ is the Borel σ-algebra over $U = \mathbb{R}$, then (1.8) is equivalent to (1.7). A *random variable* is a measurable function from Ω to U when $(\Omega, \mathcal{F}, P)$ is a probability space. The set U is often $\mathbb{R}$ equipped with the Borel σ-algebra, but for this work we will use the more general definition of a random variable that allows for a general target space U.

For a random variable $\phi : \Omega \to U$, the *σ-algebra generated by ϕ*, denoted by $\sigma(\phi)$, is defined to be the σ-algebra generated by the collection of sets $\{\phi^{-1}(B) \mid B \in \mathcal{U}\}$. By definition, $\sigma(\phi)$ will always be a sub-σ-algebra of $\mathcal{F}$, and it is possible for $\sigma(\phi)$ to be equal to $\mathcal{F}$.

One simple example of a random variable is the *Bernoulli random variable*. This is a random variable $B : \Omega \to \{a, b\}$ taking values in a binary set such that $P(\{\omega \in \Omega \mid B(\omega) = a\})$ and $P(\{\omega \in \Omega \mid B(\omega) = b\})$ are both nonzero. By convention, an event of the form $\{\omega \in \Omega \mid B(\omega) = x\}$ can be notated $\{B = x\}$, and the probability of this event can be written $P(B = x)$, referred to as "the probability that B equals x." It is easy to check that the σ-algebra $\sigma(B)$ is $\{\emptyset, B^{-1}(a), B^{-1}(b), \Omega\}$. Bernoulli random variables appear frequently throughout Chapters 2 and 3. Bernoulli random variables can also be related to the important concept of the *simple function*. Generally, $\phi : \Omega \to \mathbb{R}$ is defined to be a *simple function* if ϕ is measurable and it only takes finitely many values. This implies that any simple function can be expressed in the following manner:

$$\phi = \sum_{i=1}^{n} c_i I_{S_i}, \tag{1.9}$$

where $\{c_i\}_{i=1}^n$ is a collection of real numbers, $\{S_i\}_{i=1}^n$ is a collection of sets in $\mathcal{F}$, and I_X is the indicator function:

$$I_X(x) = \begin{cases} 1 & \text{if } x \in X, \\ 0 & \text{if } x \notin X. \end{cases} \tag{1.10}$$

As an illustrative example, if a and b are real numbers, then a Bernoulli random variable $B : \Omega \to \{a, b\}$ will be a simple function, and can be represented as $aI_{\{B=a\}} + bI_{\{B=b\}}$. Simple functions are required for the Lebesgue definition of the integral, which is used extensively in this dissertation.

Definition 1.3.5. (Lebesgue integral) Let Ω be a measure space equipped with a σ-algebra $\mathcal{F}$ and a measure $P : \mathcal{F} \to \mathbb{R}$. If $f : \Omega \to \mathbb{R}$ is a nonnegative measurable function, than the *integral of f with respect to P* is defined to be

$$\int f dP := \sup_{\phi} \sum_{i=1}^{n} c_i P(S_i),$$

where the supremum is taken over all simple functions $\phi = \sum_{i=1}^{n} c_i I_{S_i}$ satisfying $\phi(\omega) \leq f(\omega)$ for all $\omega \in \Omega$. If $g : \Omega \to \mathbb{R}$ is a (general) measureable function, then the integral of g with respect to P is defined to be

$$\int g dP := \int g^+ dP - \int g^- dP,$$

where g^+ and g^- are the following nonnegative functions:

$$g^+(\omega) = \max\{0, g(\omega)\} \qquad\qquad g^-(\omega) = \max\{0, -g(\omega)\}.$$

The derivation of basic properties of the Lebesgue integral (linearity, etc.) can be found in [8], and we will freely make use of such properties. We also will use of the following notation:

$$\int_A g dP := \int I_A g dP.$$

With this machinery in hand, we can now define the fundamental concepts of probability. For the rest of this section, we assume that we are working in a probability space $(\Omega, \mathcal{F}, P)$.

Definition 1.3.6. Two events A and B are *independent*, notated $A \perp\!\!\!\perp B$, if $P(A \cap B) = P(A)P(B)$. An event A is independent from a σ-algebra $\mathcal{T}$ ($A \perp\!\!\!\perp \mathcal{T}$) if for every event $T \in \mathcal{T}$, $P(A \cap T) = P(A)P(T)$. Two σ-algebras $\mathcal{T}$ and $\mathcal{U}$ are independent ($\mathcal{T} \perp\!\!\!\perp \mathcal{U}$) if for every event $T \in \mathcal{T}$ and $U \in \mathcal{U}$, $P(T \cap U) = P(T)P(U)$. Two random variables ϕ and ψ are independent ($\phi \perp\!\!\!\perp \psi$) if $\sigma(\phi)$ and $\sigma(\psi)$ are independent.

Independence is a mathematical definition that attempts to capture the intuitive notion

of two random processes that don't affect each other, such as the outcome of two different coin tosses. The notion of independence can be extended to larger collections of random processes. For a collection of random variables $\{\phi_i\}_{i=1}^{\infty}$, the collection is (pairwise) independent if and only if $\forall i \neq j, \sigma(\phi_i) \perp\!\!\!\perp \sigma(\phi_j)$. The stronger notion of *mutual* independence requires that for any finite indexing set $J \subseteq \mathbb{N}$ and for any choice of sets $\{S_j\}_{j\in J}$ such that $S_j \in \sigma(\phi_j)$, $P(\cap_{j\in J} S_j) = \Pi_{j\in J} P(S_j)$. If a sequence of random variables is just said to be "independent," it is understood that this refers to the stronger notion of mutual independence. A sequence of random variables $\{\phi_i\}_{i=1}^{\infty}$ is *identically distributed* if each ϕ_i maps to the same outcome space U and for each set $S \in \mathcal{U}$, $P(\phi_i \in S)$ takes on the same value independently of the choice of i. Sequences of random variables that are both independent and identically distributed, abbreviated i.i.d., are of great importance in probability theory.

To illustrate the i.i.d. concept, consider a sequence of Bernoulli random variables $\{B_i\}_{i=1}^{\infty}$, $B_i : \Omega \to \{0, 1\}$. Then the sequence is (mutually) independent if for any $J \subseteq \mathbb{N}$ and for any assignment $\{x_j\}_{j\in J} \to \{0, 1\}$, $P\big(\cap_{j\in J} \{B_j = x_j\}\big) = \Pi_{j\in J} P(B = x_j)$. The sequence is identically distributed if there is a fixed number p for which

$$p = P(B_1 = 1) = P(B_2 = 1) = \cdots = P(B_n = 1) = \cdots .$$

Assuming that $\{B_i\}_{i=1}^{\infty}$ is i.i.d., we can now define a very useful random variable as follows:

$$Bin_{n,p} := \sum_{i=1}^{n} B_i, \tag{1.11}$$

where p is the probability $P(B_i = 1)$. Note that $Bin_{n,p}$ is a measurable function from Ω to $\mathbb{R}$, and hence a legitimate random variable, because sums of measurable functions are also measurable. (The proof that sums, products, compositions etc. of measurable functions are also measurable can be found in [8].) $Bin_{n,p}$ is called the *Binomial* random variable with n trials and probability of success p, and many familiar real-world processes can be modeled with this random variable. For instance, if you flip a fair coin 100 times, the probabilities of getting various numbers of "heads" can be calculated from the distribution of the Binomial random variable $Bin_{100,\frac{1}{2}}$. The general distribution of the Binomial random variable is given by the following formula, which can be derived combinatorically:

$$P(Bin_{n,p} = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

where $\binom{n}{k}$ is the choose function, $\frac{n!}{k!(n-k)!}$.

Continuing our discussion of probability, we also require the following basic notions.

Definition 1.3.7. (Conditional Probability, Expectation, Variance) For two events A and B for which $P(B) > 0$, the *probability of A conditional on B*, which can be called the *conditional probability*, is

$$P(A \mid B) := \frac{P(A \cap B)}{P(B)}.$$

For a random variable $\phi : \Omega \to \mathbb{R}$, the *expectation* of ϕ, also known as the *mean*, is

$$E(\phi) := \int \phi dP.$$

The *expectation of ϕ conditioned on the event B*, which can be called the *conditional expectation*, is

$$E(\phi \mid B) := \frac{\int_B \phi dP}{P(B)}.$$

Suppose $E(\phi) = \mu$ and μ is not ∞ or $-\infty$. Then $(\phi - \mu)^2$ is also a random variable from Ω to $\mathbb{R}$, and we define the *variance of ϕ* to be

$$Var(\phi) := \int (\phi - \mu)^2 dP.$$

If $Var(\phi)$ is not ∞, then the *standard deviation of ϕ* is defined to be $\sqrt{Var(\phi)}$.

In addition to these straightforward concepts, we also require some more-involved measure-theoretic definitions.

Definition 1.3.8. Given a probability space $(\Omega, \mathcal{F}, P)$, a sub-σ-algebra $\mathcal{T} \subseteq \mathcal{F}$, and a random variable $\phi : \Omega \to \mathbb{R}$, the *conditional expectation of ϕ with respect to $\mathcal{T}$*, denoted $E(\phi \mid \mathcal{T})$, is defined to be a function $E(\phi \mid \mathcal{T}) : \Omega \to \mathbb{R}$ for which

1. $E(\phi \mid \mathcal{T})$ is measurable with respect to $\mathcal{T}$.
2. For all $T \in \mathcal{T}$, $\int_T E(\phi \mid \mathcal{T}) dP = \int_T \phi dP$.

For any event $A \in \mathcal{F}$, the *conditional probability of A with respect to $\mathcal{T}$*, notated $P(A \mid \mathcal{T})$, is defined to be $E(I_A \mid \mathcal{T})$. If $\mathcal{T} = \sigma(\psi)$ for some random variable ψ, by convention the conditional expectation and conditional probability can be written $E(\phi \mid \psi)$ and $P(A \mid \psi)$, respectively.

The definition of conditional expectation is a little strange because it is non-constructive: it says that *if* a function exists satisfying certain properties, then this function can be taken to be the conditional expectation. From the definition alone, we cannot determine whether such a function exists, or if there might be more than one such function. Luckily, it can be shown (with some involved mathematical arguments; see [9] or [10]) that for a wide class of random variables and sub-σ-algebras, the conditional expectation exists. For instance, a sufficient condition for $E(\phi \mid \mathcal{T})$ to exist is that $E(|\phi|) < \infty$, and this implies that conditional probabilities $P(A \mid \mathcal{T})$ *always* exist. As for the question of uniqueness, it is not too hard to demonstrate that if two functions f_1 and f_2 satisfy the definitional requirements for $E(\phi \mid \mathcal{T}) : \Omega \to \mathbb{R}$, then

$$P\big(\{x \in \Omega \mid f_1(x) \neq f_2(x)\}\big) = 0. \tag{1.12}$$

Hence the set of points where the functions disagree is "small" in a sense, and generally unproblematic, so f_1 and f_2 are essentially interchangeable. Thus the conditional expectation can be thought to be unique for all intents and purposes. If two functions satisfy (1.12), they are said to be equivalent *almost surely*, or *a.s.*, and sets of functions that are equivalent *a.s.* form equivalence classes. More generally, any property holds *a.s.* if the set of points for which it does not hold has measure zero.

Expectation and conditional expectation can be related by the following well-known property, which is a direct consequence of the definition of conditional expectation:

Fact 1.3.1. *(Law of Iterated Expectation) For a random variable ϕ and a sub-σ-algebra $\mathcal{T}$ of $\mathcal{F}$ ($\mathcal{T} \subseteq \mathcal{F}$), if $E(|\phi|) < \infty$, then the following equation holds:*

$$\begin{aligned} E(\phi) &= E\big(E(\phi \mid \mathcal{T})\big) \\ \left(\int \phi dP &= \int E(\phi \mid \mathcal{T}) dP\right) \end{aligned}$$

We will make frequent use of Fact 1.3.1 in Chapters 2 and 3. Another useful property is given by the following lemma. The proof is a good illustration of how to work with conditional expectations.

Lemma 1.3.1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra of $\mathcal{F}$, and let*

$\{B_i\}_{i\in I}$ *be a countable indexed set of pairwise-disjoint events in* $\mathcal{F}$*. Then*

$$P(\cup_{i\in I} B_i \mid \mathcal{G}) = \sum_{i\in I} P(B_i \mid \mathcal{G}), \textit{ almost surely.}$$

Proof. We show that the RHS of the above equation satisfies the two definitional requirements for being the conditional probability $P(\cup_{i\in I} B_i \mid \mathcal{G})$. The first requirement is to be $\mathcal{G}$-measurable; as $\sum_{i\in I} P(B_i \mid \mathcal{G})$ is a sum of $\mathcal{G}$-measurable functions, this is immediate. The second requirement dictates that for any set $G \in \mathcal{G}$, the following must hold:

$$\int_G \sum_{i\in I} P(B_i \mid \mathcal{G}) dP = \int_G I_{\cup_{i\in I} B_i} = P\big(G \cap (\cup_{i\in I} B_i)\big).$$

To demonstrate this, we see that

$$\int_G \sum_{i\in I} P(B_i \mid \mathcal{G}) dP = \sum_{i\in I} \int_G P(B_i \mid \mathcal{G}) dP = \sum_{i\in I} P(G \cap B_i),$$

where the second equality holds by the definition of the conditional probability expressions $P(B_i \mid \mathcal{G})$. Continuing, we thus have

$$\sum_{i\in I} P(G \cap B_i) = P\big(\cup_{i\in I} (G \cap B_i)\big) = P\big(G \cap (\cup_{i\in I} B_i)\big),$$

which holds due to the $\{B_i\}_{i\in I}$ being pairwise disjoint. □

One last notion is needed: that of conditional independence.

Definition 1.3.9. For three random variables ϕ_1, ϕ_2, and ϕ_3, we say that ϕ_1 and ϕ_2 are *conditionally independent given* ϕ_3, denoted $(\phi_1 \perp\!\!\!\perp \phi_2) \mid \phi_3$, if for every $A \in \sigma(\phi_1)$ and $B \in \sigma(\phi_2)$, the following equation holds:

$$P(A \cap B \mid \phi_3) = P(A \mid \phi_3) P(B \mid \phi_3). \tag{1.13}$$

Conditional independence encapsulates the notion of two random variables ϕ_1 and ϕ_2 being related, but only by their shared relationship with a third random variable ϕ_3. This covers all of the basic probabilistic concepts that will be used. We now introduce two famous theorems.

Theorem 1.3.1. *(Weak Law of Large Numbers) Let* $\{\phi\}_{i=1}^{\infty}$ *be an i.i.d. sequence of random variables with finite expectation* $E(\phi_i) = \mu$*, and define the* sample mean $\overline{\phi_n}$ *to be* $\sum_{i=1}^{n} \phi_i / n$*. Then for any*

$\epsilon > 0$,

$$\lim_{n\to\infty} P(|\overline{\phi_n} - \mu| < \epsilon) = 1. \tag{1.14}$$

The Weak Law of Large Numbers captures the intuitive notion that sample means converge to underlying expectations "in the long run." For instance, if somebody flips a fair coin repeatedly, after a long time it is likely that the proportion of heads will be close to $\frac{1}{2}$. This can be expressed mathematically by considering an i.i.d. sequence of Bernoulli random variables $\{B_i\}_{i=1}^{\infty}$, $B_i : \Omega \to \{0, 1\}$ for which $P(B_i = 1) = \frac{1}{2}$ and therefore $E(B_i) = \frac{1}{2}$. Then $\overline{B_n}$ will equal $Bin_{n,\frac{1}{2}}/n$, and the Weak Law of Large Numbers implies that

$$\begin{aligned} \lim_{n\to\infty} P(\text{The proportion of "heads" after } n \text{ tosses is within } \epsilon \text{ of } \frac{1}{2}) &= \lim_{n\to\infty} P\left(\left|\frac{Bin_{n,\frac{1}{2}}}{n} - \frac{1}{2}\right| < \epsilon\right) \\ &= 1. \end{aligned}$$

Thus the proportion of heads will "eventually" be close to $\frac{1}{2}$ with almost certain probability, but the result does not indicate how large the number of trials n must be before we begin to see this effect. A second theorem – the Central Limit Theorem – can be used to address this question. The Central Limit Theorem requires the notion of a *normal random variable*.

Definition 1.3.10. A *normal random variable* $N(\mu, \sigma^2) : \Omega \to \mathbb{R}$ is a random variable with mean μ and variance σ^2 obeying the following condition: for any $(a, b) \subseteq \mathbb{R}$,

$$P\big(N \in (a, b)\big) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx.$$

$N(0, 1)$ is known as the *standard normal random variable*.

Theorem 1.3.2. *(Central Limit Theorem) Let $\{\phi\}_{i=1}^{\infty}$ be an i.i.d. sequence of random variables with finite expectation $E(\phi_i) = \mu$ and variance $Var(\phi_i) = \sigma^2 > 0$, and let V_n be the random variable defined in the following manner:*

$$V_n = \frac{\overline{\phi_n}}{\sigma/\sqrt{n}}.$$

Then V_n converges in distribution to the standard normal random variable $N = N(0, 1)$. That is, for any $(a, b) \subseteq \mathbb{R}$,

$$\lim_{n\to\infty} P\big(V_n \in (a, b)\big) = P\big(N \in (a, b)\big).$$

The Central Limit Theorem tells us that V_n can be approximated with $N(0, 1)$ for large n,

a fact that we express by writing $V_n \sim N(0,1)$. In many practical situations, the approximation can be suitably accurate when n is as small as 15 or 20. For instance, if ϕ_i is a sequence of Bernoulli random variables with the probability of the two outcomes given by p and $1-p$, one common rule of thumb is that the Central Limit Theorem approximation can be used for n exceeding $\frac{5}{p}$.

The Central Limit Theorem can be used to experimentally test hypotheses. If you want to determine whether a coin has a 50 percent probability of coming up "heads," you can flip it, say, 100 times. Then the Central Limit Theorem implies that the probability that the number of "heads" will be between 40 and 60 is in excess of 95%. Thus if you were to see something like 65 heads, the hypothesis that $P(\text{heads}) = \frac{1}{2}$ could be ruled out with high confidence.

This type of experimental protocol is known as a *statistical hypothesis test*. We want to formalize and make precise this type of reasoning. The following definition of hypothesis testing is not completely general, but will suffice for our purposes.

Definition 1.3.11. (Hypothesis Testing) Suppose that a process is modeled as a collection of random variables $\{\phi_i\}_{i\in I}$. If there are two competing mutually exclusive hypotheses about characteristics of the collection $\{\phi_i\}_{i\in I}$, these can be called the *null hypothesis* H_0 and the *alternative hypothesis* H_A. A *test statistic* is a function T of the random variables $\{\phi_i\}_{i\in I}$ taking an output in $\mathbb{R}$. For any possible value x of T, the corresponding *p-value* is the maximum possible probability, given that the distributions of the $\{\phi_i\}_{i\in I}$ are compatible with the null hypothesis, that T exceeds x.

More general notions for the extremity of T can be used to formulate a different definition for the p-value, but Definition 1.3.11 can frequently be used by defining T judiciously. If the experimenter decides to reject the null hypothesis in favor of the alternative hypothesis if T exceeds some critical value z, then the *significance level* of the test is defined to be the maximum possible probability of (erroneously) rejecting H_0 if H_0 is true. The *power* of the test is probability of (correctly) rejecting H_0 if H_A is true.

In hypothesis tests, the random variables $\{\phi_i\}_{i\in I}$ are frequently presumed to be i.i.d. If H_0 and H_A are concerned with competing possibilities for $E(\phi_i)$, results like Theorem 1.3.2, which require the assumption of i.i.d., can then be used to directly compute the significance level and power of the test. A large proportion of statistical theory falls under this rubric. However, there are compelling reasons not to make the i.i.d. assumption in the experimental scenarios explored in this dissertation. We thus spend some time in Chapters 2 and 3 exploring how to compute significance

and power when the i.i.d. assumption cannot be made.

1.4 Order Theory and Domain Theory

In Chapter 4, we will study an application of the Kochen-Specker theorem to a partial order known as the *spectral order*. The spectral order is a type of partial order known as an *exact domain*, and we will need to be familiar with some of the basics of Domain Theory.

We first start with the notion of a partial order, which we can build up by starting with the definition of a relation.

Definition 1.4.1. A *relation* R on a set S is a collection of ordered pairs with elements in S. For $a, b \in S$, we say that "a bears relation R to b" iff $(a, b) \in R$.

The statement "a bears relation R to b" is notated aRs, where R is generally replaced with a different symbol, such as $\sqsubseteq$. The symbol can be used to refer to the relation itself, as we do in the following definition:

Definition 1.4.2. A relation $\sqsubseteq$ on a set S is:

$$\begin{aligned} \textit{Reflexive} \quad & \text{if} \quad \forall a \in S, a \sqsubseteq a \\ \textit{Transitive} \quad & \text{if} \quad \forall a, b, c \in S, \text{ if } a \sqsubseteq b \text{ and } b \sqsubseteq c, \text{ then } a \sqsubseteq c \\ \textit{Antisymmetric} \quad & \text{if} \quad \forall a, b \in S, \text{ if } a \sqsubseteq b \text{ and } b \sqsubseteq a, \text{ then } a = b \end{aligned}$$

Definition 1.4.3. A *preorder* is a relation that is reflexive and transitive. A *partial order* is a preorder that is antisymmetric.

When working with a relation $\sqsubseteq$ that is either a preorder or a partial order, we will sometimes use the following notation:

$$\uparrow a = \{b \mid b \sqsupseteq a\} \qquad\qquad \downarrow a = \{b \mid b \sqsubseteq a\}.$$

We can also extend this notation to sets in a natural way: if A is a set, $\uparrow A = \{b \mid \exists a \in A, b \sqsupseteq a\}$, and $\downarrow A = \{b \mid \exists a \in A, b \sqsubseteq a\}$.

A *domain* is a partial order satisfying some additional conditions. These additional conditions require some new definitions, given here:

Definition 1.4.4. If $\sqsubseteq$ is a partial order on a set S, we say that a subset $D \subseteq S$ is *directed* if D is nonempty and $\forall a, b \in D$, $\exists c \in D$ for which $a \sqsubseteq c$ and $b \sqsubseteq c$. If every directed set $D \subseteq S$ has a least upper bound in S, we say that $(S, \sqsubseteq)$ is a *directed-complete partial order*, abbreviated *dcpo*.

Definition 1.4.5. Let $(S, \sqsubseteq)$ be a dcpo. We define a new relation $\ll_e$ on S as follows. We say that $a \ll_e b$ if for any directed $D \subseteq S$ for which $\sup D = b$, there exists a $d \in D$ such that $a \sqsubseteq d$. Furthermore, we will use the following double-arrow notations:

$$\Uparrow a = \{b \mid b \gg_e a\} \qquad\qquad \Downarrow a = \{b \mid b \ll_e a\}.$$

Definition 1.4.6. A partial order $(S, \sqsubseteq)$ is an *exact domain* if it is a dcpo, and for all $y \in S$, $\Downarrow y$ is a directed set for which $\sup \Downarrow y = y$.

Remark 1.4.1. In Domain Theory, one often works with the relation "$\ll$" instead of $\ll_e$. The definition of $\ll$ is obtained from Definition 1.4.5 by replacing "$\sup D = b$" with "$\sup D \geq b$", and a *domain* can be defined by replacing "$\Downarrow y$" in Definition 1.4.6 with "$\{b \mid b \gg a\}$". Domains are more commonly studied than exact domains, but the two domain-theoretic partial orders which we will study in Chapter 4 – the Bayesian and spectral order – are exact domains, and so dcpos satisfying Definition 1.4.5 are the primary focus in this work.

The last notion that we define is *Scott continuity*. For a maps between domains, Scott continuity is a highly desirable property. For instance, if P is a dcpo with a least element $\perp$ and $f : P \to P$ is a map from P to itself, then if f is Scott continuous, it can be shown that f has a least fixed point given by $\sup_n f^n(\perp)$, by the Knaster-Tarski-Scott theorem. The definition of Scott continuity is as follows.

Definition 1.4.7. Let S and T be exact domains. A map $\phi : D \to E$ is *Scott continuous* if

1. ϕ is *monotone*: $\forall x, y \in S$, $x \sqsubseteq y \Rightarrow \phi(x) \sqsubseteq \phi(y.)$
2. ϕ preserves directed suprema: For a directed set $D \subseteq S$, $\phi(\sup D) = \sup \phi(D)$.

Chapter 2

The Clauser-Horne-Shimony-Holt Inequality

2.1 The Setting And The Mathematical Model

In this chapter, we examine the Clauser-Horne-Shimony-Holt (CHSH) Bell experiment in a mathematically rigorous manner. We will set up a precise mathematical model for the experiment with a set of axioms/assumptions about the behavior of variables in the experiment. From these assumptions, we will be able to derive the CHSH inequality in a fully general, measure-theoretic setting, as well as a method for constructing a physical test of the inequality that does not depend on any tacit assumptions (such as independent, identically distributed trials). If a violation of the inequality is observed in an experiment, one could conclude that one of the axioms/assumptions must not hold in the physical world.

The first task is to construct the model. To visualize what must be modeled, let us describe the set up of an ideal CHSH-Bell experiment, which is depicted in Figure 2.1. A photon source, such as a low-powered laser, is pointed at a beamsplitter. The source emits a single photon, and when it strikes the beamsplitter, it will split into two photons travelling on separate trajectories towards detectors at opposite ends of the laboratory. If the experiment is set up properly, these two photons will be in an entangled state. Upon arrival, each of these photons is subjected to a measurement by the detector. The time of detection of the photons is calibrated so that the detection events are as close to simultaneous as possible. The goal is to ensure that the detection events are *spacelike separated* – so if some sort of (non-superluminal) signal were sent from detector 1 at the time of detection to detector 2, the signal would not get to detector 2 until after detector 2 had already made its corresponding measurement.

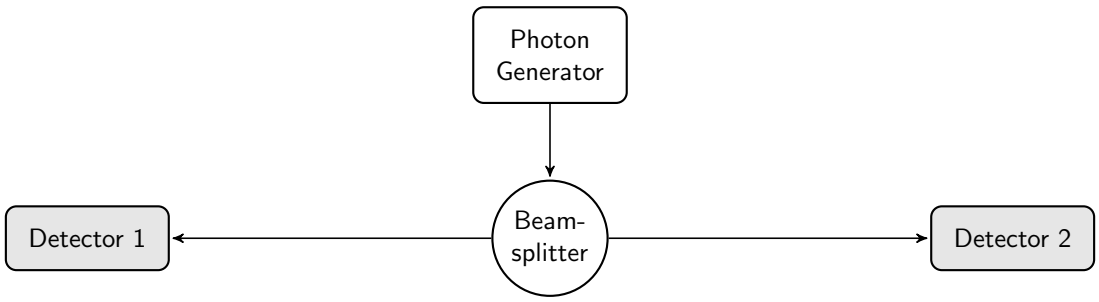

Figure 2.1: Diagram of a Bell test experiment

As depicted in Figure 2.2, Detector 1 has two measurement settings and two possible outputs. We have two setting choices (a or a') at detector 1, which can be freely set by the experimenter. Whichever setting is chosen, the detector has two choices of output, $+1$ or -1. Detector 2 has a

very similar scheme; the only difference is that we label its settings as b and b', to distinguish them from the settings of Detector 1, which need not be the same. Detector 2 also yields an output of $+1$ or -1. Sometimes these outcomes are referred to as "spin up" and "spin down."

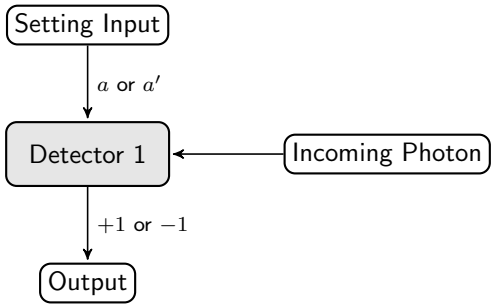

Figure 2.2: Detail at detector 1

This is a very abstract description of what is going on at the detectors. To get some sort of an intuitive picture of the process, the detector can be visualized as a transparent object that sorts photons based on polarizations. When a photon travels through the object, it is forced to travel through one of two allowable paths, the "up-down" or "left-right" path. This choice would be recorded by the detector as $+1$ or -1, respectively. Figure 2.3 is an illustration of this idea.

Thinking of the detector setting as an angle at which we orient a sorter is intuitively helpful. However, it is not necessary to know anything about the inner workings of the detectors to understand the implications of Bell's theorem. Indeed, we only need to know that there are two settings and two possible outputs at each of the two detectors – nothing more than the scheme of Figure 2.2. For the purposes of the CHSH inequality, a detector is nothing more than a mysterious box with a 2-setting dial on the outside that has the option of, say, flashing a green LED ($+1$) or a red LED (-1) upon a detection event. Even the notion that a photon causes this detection event is not necessary for mathematical model.

Our mathematical model will be expressed in the language of probability. This is because, as we will see, any *single* execution of the Bell test experiment cannot tell us anything about the incompatibility of quantum mechanics with locality – this is for essentially the same reason that the outcome of any *single* coin toss will not tell you anything about whether or not the coin is biased. The assertions of Bell's theorem are assertions about probabilities, and thus they can only be demonstrated statistically, with the results being tallied over the course of multiple repetitions of the experiment.

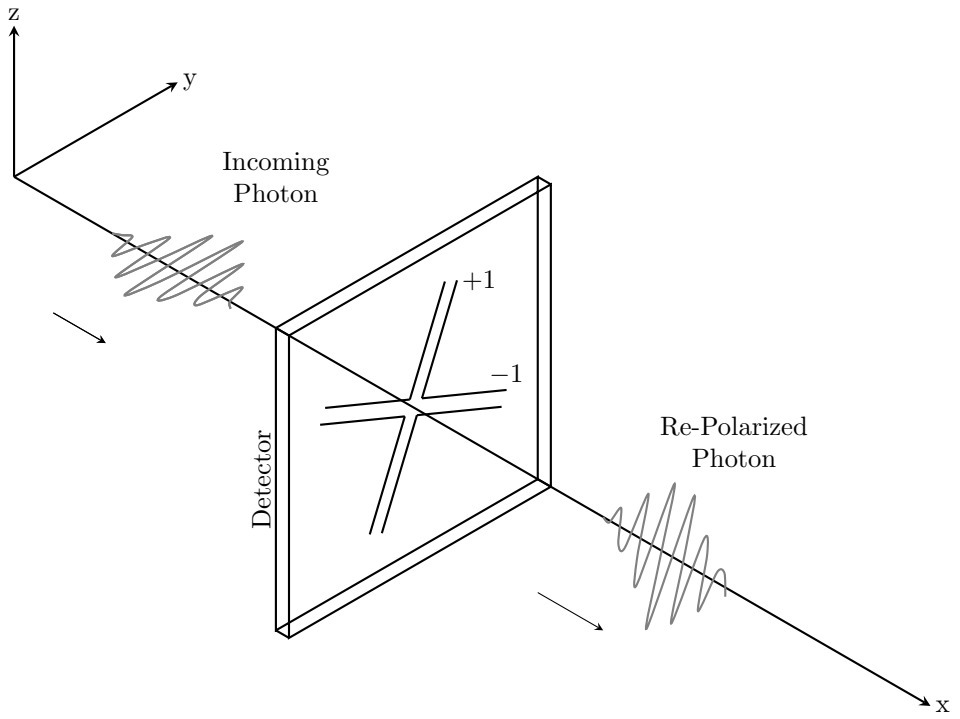

Figure 2.3: A photon with polarization in the x-y plane is measured. The detector reads "+1" after the photon is re-polarized in a new direction.

The following definition contains the elements of our model of the experiment. This is expressed in the language of measure-theoretic probability introduced in the previous chapter.

Definition 2.1.1. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let λ, A, B, D_1, D_2 be the random variables

$$\lambda : \Omega \to \Lambda, \quad \Lambda \text{ is a measurable space,}$$
$$A : \Omega \to \{a, a'\}, \quad B : \Omega \to \{b, b'\},$$
$$D_1 : \Omega \to \{-1, 1\}, \quad D_2 : \Omega \to \{-1, 1\}.$$

We call λ the joint state of the system. A and B are detector 1's and detector 2's settings, respectively, and D_1 and D_2 are detector 1's and detector 2's output, respectively. We label events in $\mathcal{F}$ corresponding to the outputs of D_1 and D_2 with the following notation:

$$+_1 = \{D_1 = +1\}, \quad -_1 = \{D_1 = -1\}, \quad +_2 = \{D_2 = +1\}, \quad -_2 = \{D_2 = -1\}.$$

We now talk a little about the empirical meaning of the objects described in Definition 2.1.1. The most cryptic of the five random variables above is λ, the "joint state of the system." This mysteriousness is fitting, because λ describes the portion of the experiment that we don't directly observe: the photon pair that is theorized to be travelling towards the detectors. We do not observe the photon as it travels; indeed, the only way we infer that anything is happening at all is that our detectors "flash an L.E.D.," so to speak. Quantum mechanics has a well-defined theory for λ and how it triggers the detectors, as we will discuss in Section 2.6. But our first goal is to derive general constraints that any "local" physical theory must obey (the notion of "locality" will be defined soon), and so λ is defined generally.

One possible conception of λ could be, for instance, a pair of random vectors, each one containing attributes such as wavelength, energy, and other quantities for two photons of light travelling in opposite directions. However, it should be emphasized that as we do not observe the photons as they travel from the beamsplitter to the detectors, we can imagine a plausible alternative local theory such that the supposed photon source is not generating photons at all, but is generating two unknown subatomic particles that trigger the detectors in some unknown way, and λ would represent the entire trajectory and momentum of these mysterious particles. To think as generally as possible, it would be best refer back to Figure 2.1, erase everything between the two detectors (i.e, erase "Photon Generator" and "Beam Splitter"), and replace it with a big mysterious cloud with "λ" printed on it. Now, λ represents anything and everything that happens between the two detectors, prior to the detection event, that can have an effect on both detection events.

The other four random variables are more straightforward, because they model aspects of the experiment that we observe on a macroscopic level. The only point of interest is the modeling of the detector settings as random variables. When photon 1 arrives at detector 1, the detector will be in either setting a or setting a', and similarly for detector 2 with b and b', and the experimenter controls these settings. In the Bell test experiment, it turns out to be important that the experimenter randomly selects the detector settings for each trial, randomly toggling between a and a' on one end, and between b and b' on the other end. This is why the settings are modeled as random variables.

This leads us towards the axioms/assumptions about the experiment. For instance, the random processes governing the measurement settings should occur right before the detection event, and should be independent of each other. This statement about the experiment can be made into a precise mathematical assumption. The next three assumptions encapsulate the set of requirements

that an experimenter must satisfy in order to properly test Bell's theorem.

Experimental Assumption 1:

$$A \perp\!\!\!\perp B. \tag{2.1}$$

In practice, the independent choice of measurement settings could be achieved, for example, by attaching a random number generator to each detector. The next assumption rules out uninteresting trivial cases.

Experimental Assumption 2:

$$P(A = a)P(A = a')P(B = b)P(B = b') > 0. \tag{2.2}$$

The final experimental assumption is as follows:

Experimental Assumption 3:

$$A \perp\!\!\!\perp \lambda, \quad B \perp\!\!\!\perp \lambda. \tag{2.3}$$

(2.3) captures the notion that the detector setting can be chosen, by the experimenter, in a random way that is not affected by the state of the approaching particles. This is sometimes called the "free will" assumption, or the "λ-independence" assumption. The experimenter should be able to satisfy (2.3) by ensuring that the random process that sets the detector settings has nothing to do with the approaching photons, or anything else occurring between the two detectors, prior to detection. Unless something very strange is going on, one would expect the aforementioned random number generators to not be correlated with the state of some photon emitted by a laser some distance away. But if an even more trustworthy source of uncorrelated randomness is desired, the detector setting could be determined by the random fluctuations in the intensity of cosmic background radiation, coming from the opposite direction of the approaching photon.

Other axiomatizations of the Bell model supplement (2.3) with a stronger assumption, that the *joint* distribution of A and B is independent from λ, or $(A, B) \perp\!\!\!\perp \lambda$ (See, for example, [11] or [12]). This goes further toward ruling out any dependencies between λ and A and B that shouldn't exist if the experimenter uses a sufficient random mechanism to choose the measurement setting. In these other axiomatizations of the Bell setting, (2.3) may not be sufficient to prove the CHSH inequality. However, as demonstrated in Section 2.2, our axiomatization does not require us to assume $(A, B) \perp\!\!\!\perp \lambda$.

Finally, we model the output of the detectors as random variables. This merits an explanation – after all, one might think that D_1 could be deterministic, given λ and A: if one has complete knowledge of the state of the photon λ and the setting of the photon detector A, shouldn't a physical theory be able to predict D_1 exactly, at least in principle? Then D_1 and D_2 could be defined as functions of the other random variables. This line of thought makes use of the assumptions of locality and *determinism*. However, we do not need to assume determinism to show that Quantum Mechanics is incompatible with locality. Bell's original paper [1] assumed that D_1 and D_2 were deterministic, given λ, and A or B, respectively. Later work, such as the 1974 paper of Clauser and Horne [13], generalized the result to the case where the D_1 and the D_2 could still be random even with knowledge of λ and A and B. In keeping with this generality, we allow D_1 and D_2 to be random, even given λ, A and B. Of course, a random variable can be constant or a function of other random variables; we are not *ruling out* determinism – we are just being careful not to unnecessarily appeal to it in the proof that Quantum Mechanics is incompatible with locality.

We now have a complete mathematical description of the experiment. This gives us a framework to discuss a local theory and the conditions it must satisfy, and we can later contrast that with the quantum explanation, which can also be presented in the framework presented thus far.

The derivation of the locality constraint for the Bell test experiment depends on an additional assumption, called the locality assumption. While the previous three assumptions are conditions on experimental parameters that the experimentalist tries to meet, the locality assumption is a little different: it is an assumption about the nature of the physical theory governing the experiment.

The locality assumption concerns the relationship between the state, the outcomes, and the settings. To formulate it in a concise manner, we define the random vectors

$$V_1 = (D_1, A), \qquad V_2 = (D_2, B).$$

Then the assumption is as follows:

Locality Assumption:

$$(V_1 \perp\!\!\!\perp V_2) \mid \lambda. \tag{2.4}$$

Theories satisfying (2.4) are sometimes called *local hidden variable theories*, or just *hidden variable theories*, and λ is then referred to as the local "hidden variable." The following question may be

asked: how is Equation (2.4) an expression of locality? Remember that λ represents what is going on between the two detectors, prior to the measurement event, that can affect the detection events. Once we condition on this knowledge, what occurs at detector 1 should be independent of what occurs at detector 2. Equation (2.4) is basically saying that the events at detector 1 cannot be correlated with the events at detector 2 *beyond* the effects of the shared history of what happened between them and prior to detection, represented by λ.

Since V_1 and V_2 each only have four possible outputs, (2.4) is essentially a statement of the conditional independence of a collection of sixteen pairs of events. For instance, one of the sixteen consequences of (2.4) would be

$$P\left[(+_1 \cap \{A = a'\}) \bigcap (-_2 \cap \{B = b\}) \;\middle|\; \lambda\right] = P\big(+_1 \cap \{A = a'\} \mid \lambda\big) \cdot P\big(-_2 \cap \{B = b\} \mid \lambda\big).$$

It is important to note that (2.4) does not rule out the possibility that D_1 and D_2 are correlated: after all, the photon generator could generate photon pairs that, due to some aspect of their frequency or intensity, always result in $+1$ at both detectors; then D_1 and D_2 would be *perfectly* correlated. However, this correlation is due to the dependence of D_1 and D_2 on the same λ: their correlation comes from the state of the system. The dependence is *not* from an instantaneous, non-local communication between detectors 1 and 2, and once we account for λ, the correlation will disappear; i.e., the conditional independence in (2.4) is still satisfied.

2.2 Analysis of λ-Independence

The four assumptions of the previous section suffice to prove the basic case of Bell's theorem. Only slight modifications are required to prove more general cases and to perform statistical analysis. As these assumptions are so central to the work of Chapters 2 and 3, a deeper analysis of the assumptions is merited, and the following three sections are devoted to this analysis. In this section, we will examine the stripped-down appearance of the λ-independence assumption (equation (2.3)), and contrast it with the stronger version that is sometimes used. In Section 2.3 we will explore what happens to the axioms if λ is taken to be a finite random variable, and Section 2.4 will explore the relationship between the assumptions set forth here and another measure-theoretic axiomatization of the Bell experiment in a 2012 paper of Brandenburger and Keisler [12]. The reader who is already satisfied with the assumptions as presented may skim or skip these three sections, as the mathematical development of the subsequent sections does not depend on anything derived here.

We start with a discussion of the alternate version of (2.3):

Experimental Assumption 3*:

$$(A, B) \perp\!\!\!\perp \lambda. \tag{2.5}$$

The first thing to notice about (2.5) is that it is a stronger assumption than (2.3): (2.5) directly implies (2.3). To demonstrate, we will use a mathematical tool introduced in Chapter 1, Lemma 1.3.1.

To see the implication, note that (2.5) states that the σ-algebra generated by the random vector (A, B) is independent from the σ-algebra generated by λ, and that this is equivalent to the statement

$$\forall \mathbf{a} \in \{a, a'\} \text{ and } \mathbf{b} \in \{b, b'\}, \quad P\big((A, B) = (\mathbf{a}, \mathbf{b}) \mid \lambda\big) = P\big((A, B) = (\mathbf{a}, \mathbf{b})\big). \tag{2.6}$$

(The equivalence of (2.6) and (2.5) is an immediate consequence of the measure-theoretic definition of conditional probabilities.) Now, we can apply Lemma 1.3.1 to see that

$$P(A = a \mid \lambda) = P(A = a \cap (B = b \cup B = b') \mid \lambda) = P(A = a \cap B = b \mid \lambda) + P(A = a \cap B = b' \mid \lambda),$$

and by (2.6), this equals

$$P(A = a \cap B = b) + P(A = a \cap B = b') = P(A = a).$$

Similarly, we can prove that $P(A = a' \mid \lambda) = P(A = a')$, $P(B = b \mid \lambda) = P(B = b)$, and $P(B = b' \mid \lambda) = P(B = b')$. Taken together, these equivalences imply that $A \perp\!\!\!\perp \lambda$ and $B \perp\!\!\!\perp \lambda$, which is of course just the statement of (2.3).

As (2.3) will suffice for our purposes, there is no reason to make the stronger assumption of (2.5). However, other authors frequently *need* to assume (2.5) to prove Bell inequalities. Pathological counterexamples that violate Bell's inequality while satisfying (2.3) but not (2.5) may exist, depending on the formulation of the other axioms. This possibility – which is not a feature of our system – can be traced to a different formulation of the locality assumption. In fact, in our formulation, the stronger formula (2.5) can be *derived* from (2.3) in conjunction with the other assumptions, which rules out the aforementioned pathological counterexamples. Once again, Lemma 1.3.1 will be used.

Proposition 2.2.1. *The condition* (2.3)*, in conjunction with conditions* (2.1) *and* (2.4)*, imply* (2.5)*.*

Proof. We show that $P\big((A,B)=(a,b)\mid\lambda\big)=P\big((A,B)=(a,b)\big)$; the proof for the three other cases (a,b'), (a',b), and (a',b') is the same. We have

$$\begin{aligned} P\big((A,B)=(a,b)\mid\lambda\big) &= P\big(\{A=a\}\cap\{B=b\}\cap(\cup_{i,j\in\{+1,-1\}}\{D_1=i\}\cap\{D_2=j\})\mid\lambda\big) \\ &= \sum_{i=\pm1}\sum_{j=\pm1}P\big(\{A=a\}\cap\{B=b\}\cap(\{D_1=i\}\cap\{D_2=j\})\mid\lambda\big), \end{aligned}$$

by Lemma 1.3.1. Applying (2.4) to the above expression yields

$$\sum_{i=\pm1}\sum_{j=\pm1}P\big(\{D_1=i\}\cap\{A=a\}\mid\lambda\big)\cdot P\big(\{D_2=j\}\cap\{B=b\}\mid\lambda\big).$$

Factoring the expression and applying Lemma 1.3.1, we obtain

$$\begin{aligned} &P\big(\{D_2=+1\}\cap\{B=b\}\mid\lambda\big)\big[P\big(\{D_1=+1\}\cap\{A=a\}+P\big(\{D_1=-1\}\cap\{A=a\}\big] \\ &\quad+P\big(\{D_2=-1\}\cap\{B=b\}\mid\lambda\big)\big[P\big(\{D_1=+1\}\cap\{A=a\}+P\big(\{D_1=-1\}\cap\{A=a\}\big] \\ &=P\big(\{D_2=+1\}\cap\{B=b\}\mid\lambda\big)\cdot P(A=a\mid\lambda)+P\big(\{D_2=-1\}\cap\{B=b\}\mid\lambda\big)\cdot P(A=a\mid\lambda) \\ &\quad=P(A=a\mid\lambda)\big[P\big(\{D_2=+1\}\cap\{B=b\}\mid\lambda\big)+P\big(\{D_2=+1\}\cap\{B=b\}\mid\lambda\big)\big] \\ &\qquad=P(A=a\mid\lambda)P(B=b\mid\lambda). \end{aligned}$$

Now, by applying (2.3) and then (2.1), we have

$$P(A=a\mid\lambda)P(B=b\mid\lambda)=P(A=a)P(B=b)=P\big((A,B)=(a,b)\big).$$

□

2.3 The Axioms When "λ" Is Finite

One facilitating strategy for working with hidden variables is to assume that they are finite. This is done, for example, in [11]. This makes the exposition cleaner and easier to understand, while usually not sacrificing much, in terms of what a hidden-variable model can do: in many cases,

finite hidden-variable models can model any empirical distribution that can be modeled by a more general hidden variable. (One such case will be explored in Chapter 3.) In this section, we explore the consequences of assumption that λ is finite, and show how this relates our set of assumptions to the system of [11]. The application of the finite-λ assumption also has the benefit of allowing us to reformulate the axiomatization of Section 2.1 in a way that makes the locality assumption far more intuitive and clear.

For this section, we make the assumption that the random variable λ, introduced in Section 2.1, is of the form

$$\lambda : \Omega \to \Lambda = \{\lambda_1, ..., \lambda_n\}, \tag{2.7}$$

so the change is that Λ is now taken to be a finite set containing n elements. (The nature of its constituent elements λ_i is not characterized, or important.) We also assume that for all i, $P(\lambda = \lambda_i) > 0$; if any of these events has zero probability, it will have no observable effect on the behavior of the model, so such events can be removed from consideration. Henceforth we will use λ_i to refer to the set $\lambda^{-1}(\lambda_i) \subseteq \Omega$, unless doing so could create ambiguity.

Assumption (2.7) simplifies the way we can express two of the other expressions. First, the expression (2.3) is now equivalent to

$$\forall i \in \{1, ..., n\}, \mathbf{a} \in \{a, a'\}, \text{ and } \mathbf{b} \in \{b, b'\},$$
$$P(A = \mathbf{a} \cap \lambda = \lambda_i) = P(A = \mathbf{a})P(\lambda = \lambda_i) \text{ and } P(B = \mathbf{b} \cap \lambda = \lambda_i) = P(B = \mathbf{b})P(\lambda = \lambda_i).$$

Additionally, the assumption that Λ is finite simplifies the interpretation of (2.4), as for any event E, $P(E \mid \lambda)$ will now just be a simple function that is constant on each set λ_i, being equal to $P(E \mid \lambda = \lambda_i)$. Now, (2.4) is equivalent to the condition

$$\forall i \in \{1, ..., n\}, \mathbf{a} \in \{a, a'\}, \mathbf{b} \in \{b, b'\}, i_a \in \{+_1, -_1\}, \text{ and } i_b \in \{+_2, -_2\},$$
$$P(i_a, \mathbf{a}, i_b, \mathbf{b} \mid \lambda_i) = P(i_a, \mathbf{a} \mid \lambda_i)P(i_b, \mathbf{b} \mid \lambda_i). \tag{2.8}$$

So we see that whereas before, some of the assumptions involved unwieldy measure-theoretic concepts, now all of the assumptions (2.1) - (2.4) can be expressed in terms of elementary probabilistic statements concerning a finite collection of events. This makes it vastly easier to work with the assumptions.

For example, in this setting it is not that hard to show that the axiomatization of Section 2.1 is essentially equivalent to the axiomatization of Brandenburger and Yanofsky [11], applied to the relevant setting. To do this, note that if we assume (2.1), (2.2), and replace (2.3) with the stronger (2.5), then the following statement,

$$\forall i, \mathbf{a}, \mathbf{b}, i_a, i_b, \quad \frac{P(i_a, \mathbf{a}, i_b, \mathbf{b}, \lambda_i)}{P(\lambda_i)} = \frac{P(i_a, \mathbf{a}, \lambda_i)P(i_b, \mathbf{b}, \lambda_i)}{P(\lambda_i)^2} \tag{2.9}$$

is equivalent to

$$\forall i, \mathbf{a}, \mathbf{b}, i_a, i_b, \quad \frac{P(i_a, \mathbf{a}, i_b, \mathbf{b}, \lambda_i)}{P(\mathbf{a}, \mathbf{b}, \lambda_i)} = \frac{P(i_a, \mathbf{a}, \lambda_i)P(i_b, \mathbf{b}, \lambda_i)}{P(\mathbf{a}, \lambda_i)P(\mathbf{b}, \lambda_i)}. \tag{2.10}$$

Demonstrating the above biconditional is a straightforward exercise. Note that (2.9) is equivalent to (2.8). Now, recall that by Proposition 2.2.1, we have the following logical relationship between assumptions,

$$(2.1), (2.2), (2.3), (2.4) \quad \Leftrightarrow \quad (2.1), (2.2), (2.5), (2.4),$$

and in the finite-λ setting, we have (2.4) $\Leftrightarrow$ (2.8), so we can say that

$$(2.1), (2.2), (2.5), (2.4) \quad \Leftrightarrow \quad (2.1), (2.2), (2.5), (2.10).$$

The collection of assumptions on the right side of the above equivalence is closely related to the framework of Brandenburger and Yanofsky [11] as it would apply to the 2-detector, 2-setting, 2-outcome scenario. (2.5) is equivalent to Definition 2.4 ("λ-independence") in [11], and (2.10) is equivalent to Definition 2.10 ("locality") in [11]. So we see that in a finite-λ setting, our framework – i.e., the set of conditions (2.1)-(2.4) – is equivalent to the Brandenburger/Yanofsky framework applied to a 2-detector, 2-setting, 2-outcome scenario where measurement choices are independent from each other (the condition (2.1)) and none of the measurement settings have zero probability (the condition (2.2)).

2.4 Comparison To An Alternate Axiomatization

Recently, Brandenburger & Keisler introduced in [12] a measure-theoretic framework for analyzing hidden variable models. In this section, we demonstrate that the Brandenburger & Keisler (B/K) framework can be related to the framework of Section 2.1. The results are similar to the previous section: if we restrict the B/K formalism to the 2-detector, 2-setting, 2-outcome scenario,

we can encode their system within the system of Section 2.1. Once this is done, if we assume (2.1) and (2.2) (independence of measurement settings from each other and non-trivial setting probabilities), then conditions (2.3) and (2.4) can be shown to be equivalent to the corresponding λ-independence and locality notions in the B/K system.

Let us examine the system of Brandenburger & Keisler. The subscript "bk" is used to distinguish these definitions from the definitions of Section 2.1, wherever symbols happen to overlap.

Definition 2.4.1. Let $\Omega_{bk} = \mathcal{X} \otimes \mathcal{Y} \otimes \mathcal{L}$, where $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{L}$ are σ-algebras over sets X, Y, and Λ_{bk}, respectively. X, Y, and Λ_{bk} are called the *outcome space*, the *measurement setting space*, and the *hidden variable space*. A *hidden variable model* is a probability measure P on Ω_{bk}.

The B/K system can apply generally to Bell scenarios for various different numbers of outcomes and/or measurement settings. For our purposes, we want to apply the above definition to the scenario being modeled in Section 2.1. Hence we want to realize the four possibilities for D_1 and D_2 in the outcome space X and the four possibilities for A and B in the measurement space Y:

$$\begin{aligned} X &= \{(+1,+1),(+1,-1),(-1,+1),(-1,-1)\} \\ Y &= \{(a,b),(a,b'),(a',b),(a',b')\}. \end{aligned}$$

We write these as ordered pairs, so X and Y can also be represented as Cartesian products:

$$\begin{aligned} X = X_a \times X_b, \qquad X_a = \{+1,-1\}, \quad X_b = \{+1,-1\} \\ Y = Y_a \times Y_b, \qquad Y_a = \{a,a'\}, \quad Y_b = \{b,b'\}. \end{aligned} \tag{2.11}$$

This enables us to separate what is going on at detector 1 (subscript a) from what is going on at detector 2 (subscript b). This is necessary in order to write the B/K formulation of the locality assumption. We also need the following notation, which is used extensively in [12]:

Definition 2.4.2. Suppose P is a probability measure on a product space $\mathcal{X} \otimes \mathcal{Y} \otimes \mathcal{Z}$ and $J \in \mathcal{X}$. Then

$$p[J||Z] := E[I_{J\times Y\times Z} \mid \{X \times Y, \emptyset\} \otimes \mathcal{Z}].$$

Note that $\{X \times Y, \emptyset\} \otimes \mathcal{Z}$ is the σ-algebra consisting of exactly the sets $X \times Y \times Z_i$, where

$Z_i \in \mathcal{Z}$. We introduce a helpful shorthand $\mathcal{Z}^*$ for this type of σ-algebra:

$$\mathcal{Z}^* = \{X \times Y, \emptyset\} \otimes \mathcal{Z}.$$

Now we can present the axioms of the B/K system.

Definition 2.4.3. A hidden variable model p is λ-*independent*$_{bk}$ if for every event $L \in \mathcal{L}$,

$$p[L||\mathcal{Y}] = P(X \times Y \times L). \tag{2.12}$$

A hidden variable model p is *local*$_{bk}$ if for every $(i_a, i_b) \in X$, we have

$$p[(i_a, i_b)||\mathcal{Y} \otimes \mathcal{L}] = p[i_a||\mathcal{Y}_a \otimes \mathcal{L}] \times p[i_b||\mathcal{Y}_b \otimes \mathcal{L}]. \tag{2.13}$$

Our task for the rest of this section is to model the B/K system within the system of Section 2.1, and then to show that if we assume (2.1) and (2.2), the notions of λ-independence$_{bk}$ and locality$_{bk}$ are equivalent to assumptions (2.3) and (2.4), respectively.

To model the B/K system within the system of Section 2.1, return to Definition 2.1.1. Here, take Ω to be $X \times Y \times \Lambda_{bk}$. Then, let Λ equal Λ_{bk} so λ is a map from Ω to $\Lambda = \Lambda_{bk}$. With Ω now equaling $X \times Y \times \Lambda$, take λ to be the projection onto Λ, so $\lambda : \Omega \to \Lambda$ is equal to $\pi_\Lambda(x, y, l) = l$, where x, y, and l are elements of X, Y, and Λ. It is straighforward to check that $\sigma(\lambda)$ will be equal to the collection of sets $\{X \times Y \times L \mid L \in \mathcal{L}\} = \mathcal{L}^*$. This is the key to switching between the system of Section 2.1, where λ-independence and locality are defined in terms of $\sigma(\lambda)$, and the B/K system, where corresponding notions are defined using the $p[J||Z]$ operator.

Now that Ω is taken to be a Cartesian product, we want to make the following associations:

$$\begin{aligned}
\forall i_a, i_b \in \{+1, -1\}, \mathbf{a} \in \{a, a'\}, \mathbf{b} \in \{b, b'\}, & \\
\{D_1 = i_a\} &= \{i_a\} \times X_b \times Y \times \Lambda \\
\{D_2 = i_b\} &= X_a \times \{i_b\} \times Y \times \Lambda \\
\{A = \mathbf{a}\} &= X \times \{\mathbf{a}\} \times Y_b \times \Lambda \\
\{B = \mathbf{b}\} &= X \times Y_a \times \{\mathbf{b}\} \times \Lambda.
\end{aligned}$$

It is straightforward to achieve this by taking D_1, D_2, A, and B to be projection functions onto the

appropriate component of $\Omega = X \times Y \times \Lambda$.

The previous two paragraphs encode the B/K system into the system of Section 2.1. Note that this requires putting some constraints on the space Ω, which implies that the system of Section 2.1 is at least as general as the B/K system, and possibly more general. There may be a different construction could encode the system of Section 2.1 within the B/K system, in which case both systems would be equally expressive. Whether such an encoding exists is an open problem.

It remains to be demonstrated that the corresponding definitions of λ-independence and locality are equivalent. To show this, we start by using the current definitions of Ω, D_1, D_2, A, and B to re-write the assumptions (2.1)-(2.4) in a way that exploits the new structure imposed on Ω. In doing this, we will freely use some facts relating Cartesian products and set operations: in a product set $X \times Y$, if $A \subseteq X$ and $B, C \subseteq Y$, then

$$\begin{aligned} A \times B &= (X \times B) \cap (A \times Y) \\ A \times (B \cap C) &= (A \times B) \cap (A \times C). \end{aligned}$$

These facts are basic, but we point them out explicitly as they are used frequently throughout the rest of this section.

Now assumptions (2.1)-(2.4) are now equivalent to, in order,

$$\begin{gathered} \forall \mathbf{a} \in \{a, a'\}, \mathbf{b} \in \{b, b'\}, \\ P\left(X \times (\mathbf{a}, \mathbf{b}) \times \Lambda\right) = P(X \times \{\mathbf{a}\} \times Y_b \times \Lambda) P(X \times Y_a \times \{\mathbf{b}\} \times \Lambda) \end{gathered} \tag{2.14}$$

$$\begin{aligned} \forall \mathbf{a} \in \{a, a'\}, \quad P(X \times \{\mathbf{a}\} \times Y_b \times \Lambda) &> 0 \\ \forall \mathbf{b} \in \{b, b'\}, \quad P(X \times Y_a \times \{\mathbf{b}\} \times \Lambda) &> 0 \end{aligned} \tag{2.15}$$

$$\begin{aligned} \forall \mathbf{a} \in \{a, a'\}, L \in \Lambda, \quad P(X \times \{\mathbf{a}\} \times Y_b \times L) &= P(X \times \{\mathbf{a}\} \times Y_b \times \Lambda) P(X \times Y \times L) \\ \forall \mathbf{b} \in \{b, b'\}, L \in \Lambda, \quad P(X \times Y_a \times \{\mathbf{b}\} \times L) &= P(X \times Y_a \times \{\mathbf{b}\} \times \Lambda) P(X \times Y \times L) \end{aligned} \tag{2.16}$$

$$\forall i_a, i_b \in \{+1, -1\}, \mathbf{a} \in \{a, a'\}, \mathbf{b} \in \{b, b'\},$$

$$p[(i_a, i_b) \times (\mathbf{a}, \mathbf{b})||\mathcal{L}] = p[\{i_a\} \times \{\mathbf{a}\}||\mathcal{L}] \cdot p[\{i_b\} \times \{\mathbf{b}\}||\mathcal{L}]. \tag{2.17}$$

If (2.14) and (2.15) hold, then conditions (2.12) and (2.13) are equivalent to (2.16) and (2.17). The proof of this is involved, so for clarity we divide it into two propositions.

Proposition 2.4.1. *If* (2.14) *and* (2.15) *hold, then conditions* (2.12) *and* (2.13) *imply* (2.16) *and* (2.17).

Proof. We first derive (2.16). To do this, note that for any fixed choice $(\mathbf{a}, \mathbf{b}) \in Y$,

$$\begin{aligned}
P(X \times (\mathbf{a}, \mathbf{b}) \times L) &= \int I_{X \times (\mathbf{a},\mathbf{b}) \times L} dP \\
&= \int I_{X \times (\mathbf{a},\mathbf{b}) \times \Lambda} I_{X \times Y \times L} dP \\
&= \int_{X \times (\mathbf{a},\mathbf{b}) \times \Lambda} I_{X \times Y \times L} dP \\
&= \int_{X \times (\mathbf{a},\mathbf{b}) \times \Lambda} E(I_{X \times Y \times L} \mid \mathcal{Y}^*) dP \\
&= \int_{X \times (\mathbf{a},\mathbf{b}) \times \Lambda} p[L||\mathcal{Y}] dP \\
&= \int_{X \times (\mathbf{a},\mathbf{b}) \times \Lambda} P(X \times Y \times L) dP,
\end{aligned}$$

where the last equality follows from (2.12). This implies that

$$P(X \times (\mathbf{a}, \mathbf{b}) \times L) = P(X \times (\mathbf{a}, \mathbf{b}) \times \Lambda) P(X \times Y \times L), \tag{2.18}$$

which will be useful. Now, starting with the left side of (2.16), we can use (2.18) to obtain

$$\begin{aligned}
P(X \times \{\mathbf{a}\} \times Y_b \times L) &= P(X \times (\mathbf{a}, b) \times L) + P(X \times (\mathbf{a}, b') \times L) \\
&= P(X \times (\mathbf{a}, b) \times \Lambda) P(X \times Y \times L) + P(X \times (\mathbf{a}, b') \times \Lambda) P(X \times Y \times L) \\
&= P(X \times \{\mathbf{a}\} \times Y_b \times \Lambda) P(X \times Y \times L).
\end{aligned}$$

The same argument works for $P(X \times Y_a \times \{\mathbf{b}\} \times L)$, so we see that (2.16) holds. Deriving (2.17) is more involved. To do this, we first assert that for any fixed choice of $(i_a, i_b, \mathbf{a}, \mathbf{b})$, the

function $p[(i_a, i_b) \times (\mathbf{a}, \mathbf{b})||\mathcal{L}]$ can be taken to be the function $\eta_{i_a i_b \mathbf{ab}} : \Omega \to \mathbb{R}$ defined as follows:

$$\forall (x, y, l) \in X \times Y \times \Lambda, \qquad \eta_{i_a i_b \mathbf{ab}}(x, y, l) := P(X \times (\mathbf{a}, \mathbf{b}) \times \Lambda) \eta_{i_a i_b}(x, (\mathbf{a}, \mathbf{b}), l), \tag{2.19}$$

where $\eta_{i_a i_b} : \Omega \to \mathbb{R}$ is the function $p[(i_a, i_b)||\mathcal{Y} \otimes \mathcal{L}]$. Note that in (2.19), the free coordinate y is replaced with the fixed choice $(\mathbf{a}, \mathbf{b})$ when the map $\eta_{i_a i_b}$ is applied. To prove the assertion that $p[(i_a, i_b) \times (\mathbf{a}, \mathbf{b})||\mathcal{L}]$ can be defined by (2.19), we must show that $\eta_{i_a i_b \mathbf{ab}}$ is measurable with respect to $\mathcal{L}^*$ and that $\int_{X \times Y \times L} \eta_{i_a i_b \mathbf{ab}} dP = \int_{X \times Y \times L} I_{(i_a, i_b) \times (\mathbf{a}, \mathbf{b}) \times \Lambda} dP$ for all $L \in \mathcal{L}$. To see the measurability, first note that for any fixed $\beta \in \mathbb{R}$, $\eta^{-1}_{i_a i_b}(\beta, \infty)$ will be a set $S_\beta \in (\mathcal{Y} \otimes \mathcal{L})^*$. It is straightforward to check that any set in $(\mathcal{Y} \otimes \mathcal{L})^*$ can be decomposed into a disjoint union as follows:

$$S_\beta = [X \times (a, b) \times L_{ab}] \cup [X \times (a, b') \times L_{ab'}] \cup [X \times (a', b) \times L_{a'b}] \cup [X \times (a', b') \times L_{a'b'}], \tag{2.20}$$

where each L_{xy} is a set in $\mathcal{L}$. This decomposition will be handy. Now, let us examine the set $\eta^{-1}_{i_a i_b \mathbf{ab}}(\alpha, \infty)$. For a fixed point $(x, y, l) \in X \times Y \times L$, $\eta_{i_a i_b \mathbf{ab}}$ maps (x, y, l) into the interval $(\alpha, \infty) \subseteq \mathbb{R}$ if and only if $\eta_{i_a i_b}$ maps $(x, (\mathbf{a}, \mathbf{b}), l)$ into the interval (β, ∞) where $\beta = \alpha \cdot P(X \times (\mathbf{a}, \mathbf{b}) \times \Lambda)^{-1}$. Hence if we look back to expression (2.20), we will have

$$\begin{aligned} &\eta^{-1}_{i_a i_b \mathbf{ab}}(\alpha, \infty) \\ &\quad = [X \times (a, b) \times L_{\mathbf{ab}}] \cup [X \times (a, b') \times L_{\mathbf{ab}}] \cup [X \times (a', b) \times L_{\mathbf{ab}}] \cup [X \times (a', b') \times L_{\mathbf{ab}}], \end{aligned} \tag{2.21}$$

where $L_{\mathbf{ab}}$ will be a particular choice of the four L_{xy} sets appearing in (2.20), determined by whether $\mathbf{ab}$ is ab, ab', $a'b$, or $a'b'$. From (2.21), it follows that

$$\eta^{-1}_{i_a i_b \mathbf{ab}}(\alpha, \infty) = X \times Y \times L_{\mathbf{ab}}, \tag{2.22}$$

and $X \times Y \times L_{\mathbf{ab}}$ is an element of $\mathcal{L}^*$.

Now we need to show that $\int_{X \times Y \times L} \eta_{i_a i_b \mathbf{ab}} dP = \int_{X \times Y \times L} I_{(i_a, i_b) \times (\mathbf{a}, \mathbf{b}) \times \Lambda} dP$ for all $L \in \mathcal{L}$. We have

$$\begin{aligned} \int_{X \times Y \times L} \eta_{i_a i_b \mathbf{ab}} dP &= \sum_{y_1 \in \{a, a'\}, y_2 \in \{b, b'\}} \int_{X \times (y_1, y_2) \times L} \eta_{i_a i_b \mathbf{ab}} dP \\ &= P(X \times (\mathbf{a}, \mathbf{b}) \times \Lambda) \sum_{y_1 \in \{a, a'\}, y_2 \in \{b, b'\}} \int_{X \times (y_1, y_2) \times L} \eta^{(\mathbf{ab})}_{i_a i_b} dP, \end{aligned} \tag{2.23}$$

where $\eta^{(\mathbf{ab})}_{i_a i_b}(x,y,l) := \eta_{i_a i_b}(x,(\mathbf{a},\mathbf{b}),l)$. Now, let $\{\eta_j\}$ be a $(\mathcal{Y}\otimes\mathcal{L})^*$-measurable sequence of simple functions converging monotonically (from below) to $\eta_{i_a i_b}$. The existence of such a sequence is guaranteed by Lemma 2.11 of [8], given the almost-everywhere nonnegativity of $\eta_{i_a i_b}$. Hence each η_j can be written as a finite sum, $\sum_k c_k I_{S^k}$, where S^k is a set in $(\mathcal{Y}\otimes\mathcal{L})^*$ and therefore can be decomposed as is done in (2.20). (The index k clearly depends on j, but writing k_j everywhere that k occurs would create a lot of clutter, so we omit writing the subscript j every time with the tacit understanding that it is there.)

With the $\{\eta_j\}$ sequence so defined, the Monotone Convergence Theorem ([8], Theorem 4.6) will allow us to exchange limits and integrals, so we have

$$\begin{aligned}\int_{X\times(y_1,y_2)\times L}\eta^{(\mathbf{ab})}_{i_a i_b}dP &= \int_{X\times(y_1,y_2)\times L}\lim_j \eta^{(\mathbf{ab})}_j dP \\ &= \lim_j \int_{X\times(y_1,y_2)\times L}\eta^{(\mathbf{ab})}_j dP \\ &= \lim_j \int_{X\times(y_1,y_2)\times L}\sum_k c_k I^{(\mathbf{ab})}_{S^k} dP.\end{aligned}$$

We read $I^{(\mathbf{ab})}_{S^k}$ as $I^{(\mathbf{ab})}_{S^k}(x,y,l) := I_{S^k}(x,(\mathbf{a},\mathbf{b}),l)$. As each S^k set can be decomposed in the manner of (2.20), we can conclude that $I^{(\mathbf{ab})}_{S^k}$ is equal to $I_{X\times Y\times L^k_{\mathbf{ab}}}$, by the same type of reasoning that resulted in expression (2.22). Hence

$$\begin{aligned}\int_{X\times(y_1,y_2)\times L}\eta^{(\mathbf{ab})}_{i_a i_b}dP &= \lim_j \sum_k c_k \int_{X\times(y_1,y_2)\times L} I_{X\times Y\times L^k_{\mathbf{ab}}}dP \\ &= \lim_j \sum_k c_k P\left(X\times(y_1,y_2)\times[L^k_{\mathbf{ab}}\cap L]\right) \\ &= \lim_j \sum_k c_k P\left(X\times Y\times[L^k_{\mathbf{ab}}\cap L]\right) P\left(X\times(y_1,y_2)\times\Lambda\right),\end{aligned}$$

where the last equality follows from (2.18). Now, returning to (2.23), note that

$$P(X\times(\mathbf{a},\mathbf{b})\times\Lambda)P\left(X\times Y\times[L^k_{\mathbf{ab}}\cap L]\right) = P\left(X\times(\mathbf{a},\mathbf{b})\times[L^k_{\mathbf{ab}}\cap L]\right),$$

so we can write

$$\begin{aligned}
\int_{X\times Y\times L}\eta_{i_a i_b\mathbf{ab}}dP &= \sum_{y_1\in\{a,a'\},y_2\in\{b,b'\}}\lim_j\sum_k c_k P\left(X\times(\mathbf{a},\mathbf{b})\times[L^k_{\mathbf{ab}}\cap L]\right)P\left(X\times(y_1,y_2)\times\Lambda\right)\\
&= \sum_{y_1\in\{a,a'\},y_2\in\{b,b'\}}P\left(X\times(y_1,y_2)\times\Lambda\right)\lim_j\sum_k c_k P\left(X\times(\mathbf{a},\mathbf{b})\times[L^k_{\mathbf{ab}}\cap L]\right)\\
&= \lim_j\sum_k c_k P\left(X\times(\mathbf{a},\mathbf{b})\times[L^k_{\mathbf{ab}}\cap L]\right)\\
&= \lim_j\sum_k c_k\int I_{X\times(\mathbf{a},\mathbf{b})\times[L^k_{\mathbf{ab}}\cap L]}dP.
\end{aligned}$$

Now, note that

$$\begin{aligned}
X\times(\mathbf{a},\mathbf{b})\times[L^k_{\mathbf{ab}}\cap L] &= (X\times(\mathbf{a},\mathbf{b})\times L)\cap(X\times(\mathbf{a},\mathbf{b})\times L^k_{\mathbf{ab}})\\
&= (X\times(\mathbf{a},\mathbf{b})\times L)\cap(S^k),
\end{aligned}$$

which again follows from the disjoint decomposition of S^k in the manner of (2.20), and the fact that only the term with $L^k_{\mathbf{ab}}$ remains from S^k if we intersect S^k with a set of the form $(X\times(\mathbf{a},\mathbf{b})\times L)$. This allows us to write

$$\begin{aligned}
\int_{X\times Y\times L}\eta_{i_a i_b\mathbf{ab}}dP &= \lim_j\int_{X\times(\mathbf{a},\mathbf{b})\times L}\sum_k c_k I_{S^k}dP\\
&= \lim_j\int_{X\times(\mathbf{a},\mathbf{b})\times L}\eta_j dP\\
&= \int_{X\times(\mathbf{a},\mathbf{b})\times L}\lim_j\eta_j dP\\
&= \int_{X\times(\mathbf{a},\mathbf{b})\times L}\eta_{i_a i_b}dP\\
&= \int_{X\times(\mathbf{a},\mathbf{b})\times L}I_{(i_a,i_b)\times Y\times\Lambda}dP\\
&= \int I_{(i_a,i_b)\times(\mathbf{a},\mathbf{b})\times L}dP\\
&= \int_{X\times Y\times L}I_{(i_a,i_b)\times(\mathbf{a},\mathbf{b})\times\Lambda}dP.
\end{aligned}$$

This completes the proof that $p[(i_a,i_b)\times(\mathbf{a},\mathbf{b})||\mathcal{L}]$ can be taken to be $\eta_{i_a i_b\mathbf{ab}}$.

With this definition of $p[(i_a, i_b) \times (\mathbf{a}, \mathbf{b})||\mathcal{L}]$, we can demonstrate (2.16). We have

$$\begin{aligned} p[(i_a, i_b) \times (\mathbf{a}, \mathbf{b})||\mathcal{L}] &= P(X \times (\mathbf{a}, \mathbf{b}) \times \Lambda) \cdot p[(i_a, i_b)||\mathcal{Y} \otimes \mathcal{L}]^{(\mathbf{a},\mathbf{b})} \\ &= P(X \times \{\mathbf{a}\} \times Y_b \times \Lambda) P(X \times Y_a \times \{\mathbf{b}\} \times \Lambda) p[(i_a||\mathcal{Y}_a \otimes \mathcal{L}]^{(\mathbf{a},\mathbf{b})} p[(i_b||\mathcal{Y}_b \otimes \mathcal{L}]^{(\mathbf{a},\mathbf{b})}, \end{aligned}$$

which follows from(2.13) and (2.14). Then it is true that

$$p[(i_a||\mathcal{Y}_a \otimes \mathcal{L}]^{(\mathbf{a},\mathbf{b})} = p[(i_a||\mathcal{Y}_a \otimes \mathcal{L}]^{\mathbf{a}} \tag{2.24}$$

$$p[(i_b||\mathcal{Y}_b \otimes \mathcal{L}]^{(\mathbf{a},\mathbf{b})} = p[(i_b||\mathcal{Y}_b \otimes \mathcal{L}]^{\mathbf{b}}. \tag{2.25}$$

This is because the $(\mathcal{Y}_{y_i} \otimes \mathcal{L})^*$-measurability of the respective functions implies that the $\mathbf{b}$ coordinate does not affect the value of (2.24) and the $\mathbf{a}$ coordinate does not affect the value of (2.25). To illustrate why this is, consider a function $f : \{y, y'\} \times Z \to \mathbb{R}$ that is measurable with respect to a σ-algebra $\mathcal{Z}^*$, which consists of sets of the form $\big\{\{y, y'\} \times Z_i \mid Z_i \in \mathcal{Z}\big\}$ where $\mathcal{Z}$ is a σ-algebra over Z. Now for any point (y, z), $z \in Z$, we can see that for any $r \in \mathbb{R}$, $f(y, z) > r \Leftrightarrow f(y', z) > r$, because $f^{-1}(r, \infty)$ is a set of the form $\{y, y'\} \times Z_i$ by the $\mathcal{Z}^*$-measurability of f. This implies that $f(y, z) = f(y', z)$, so swapping y for y' cannot change the value of f. By this same argument, we can freely toggle the appropriate coordinate in (2.24) and (2.24).

Finally, we have

$$\begin{aligned} P(X \times \{\mathbf{a}\} \times Y_b \times \Lambda) p[i_a||\mathcal{Y}_a \otimes \mathcal{L}]^{\mathbf{a}} &= p[\{i_a\} \times \{\mathbf{a}\}||\mathcal{L}] \\ P(X \times Y_a \times \{\mathbf{b}\} \times \Lambda) p[i_b||\mathcal{Y}_b \otimes \mathcal{L}]^{\mathbf{b}} &= p[\{i_b\} \times \{\mathbf{b}\}||\mathcal{L}], \end{aligned}$$

because $p[\{i_a\} \times \{\mathbf{a}\}||\mathcal{L}]$ and $p[\{i_b\} \times \{\mathbf{b}\}||\mathcal{L}]$ can be defined in terms of $p[i_a||\mathcal{Y}_a \otimes \mathcal{L}]$ and $p[i_b||\mathcal{Y}_b \otimes \mathcal{L}]$ using the same argument that defined $p[(i_a, i_b) \times (\mathbf{a}, \mathbf{b})||\mathcal{L}]$ in terms of $p[(i_a, i_b)||\mathcal{Y} \otimes \mathcal{L}]$. Hence, (2.16) holds. □

Proposition 2.4.2. *(Converse) If* (2.14) *and* (2.15) *hold, then conditions* (2.16) *and* (2.17) *imply* (2.12) *and* (2.13)*.*

Proof. We start by demonstrating (2.12). Note that because the set Y only consists of four elements, the σ-algebra $\mathcal{Y}^*$ will also be finite; its atomic sets are sets of the form $X \times (\mathbf{a}, \mathbf{b}) \times \Lambda$. Thus we can

demonstrate (2.12) by showing that

$$\int_{X\times(\mathbf{a},\mathbf{b})\times\Lambda} P(X\times Y\times L)dP = \int_{X\times(\mathbf{a},\mathbf{b})\times\Lambda} I_{X\times Y\times L}dP \tag{2.26}$$

for a fixed choice of $(\mathbf{a},\mathbf{b})$. First, we have

$$\int_{X\times(\mathbf{a},\mathbf{b})\times\Lambda} P(X\times Y\times L)dP = P(X\times(\mathbf{a},\mathbf{b})\times\Lambda)P(X\times Y\times L).$$

Now recall Proposition 2.2.1, which in this setting can be interpreted as stating that (2.14) and (2.17) imply the relation

$$P(X\times(\mathbf{a},\mathbf{b})\times\Lambda)P(X\times Y\times L) = P(X\times(\mathbf{a},\mathbf{b})\times L). \tag{2.27}$$

From this,

$$\begin{aligned}\int_{X\times(\mathbf{a},\mathbf{b})\times\Lambda} P(X\times Y\times L) &= P(X\times(\mathbf{a},\mathbf{b})\times L)\\ &= \int I_{X\times(\mathbf{a},\mathbf{b})\times L}dP\\ &= \int I_{X\times(\mathbf{a},\mathbf{b})\times\Lambda}I_{X\times Y\times L}dP\\ &= \int_{X\times(\mathbf{a},\mathbf{b})\times\Lambda} I_{X\times Y\times L}dP.\end{aligned}$$

This completes the derivation of (2.12). To derive (2.13), we must relate $p[(i_a,i_b)||\mathcal{Y}\otimes\mathcal{L}]$ to $p[(i_a,i_b)\times(\mathbf{a},\mathbf{b})||\mathcal{L}]$ by way of a new definition. We assert that $p[(i_a,i_b)||\mathcal{Y}\otimes\mathcal{L}]$ can be taken to be the function $\eta_{i_ai_b}$ defined as follows:

$$\eta_{i_ai_b}(x,y,l) := \begin{cases} \frac{p[(i_a,i_b)\times(a,b)||\mathcal{L}](x,y,l)}{P(X\times(a,b)\times\Lambda)} & \text{if } y=(a,b)\\ \frac{p[(i_a,i_b)\times(a,b')||\mathcal{L}](x,y,l)}{P(X\times(a,b')\times\Lambda)} & \text{if } y=(a,b')\\ \frac{p[(i_a,i_b)\times(a',b)||\mathcal{L}](x,y,l)}{P(X\times(a',b)\times\Lambda)} & \text{if } y=(a',b)\\ \frac{p[(i_a,i_b)\times(a',b')||\mathcal{L}](x,y,l)}{P(X\times(a',b')\times\Lambda)} & \text{if } y=(a',b'). \end{cases} \tag{2.28}$$

To prove the assertion, we must demonstrate that $\eta_{i_ai_b}$ is measurable with respect to $(\mathcal{Y}\times\mathcal{L})^*$ and that $\int_S \eta_{i_ai_b}dP = \int_S I_{(i_ai_b)\times Y\times\Lambda}dP$ for all $S\in(\mathcal{Y}\times\mathcal{L})^*$.

To see $(\mathcal{Y}\times\mathcal{L})^*$-measurability, we see that for fixed $\alpha\in\mathbb{R}$, $\eta_{i_ai_b}^{-1}(\alpha,\infty)$ is equal to the set

$$\begin{aligned}
&\left[p[(i_a,i_b)\times(a,b)||\mathcal{L}]^{-1}(\beta_{ab},\infty)\cap(X\times(a,b)\times\Lambda)\right]\\
\cup\ &\left[p[(i_a,i_b)\times(a,b')||\mathcal{L}]^{-1}(\beta_{a'b},\infty)\cap(X\times(a,b')\times\Lambda)\right]\\
\cup\ &\left[p[(i_a,i_b)\times(a',b)||\mathcal{L}]^{-1}(\beta_{ab'},\infty)\cap(X\times(a',b)\times\Lambda)\right]\\
\cup\ &\left[p[(i_a,i_b)\times(a',b')||\mathcal{L}]^{-1}(\beta_{a'b'},\infty)\cap(X\times(a',b')\times\Lambda)\right],
\end{aligned}\tag{2.29}$$

where $\beta_{xy}=\alpha\cdot P(X\times(x,y)\times\Lambda)$, and the four terms above are clearly disjoint. As each $p[(i_a,i_b)\times(x,y)||\mathcal{L}]^{-1}(\beta_{xy},\infty)$ is a set of the form $X\times Y\times L_{xy}$ where $L_{xy}\in\mathcal{L}$, expression (2.29) can be written as

$$(X\times(a,b)\times L_{ab})\cup(X\times(a,b')\times L_{ab'})\cup(X\times(a',b)\times L_{a'b})\cup(X\times(a',b')\times L_{a'b'}),\tag{2.30}$$

which is in $(\mathcal{Y}\otimes\mathcal{L})^*$.

To show that $\int_S\eta_{i_ai_b}dP=\int_S I_{(i_ai_b)\times Y\times\Lambda}dP$ for all $S\in(\mathcal{Y}\times\mathcal{L})^*$, note that S can be broken up as in (2.20), so

$$\begin{aligned}
\int_S\eta_{i_ai_b}dP=&\int_{X\times(a,b)\times L_{ab}}\eta_{i_ai_b}dP+\int_{X\times(a,b')\times L_{ab'}}\eta_{i_ai_b}dP\\
&+\int_{X\times(a',b)\times L_{a'b}}\eta_{i_ai_b}dP+\int_{X\times(a',b')\times L_{a'b'}}\eta_{i_ai_b}dP.
\end{aligned}\tag{2.31}$$

Then if we examine any particular one of these terms, we have

$$\begin{aligned}
\int_{X\times(\mathbf{a},\mathbf{b})\times L_{\mathbf{ab}}}\eta_{i_ai_b}dP&=\int_{X\times Y\times L_{\mathbf{ab}}}I_{X\times(\mathbf{a},\mathbf{b})\times\Lambda}\eta_{i_ai_b}dP\\
&=P(X\times(\mathbf{a},\mathbf{b})\times\Lambda)^{-1}\int_{X\times Y\times L_{\mathbf{ab}}}I_{X\times(\mathbf{a},\mathbf{b})\times\Lambda}p[(i_a,i_b)\times(\mathbf{a},\mathbf{b})||\mathcal{L}]dP.
\end{aligned}$$

Now we again make use of the fact that there is a monotone increasing sequence $\{\eta_j\}$ of simple,

$\mathcal{L}^*$-measurable functions for which η_j converges to $p[(i_a, i_b) \times (\mathbf{a}, \mathbf{b})||\mathcal{L}]$ from below. Hence

$$\begin{aligned}
\int_{X\times(\mathbf{a},\mathbf{b})\times L_{\mathbf{ab}}} \eta_{i_a i_b} dP &= P(X \times (\mathbf{a}, \mathbf{b}) \times \Lambda)^{-1} \int_{X\times Y\times L_{\mathbf{ab}}} I_{X\times(\mathbf{a},\mathbf{b})\times\Lambda} \lim_j \eta_j dP \\
&= P(X \times (\mathbf{a}, \mathbf{b}) \times \Lambda)^{-1} \lim_j \int_{X\times Y\times L_{\mathbf{ab}}} I_{X\times(\mathbf{a},\mathbf{b})\times\Lambda} \eta_j dP \\
&= P(X \times (\mathbf{a}, \mathbf{b}) \times \Lambda)^{-1} \lim_j \int_{X\times Y\times L_{\mathbf{ab}}} I_{X\times(\mathbf{a},\mathbf{b})\times\Lambda} \sum_k c_k I_{X\times Y\times L_k} dP \\
&= P(X \times (\mathbf{a}, \mathbf{b}) \times \Lambda)^{-1} \lim_j \sum_k c_k P(X \times (\mathbf{a}, \mathbf{b}) \times [L_{\mathbf{ab}} \cap L_k]) \\
&= P(X \times (\mathbf{a}, \mathbf{b}) \times \Lambda)^{-1} \lim_j \sum_k c_k P(X \times Y \times [L_{\mathbf{ab}} \cap L_k]) P(X \times (\mathbf{a}, \mathbf{b}) \times \Lambda) \\
&= \lim_j \sum_k c_k P(X \times Y \times [L_{\mathbf{ab}} \cap L_k]),
\end{aligned}$$

by way of the Monotone Convergence Theorem and (2.27). Continuing, we have

$$\begin{aligned}
\int_{X\times(\mathbf{a},\mathbf{b})\times L_{\mathbf{ab}}} \eta_{i_a i_b} dP &= \lim_j \sum_k c_k \int_{X\times Y\times L_{\mathbf{ab}}} I_{X\times Y\times L_k} dP \\
&= \int_{X\times Y\times L_{\mathbf{ab}}} \lim_j \sum_k c_k I_{X\times Y\times L_k} dP \\
&= \int_{X\times Y\times L_{\mathbf{ab}}} \lim_j \eta_j dP \\
&= \int_{X\times Y\times L_{\mathbf{ab}}} p[(i_a, i_b) \times (\mathbf{a}, \mathbf{b})||\mathcal{L}] dP \\
&= \int_{X\times Y\times L_{\mathbf{ab}}} I_{(i_a,i_b)\times(\mathbf{a},\mathbf{b})\times\Lambda} dP \\
&= P\big((i_a, i_b) \times (\mathbf{a}, \mathbf{b}) \times L_{\mathbf{ab}}\big).
\end{aligned}$$

Returning to (2.31), we can now write

$$\begin{aligned}
\int_S \eta_{i_a i_b} dP &= P\big((i_a, i_b) \times (a, b) \times L_{ab}\big) + P\big((i_a, i_b) \times (a, b') \times L_{ab'}\big) \\
&\quad + P\big((i_a, i_b) \times (a', b) \times L_{a'b}\big) + P\big((i_a, i_b) \times (a', b') \times L_{a'b'}\big) \\
&= P\big([(i_a, i_b) \times Y \times \Lambda] \cap [\cup_{x,y} X \times (x, y) \times L_{xy}]\big) \\
&= P\big([(i_a, i_b) \times Y \times \Lambda] \cap S\big) \\
&= \int_S I_{(i_a,i_b)\times Y\times\Lambda} dP.
\end{aligned}$$

This completes the proof of the claim that $p[(i_a, i_b)||\mathcal{Y} \otimes \mathcal{L}]$ can be taken to be $\eta_{i_a i_b}$ as defined in

(2.28). This allows us to derive (2.13):

$$\begin{aligned} p[(i_a, i_b)||\mathcal{Y} \otimes \mathcal{L}](x, (\mathbf{a}, \mathbf{b}), l) &= \frac{p[(i_a, i_b) \times (\mathbf{a}, \mathbf{b})||\mathcal{L}](x, (\mathbf{a}, \mathbf{b}), l)}{P(X \times (\mathbf{a}, \mathbf{b}) \times \Lambda)} \\ &= \frac{p[\{i_a\} \times \{\mathbf{a}\}||\mathcal{L}](x, (\mathbf{a}, \mathbf{b}), l) \cdot p[\{i_b\} \times \{\mathbf{b}\}||\mathcal{L}](x, (\mathbf{a}, \mathbf{b}), l)}{P(X \times \{\mathbf{a}\} \times Y_b \times \Lambda) \cdot P(X \times Y_a \times \{\mathbf{b}\} \times \Lambda)} \end{aligned}$$

by (2.16) and (2.17). The same argument relating $p[(i_a, i_b)||\mathcal{Y} \otimes \mathcal{L}]$ to $p[(i_a, i_b) \times (\mathbf{a}, \mathbf{b})||\mathcal{L}]$ can be used to relate $p[\{i_a\} \times \{\mathbf{a}\}||\mathcal{L}]$ and $p[\{i_b\} \times \{\mathbf{b}\}||\mathcal{L}]$ to $p[i_a||\mathcal{Y}_a \otimes \mathcal{L}]$ and $p[i_b||\mathcal{Y}_b \otimes \mathcal{L}]$ so that

$$\begin{aligned} \frac{p[\{i_a\} \times \{\mathbf{a}\}||\mathcal{L}](x, (\mathbf{a}, \mathbf{b}), l)}{P(X \times \{\mathbf{a}\} \times Y_b \times \Lambda)} &= p[i_a||\mathcal{Y}_a \otimes \mathcal{L}](x, (\mathbf{a}, \mathbf{b}), l) \\ \frac{p[\{i_b\} \times \{\mathbf{b}\}||\mathcal{L}](x, (\mathbf{a}, \mathbf{b}), l)}{P(X \times Y_a \times \{\mathbf{b}\} \times \Lambda)} &= p[i_b||\mathcal{Y}_b \otimes \mathcal{L}](x, (\mathbf{a}, \mathbf{b}), l), \end{aligned}$$

so (2.13) holds. □

2.5 The Mathematical Development

In this section, we work under the locality assumption (2.4), and derive a constraint on the results of the Bell test experiment. This will be the classical CHSH inequality [2] that was introduced in Chapter 1. Our derivation is a little bit more involved as it uses the more general measure-theoretic framework, whereas the proofs in physics papers tend to tacitly assume that the random variable λ is absolutely continuous (i.e., that it has a probability density function).

To start, suppose we condition on the event that detector 1 is set to "a" and detector 2 is set to "b". Then we can discuss the quantity

$$E_{ab}(D_1 D_2) := E(D_1 D_2 \mid A = a, B = b). \tag{2.32}$$

This is well-defined, in view of (2.2). Note that when writing expectations, we are using subscripts to denote the events on which we are conditioning, and ab in the subscript is a shorthand for $a \cap b$, which in turn is a shorthand for the event $\{A = a\} \cap \{B = b\}$. We will use this shorthand convention freely from now on. Furthermore, our use of subscripts in expressions like (2.32) will help distinguish conditioning on *events* from conditioning on random variables, as we saw in Section 1.3 that these are two distinct notions, and both will be used. The following notation will also be useful in both

Chapters 2 and 3:

$$\mu_X(Y \mid \xi) := \frac{1}{P(X)} P(Y \cap X \mid \xi). \tag{2.33}$$

(2.33) is introduced to approximate an intuitive notion of the probability of event Y that is conditioned simultaneously on the event X *and* the random variable ξ. Using this shorthand, we can write the following expression:

$$E_{ab}(D_1 D_2) = \int [\mu_{ab}(D_1 D_2 = +1 \mid \lambda) - \mu_{ab}(D_1 D_2 = -1 \mid \lambda)] dP. \tag{2.34}$$

The justification of equation (2.34) is given by the following lemma. Note that the proof makes no use of the locality assumption (2.4), so (2.34) will be valid for an especially large class of theories.

Lemma 2.5.1. *Let a, b, D_1, D_2 be as in Definition 2.1.1. Then, under* (2.1) *and* (2.2)*, the equation* (2.34) *holds.*

Proof. By (2.1) and (2.2), $P(a \cap b) > 0$. If we let I_{ab} denote the indicator function of the event $a \cap b$, we can write

$$E_{ab}(D_1 D_2) = \frac{E(I_{ab} D_1 D_2)}{P(ab)}$$

by the definition of conditional expectation (when we condition on events). We can apply the law of iterated expectation to get

$$E(I_{ab} D_1 D_2) = E(E(I_{ab} D_1 D_2 \mid \lambda)).$$

Note that we can be sure that the conditional expectation $E(I_{ab} D_1 D_2 \mid \lambda)$ does indeed exist: $I_{ab} D_1 D_2$ takes values in the set $\{-1, 0, +1\}$, so $E(|I_{ab} D_1 D_2|)$ must be finite. By Theorem 9.1.1 in [10] (a standard result in elementary probability theory), if the (unconditional) expectation of the absolute value of a random variable is finite, then the conditional expectation is guaranteed to exist.

We claim that

$$E(I_{ab} D_1 D_2 \mid \lambda) = E(I_{\{D_1 D_2 = +1\} \cap ab} \mid \lambda) - E(I_{\{D_1 D_2 = -1\} \cap ab} \mid \lambda), \quad a.s. \tag{2.35}$$

To justify this, we recall the definition of conditional expectation: $E(I_{ab} D_1 D_2 \mid \lambda)$ is defined to be a random variable – let us denote it as ξ – that is measurable with respect to $\sigma(\lambda)$ (the σ-algebra

generated by λ), such that for all sets $A \in \sigma(\lambda)$,

$$\int_A \xi dP = \int_A I_{ab} D_1 D_2 dP.$$

So to prove the assertion, we must show that for all $A \in \sigma(\lambda)$,

$$\int_A E(I_{\{D_1D_2=+1\}\cap ab} \mid \lambda) - E(I_{\{D_1D_2=-1\}\cap ab} \mid \lambda) dP = \int_A I_{ab} D_1 D_2 dP.$$

Indeed,

$$\begin{aligned} & \int_A E(I_{\{D_1D_2=+1\}\cap ab} \mid \lambda) - E(I_{\{D_1D_2=-1\}\cap ab} \mid \lambda) dP \\ = & \int_A I_{\{D_1D_2=+1\}\cap ab} - I_{\{D_1D_2=-1\}\cap ab} dP \\ = & (+1)P(\{D_1D_2 = +1\} \cap ab \cap A) + (-1)P(\{D_1D_2 = -1\} \cap ab \cap A) \\ = & \int (D_1D_2) I_A I_{ab} dP = \int_A I_{ab} D_1 D_2 dP, \end{aligned}$$

which proves (2.35). Using this, we have

$$\begin{aligned} E_{ab}(D_1D_2) &= E\bigg(\frac{1}{P(ab)} E(I_{ab} D_1 D_2 \mid \lambda)\bigg) \\ &= E\bigg(\frac{1}{P(ab)} \big[E(I_{\{D_1D_2=+1\}\cap ab} \mid \lambda) - E(I_{\{D_1D_2=-1\}\cap ab} \mid \lambda)\big]\bigg). \end{aligned} \tag{2.36}$$

Recall that by the definition of conditional probability (when conditioning on a random variable),

$$P(\{D_1D_2 = +1\} \cap ab \mid \lambda) = E(I_{\{D_1D_2=+1\}\cap ab} \mid \lambda).$$

Using the notation introduced in (2.33), we have

$$\mu_{ab}(D_1D_2 = +1 \mid \lambda) = \frac{1}{P(ab)} E(I_{\{D_1D_2=+1\}\cap ab} \mid \lambda),$$

so we can rewrite (2.36) as

$$E\bigg(\mu_{ab}(D_1D_2 = +1 \mid \lambda) - \mu_{ab}(D_1D_2 = -1 \mid \lambda)\bigg).$$

Thus, (2.34) holds. □

As we work toward the CHSH inequality, we will have to use the locality assumption, starting with the following proposition. It will be useful to expand expression (2.32), so we introduce a shorthand, for readability:

$$\mathbf{a} = \big[\mu_a(+_1 \mid \lambda) - \mu_a(-_1 \mid \lambda)\big], \quad \mathbf{b} = \big[\mu_b(+_2 \mid \lambda) - \mu_b(-_2 \mid \lambda)\big],$$
$$\mathbf{b}' = \big[\mu_{b'}(+_2 \mid \lambda) - \mu_{b'}(-_2 \mid \lambda)\big], \quad \mathbf{a}' = \big[\mu_{a'}(+_1 \mid \lambda) - \mu_{a'}(-_1 \mid \lambda)\big].$$

Proposition 2.5.1. *Let a, b, D_1, D_2 be as in Definition 2.1.1. Then, under* (2.1), (2.2), (2.4),

$$E_{ab}(D_1 D_2) = \int_\Omega \mathbf{ab}\, dP. \tag{2.37}$$

Proof. Note that

$$\{D_1 D_2 = 1\} = (+_1 \cap +_2) \cup (-_1 \cap -_2),$$
$$\{D_1 D_2 = -1\} = (+_1 \cap -_2) \cup (-_1 \cap +_2).$$

Lemma 1.3.1 can thus be applied to rewrite (2.34) as

$$\int_\Omega \mu_{ab}(+_1 \cap +_2 \mid \lambda) + \mu_{ab}(-_1 \cap -_2 \mid \lambda) - [\mu_{ab}(+_1 \cap -_2 \mid \lambda) + \mu_{ab}(-_1 \cap +_2 \mid \lambda)] dP. \tag{2.38}$$

We now appeal to the locality assumption. Applying (2.4), as well as (2.1), we can modify the terms in the integrand in the following way:

$$\begin{aligned} \mu_{ab}(+_1 \cap +_2 \mid \lambda) &= \frac{1}{P(ab)} P(ab \cap +_1 \cap +_2 \mid \lambda) \\ &= \frac{1}{P(a)P(b)} P(a \cap +_1 \mid \lambda) P(b \cap +_2 \mid \lambda) \\ &= \mu_a(+_1 \mid \lambda)\mu_b(+_2 \mid \lambda). \end{aligned} \tag{2.39}$$

We can obtain relationships like (2.39) for the three other terms in (2.38), and applying these to

(2.38), we get

$$\int_\Omega \mu_a(+_1 \mid \lambda)\mu_b(+_2 \mid \lambda) + \mu_a(-_1 \mid \lambda)\mu_b(-_2 \mid \lambda) - [\mu_a(+_1 \mid \lambda)\mu_b(-_2 \mid \lambda) + \mu_a(-_1 \mid \lambda)\mu_b(+_2 \mid \lambda)]dP$$
$$= \int_\Omega \big[\mu_a(+_1 \mid \lambda) - \mu_a(-_1 \mid \lambda)\big]\big[\mu_b(+_2 \mid \lambda) - \mu_b(-_2 \mid \lambda)\big]dP. \quad (2.40)$$

Thus, (2.37) holds. □

Now, let's consider the following constant:

$$CHSH := E_{ab}(D_1D_2) - E_{a'b}(D_1D_2) + E_{ab'}(D_1D_2) + E_{a'b'}(D_1D_2). \quad (2.41)$$

The quantity $CHSH$, familiar from our discussion in Section 1.1, is useful because it turns out that in the local setting, we can calculate an upper bound that $CHSH$ must obey. (Later we will see how Quantum Mechanics predicts a violation of this upper bound.) This upper bound is developed in Proposition 2.5.2. The proposition requires Lemma 2.5.2, which will also be useful in Chapter 3, so we prove a rather general version.

Lemma 2.5.2. *Let X be an event for which $P(X) > 0$ and $X \perp\!\!\!\perp \sigma(\xi)$, where ξ is a random variable. Then for any event Y,*

$$0 \leq \mu_X(Y \mid \xi) \leq 1, \quad \textit{almost surely.} \quad (2.42)$$

Proof. We have

$$\mu_X(Y \mid \xi) = \frac{1}{P(X)}P(Y \cap X \mid \xi) = \frac{P(Y \cap X \mid \xi)}{P(X \mid \xi)}.$$

In the above, $P(X \mid \xi) = P(X)$ holds because $X \perp\!\!\!\perp \sigma(\xi)$; it is straightforward to see that the conditional probability can be represented as the constant function taking value $P(X)$.

We assert that

$$P(X \mid \xi) \geq P(Y \cap X \mid \xi), \quad \textit{almost surely.}$$

To show this, let A_ϵ be the set for which $P(Y \cap X \mid \xi) > P(X \mid \xi) + \epsilon$, where $\epsilon > 0$. Since $P(Y \cap X \mid \xi)$ and $P(X \mid \xi)$ are measurable with respect to $\sigma(\xi)$, we have $A_\epsilon \in \sigma(\xi)$. Hence

$$P(X \cap A_\epsilon) \geq P(Y \cap X \cap A_\epsilon) = \int_{A_\epsilon} P(Y \cap X \mid \xi)dP \geq \int_{A_\epsilon} P(X \mid \xi) + \epsilon dP = P(X \cap A_\epsilon) + P(A_\epsilon)\epsilon,$$

which is only possible if $P(A_\epsilon) = 0$. The event $\{P(Y \cap X \mid \xi) > P(X \mid \xi)\}$ is equal to

$$\bigcup_{n=1}^{\infty} A_{1/n},$$

which is an increasing union of sets, each of measure zero. So, we can say that $P(Y \cap X \mid \xi) > P(X \mid \xi)$ occurs with probability zero, by Lemma 3.4 of [8].

Since $P(Y \cap X \mid \xi) \leq P(X \mid \xi)$, almost surely, we have

$$\mu_X(Y \mid \xi) = \frac{P(X \cap Y \mid \xi)}{P(X \mid \xi)} \leq 1 \quad a.s.$$

This proves the upper bound. To show the lower bound, it will suffice to show that $P(Y \cap X \mid \xi)$ is greater than or equal to zero, almost surely. Let B_n be the set where $P(Y \cap X \mid \xi) = E(I_{Y \cap X} \mid \xi) < -\frac{1}{n}$, which will be measurable with respect to $\sigma(\xi)$. Then we can write

$$-\frac{1}{n}P(B_n) = \int_{B_n} -\frac{1}{n} dP \geq \int_{B_n} E(I_{Y \cap X} \mid \xi) dP = \int_{B_n} I_{Y \cap X} dP \geq \int_{B_n} 0 dP = 0,$$

which can only be true if $P(B_n) = 0$. Finally, $\{P(Y \cap X \mid \xi) < 0\} = \cup_{n=1}^{\infty} B_n$, which is an increasing union of measure-zero events, so $P(Y \cap X \mid \xi)$ is negative with probability zero. □

For our current purposes, we note that by (2.3), Lemma 2.5.2 applies to expressions such as $\mu_a(+_1 \mid \lambda)$. We can now prove the CHSH inequality.

Proposition 2.5.2. *(CHSH Inequality) Let a, b, D_1, D_2 be as in Definition 2.1.1. Then, under* (2.1), (2.2), (2.3), *and* (2.4),

$$|CHSH| \leq 2. \tag{2.43}$$

Proof. By Proposition 2.5.1, we have

$$\begin{aligned} CHSH &= \int_\Omega \mathbf{ab} dP - \int_\Omega \mathbf{ab'} dP + \int_\Omega \mathbf{a'b} dP + \int_\Omega \mathbf{a'b'} dP \\ &= \int_\Omega (\mathbf{ab}) - (\mathbf{ab'}) + (\mathbf{a'b}) + (\mathbf{a'b'}) dP = \int_\Omega (\mathbf{a} + \mathbf{a'})\mathbf{b} + (\mathbf{a'} - \mathbf{a})\mathbf{b'} dP. \end{aligned}$$

By Lemma 2.5.2, $\mathbf{a}$ and $\mathbf{a'}$ must lie in the interval $[-1, +1]$. By arithmetical considerations, it follows

that

$$|\mathbf{a}+\mathbf{a}'|+|\mathbf{a}'-\mathbf{a}|\leq 2. \tag{2.44}$$

A simple case analysis demonstrates why (2.44) holds.

Case 1: $\mathbf{a}+\mathbf{a}'\geq 0$

In this case, the left side of (2.44) is equal to $\mathbf{a}+\mathbf{a}'+|\mathbf{a}'-\mathbf{a}|$, and this is either equal to $\mathbf{a}+\mathbf{a}'+(\mathbf{a}'-\mathbf{a})=2\mathbf{a}'\leq 2$, or equal to $\mathbf{a}+\mathbf{a}'-(\mathbf{a}'-\mathbf{a})=2\mathbf{a}\leq 2$.

Case 2: $\mathbf{a}+\mathbf{a}'<0$

In this case, the left side of (2.44) is equal to $-\mathbf{a}-\mathbf{a}'+|\mathbf{a}'-\mathbf{a}|$, and this is either equal to $-\mathbf{a}-\mathbf{a}'+(\mathbf{a}'-\mathbf{a})=-2\mathbf{a}\leq 2$, or equal to $-\mathbf{a}-\mathbf{a}'-(\mathbf{a}'-\mathbf{a})=-2\mathbf{a}'\leq 2$.

Hence (2.44) holds. Now, since $|\mathbf{b}|,|\mathbf{b}'|\leq 1$, we have

$$\begin{aligned} |CHSH| &= \left|\int_\Omega (\mathbf{a}+\mathbf{a}')\mathbf{b}+(\mathbf{a}'-\mathbf{a})\mathbf{b}'dP\right| \leq \int_\Omega |\mathbf{a}+\mathbf{a}'||\mathbf{b}|+|\mathbf{a}'-\mathbf{a}||\mathbf{b}'|dP \\ &\leq \int_\Omega |\mathbf{a}+\mathbf{a}'|+|\mathbf{a}'-\mathbf{a}|dP \leq \int_\Omega 2dP = 2. \end{aligned}$$

□

As a consequence of Proposition 2.5.2, in any local theory, the quantity *CHSH* must satisfy the simple inequality (2.43). If we assume that Bell test experiments are repeatable and that the results of repeated trials are independent and identically distributed (i.i.d.), we can calculate the *CHSH* quantity empirically, by tallying the relative frequencies of various experimental outcomes, and appealing to the Law of Large Numbers to assert that the relative frequencies will converge to the true *CHSH* quantity. In the next section, we will see that quantum mechanics predicts that $CHSH=2\sqrt{2}>2$, and is thus incompatible with locality. In the section after that, we will show how to compare the two theories *without* making the i.i.d. assumption.

2.6 The Quantum Alternative

Quantum Mechanics has a precise model for a CHSH experiment. The state of the system, λ, is taken to be the singlet state (1.5) that was introduced in Section 1.2,

$$\lambda_s=\frac{|01\rangle-|10\rangle}{\sqrt{2}},$$

with probability one. There are four different projective measurements that can be used to measure λ_s, and the choice of measurement is governed by the random variables A and B. If, for instance, A and B are the outputs of random number generators, then we can visualize the experimenter hooking

up the random number generator to the apparatus that performs the projective measurement, such that the apparatus performs different projective measurements depending on the particular random output. The detector outputs D_1 and D_2 are the output of the projective measurement.

To describe the projective measurements that are used, recall that λ_s exists in the tensor product space $\mathbb{C}^2 \otimes \mathbb{C}^2$. This allows us to describe projection maps in $\mathbb{C}^2$ and then tensor them to get projection maps in $\mathbb{C}^2 \otimes \mathbb{C}^2 = \mathbb{C}^4$.

Let $\{|0\rangle, |1\rangle\}$ be an orthonormal basis of $\mathbb{C}^2$, recalling the scheme of Figure 1.3. Then for any fixed angle θ, we can describe a new rotated orthonormal basis as follows:

$$\begin{aligned} |+\rangle &= \cos\theta|0\rangle + \sin\theta|1\rangle \\ |-\rangle &= -\sin\theta|0\rangle + \cos\theta|1\rangle. \end{aligned}$$

Figure 1.3 depicts a projection measurement involving projection onto $|0\rangle$ and $|1\rangle$, but if the apparatus is rotated counter-clockwise by θ radians (rotating the $|0\rangle$ direction up towards the $|1\rangle$ direction), a new projection measurement can be performed onto $|+\rangle$ and $|-\rangle$. Denote the projection maps onto $|+\rangle$ and $|-\rangle$ by P_θ^+ and P_θ^-, respectively. Since $|+\rangle$ and $|-\rangle$ are orthonormal, it follows that for any fixed θ, $P_\theta^+ + P_\theta^- = I_2$, and thus the collection $\{P_\theta^+, P_\theta^-\}$ is a valid projective measurement on $\mathbb{C}^2$.

What is needed, however, is a projective measurement in $\mathbb{C}^2 \otimes \mathbb{C}^2$, as this is the space inhabited by λ_s. To construct one, first note that if $\{P_1, P_2\}$ and $\{P_3, P_4\}$ are two different projection measurements on $\mathbb{C}^2$, then the collection

$$P_1 \otimes P_3 \qquad P_1 \otimes P_4 \qquad P_2 \otimes P_3 \qquad P_2 \otimes P_4 \tag{2.45}$$

is a valid measurement on $\mathbb{C}^2 \otimes \mathbb{C}^2$: the algebraic properties of tensors allow us to demonstrate that the elements of 2.45 are idempotent (and hence projectors), and sum to I_4.

If a two-photon system is measured by performing the individual measurement $\{P_1, P_2\}$ on one of the photons and performing the individual measurement $\{P_3, P_4\}$ on the other photon, then the quantum description of the complete measurement process is given by the projective measurement represented by (2.45). If the two photons are not interacting with each other, the measurement statistics can calculated either individually or as the joint measurement (2.45). However, if the photon pair is in an entangled state like λ_s, which can occur even if the photons are separated by

large distances, the measurement statistics must be calculated with the expression (2.45), and the statistics can display correlations that can violate the CHSH inequality. To see a violation, consider the following assignment of angles:

$$\theta_a = \frac{3\pi}{4} \qquad \theta_{a'} = 0 \qquad \theta_b = \frac{9\pi}{8} \qquad \theta_{b'} = \frac{3\pi}{8} \tag{2.46}$$

Using this assignment, the experiment is set up so that if, for instance, $A = a'$ and $B = b$, then detector 1 is rotated 0 radians, detector 2 is rotated $\frac{9\pi}{8}$ radians, and the projective measurement performed is given by

$$P^+_{\theta_a} \otimes P^+_{\theta_{b'}} \qquad P^+_{\theta_a} \otimes P^-_{\theta_{b'}} \qquad P^-_{\theta_a} \otimes P^+_{\theta_{b'}} \qquad P^-_{\theta_a} \otimes P^-_{\theta_{b'}}.$$

Depending on which outcome occurs above, the results for (D_1, D_2) will be $(+1, +1)$, $(+1, -1)$, $(-1, +1)$, or $(-1, -1)$, respectively. The experimenter will observe one of these four results, and the probability of each result is given by Expression (1.4). Thus it can be calculated that if the state of the photons is λ_s,

$$E_{a'b}(D_1 D_2) = -\frac{1}{\sqrt{2}}.$$

Similarly, we can calculate that

$$E_{ab}(D_1 D_2) = E_{ab'}(D_1 D_2) = E_{a'b'}(D_1 D_2) = \frac{1}{\sqrt{2}},$$

which implies that

$$\begin{aligned} CHSH_{\text{quantum}} &= \frac{1}{\sqrt{2}} - \left(-\frac{1}{\sqrt{2}}\right) + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} \\ &= 2\sqrt{2} \\ &\not\leq 2 \quad (!) \end{aligned}$$

This violates the inequality (2.43). Inequality (2.43) does not apply to the $CHSH$ quantity under a quantum description because Quantum Mechanics does not obey (2.4), the locality

assumption, which is used to derive (2.43). Namely, if $\lambda = \lambda_s$ with probability one,

$$P_{\text{quantum}}(+_1 +_2 ab \mid \lambda) = P_{\text{quantum}}(+_1 +_2 ab)$$
$$\neq P_{\text{quantum}}(+_1 a)P_{\text{quantum}}(+_2 b) = P_{\text{quantum}}(+_1 a \mid \lambda)P_{\text{quantum}}(+_2 b \mid \lambda),$$

where equality would need to hold if (2.4) were to be satisfied. This is counterintuitive, but it is a consequence of the fact that certain quantum systems (such as the singlet state) cannot be described by two separate states, even if the two parts of the quantum state are separated by great distances.

Quantum Mechanics allows independent, identical copies of the singlet state to be reproduced for repeated trials of the same experiment, so if we assume Quantum Mechanics is true, it is appropriate to compute sample averages that approximate $E_{xy}(D_1 D_2)$ terms and appeal to the law of large numbers to assert that the sample averages will converge to the true parameters. However, potential local theories need not have such nice properties, so in the hypothesis test described in the next section, such ideas cannot be used.

2.7 The Hypothesis Test

If we run the Bell experiment one time, we will randomly select one particular setting result for A and B, and we will observe $D_1 D_2$ equal to $+1$ or -1. This one result tells us nothing about the satisfaction or violation of (2.5.2). So, we must run the experiment many times to discern a pattern.

In the usual framework for a hypothesis test, one assumes that the repeated trials are independent and identically distributed. This may seem like a reasonable assumption to make, as the experimenter runs repeated trials in exactly the same fashion. However, we are trying to contrast quantum mechanics with a whole *class* of local theories – and a local theory could easily include a scenario in which early trial results affect the outcomes of later trials. In fact, it is not hard to imagine such a local scenario: at one of the detectors, each time a photon is measured, it generates a result in $\{+1, -1\}$, but also modifies the detector in some microscopic way such that the measurement outcome of the next photon at that detector is correlated with the measurement of the previous photon. Hence photon measurements at the same location but at different times are correlated - so repeated trials would not be i.i.d., but the theory could still be local, as the measurements for any specified pair of spatially separated photons could still satisfy (2.4).

This issue has been referred to as the *memory loophole*. Two papers [14, 15] dealing with

this issue in the early 2000s concluded that a cogent hypothesis test can still be performed. However, [14] assumes that the variable λ is absolutely continuous and uses some informal arguments. The paper [15] reaches its conclusions with more mathematical rigor, but also falls short of full generality in modeling the random variable λ, and requires the introduction of unnecessary probabilistic machinery – a random process known as the *martingale* – to make the argument. (Incidentally, we will see that martingale techniques are unavoidable in certain related scenarios explored in Chapter 3, but the use of the martingale technique for the current CHSH experiment is not necessary and in fact would lead to a suboptimal calculation of statistical strength.) This section presents a completely general measure-theoretic proof that a cogent hypothesis test can still be performed, even without the assumption of i.i.d. trials, along with a demonstration of how to exactly calculate the significance level and power of the test.

Here is a useful analogy that will illustrate how we will proceed. Suppose we were to flip 10,000 different coins, and 80% of them were to come up "heads." Then we could reasonably conclude that at least *some* of the 10,000 coins were biased towards heads. The coins needn't be identically distributed - indeed, perhaps some of the coins were fair - but it is intuitively clear that *some* of them must have been biased.

Analogously, each trial of the Bell test is like a coin flip, resulting in $+1$ and -1. In the previous sections, we showed that the assumption of locality constrains the probabilities of getting $+1$ or -1. If a theory is local, then the constraint must be satisfied on *every* trial. On the other hand, if Quantum Mechanics is obeyed, the constraint is *violated* on every trial. The locality assumption is like the assumption that all of the 10,000 coins are fair, and agreement with Quantum Mechanics will roughly correspond to getting 80% heads. The CHSH inequality is of course a little more complicated than this, but this is a good idea to keep in mind as we design the hypothesis test.

To represent repeated trials, we will have to make a slight update to our mathematical model. let us define a sequence of random vectors:

$$\{D_{1i}, D_{2i}, A_i, B_i, \lambda_i\}_{i \in \mathbf{N}^+} \tag{2.47}$$

For each i, we take the above to be as defined in Definition 2.1.1, satisfying conditions (2.1) and (2.3), and a strengthened version of (2.2). That is, we assume:

Experimental Assumption 1:

$$\forall i, \quad A_i \perp\!\!\!\perp B_i. \tag{2.48}$$

Experimental Assumption 2*:

$$\forall i, \quad P(A_i = a) = P(B_i = b) = \frac{1}{2}. \tag{2.49}$$

Remark 2.7.1. Whereas before, we only assumed that the probabilities were nonzero, we now assume that all the measurement probabilities are $\frac{1}{2}$. Condition (2.49) can of course be satisfied by appropriate calibration of the experimental apparatus. This makes the task of proving an analog of the CHSH inequality that holds over repeated trials significantly simpler.

Experimental Assumption 3:

$$A_i \perp\!\!\!\perp \lambda_i, \quad B_i \perp\!\!\!\perp \lambda_i. \tag{2.50}$$

An additional point about the λ_i needs to be made. Since λ_i models the state of the system at the ith trial, the previous $i-1$ trials have already taken place. Hence, the outcomes of previous trials are in the "history", and contribute to the present state of the system, so λ_i can contain information about them. For instance, we theorized how this could occur if the outcome of one trial left a residue in the detector to affect the state of the next incoming photon. Mathematically, this is modeled by assuming that the results of previous trials are events in $\sigma(\lambda_i)$. This is related to the notion of a *filtration* - i.e., a sequence nested σ-algebras:

$$\text{For } i < j, \quad \sigma(\lambda_i) \subseteq \sigma(\lambda_j).$$

The filtration is a standard mathematical tool for modeling a time-indexed stochastic process. The above equation is not directly used in our argument; rather, we will directly use the fact that the outcomes of previous trials are events in $\sigma(\lambda_i)$. The following assumption formalizes this.

Time Sequentiality: For any positive integer $n \geq 2$, let I be a subset of $\{1, 2, ..., n-1\}$ whose cardinality we denote with the letter m. Let $\vec{v}_1$, $\vec{v}_2$ be elements of $\{-1, +1\}^m$, let $\vec{w}_a$ be an element of $\{a, a'\}^m$, and let $\vec{w}_b$ be an element of $\{b, b'\}^m$. Then the following event is in $\sigma(\lambda_n)$:

$$\bigcap_{i \in I} \left[\{D_{1i} = \vec{v}_{1i}\} \cap \{D_{2i} = \vec{v}_{2i}\} \cap \{A_i = \vec{w}_{ai}\} \cap \{B_i = \vec{w}_{bi}\} \right]. \tag{2.51}$$

The astute reader will notice that this is mathematically equivalent to saying that all the single events such as $\{D_{1i} = \vec{v}_{1i}\}$ are individually in $\sigma(\lambda_n)$; (2.51) is written the way it is to emphasize simultaneousness of the i-indexed events. The significance of asserting (2.51) is in $\sigma(\lambda_n)$ is to encode the notion that λ_n can potentially depend on the outcomes of previous trials. However, this should not be mis-interpreted as saying that λ_n definitely *does* have some relation to the outcome of previous trials; it could be completely independent of this information. What we are doing here is only allow for this possibility.

Moving forward, we will continue to have a locality assumption corresponding to (2.4). Restated to reflect the sequential nature of the trials, we end up with the following formulation:

LOCALITY ASSUMPTION: Let $V_{i1} = (D_{i1}, A_i)$ and $V_{i2} = (D_{i2}, B_i)$. Then

$$(V_{i1} \perp\!\!\!\perp V_{i2}) \mid \lambda_i. \tag{2.52}$$

This completes the set of assumptions. Now, to formulate the hypothesis test, it will be convenient to define a random variable C_i as a function of the random variables in (2.47). So, let

$$C_i = \begin{cases} D_{1i}D_{2i}, & \text{if } (A_i, B_i) \neq (a', b), \\ -D_{1i}D_{2i}, & \text{if } (A_i, B_i) = (a', b). \end{cases}$$

C_i is useful because it distills the result of the ith trial into a single, two-output random variable. As we will see in the upcoming proposition, the CHSH inequality applies to C_i to cap the probability that $C_i = +1$ at 75%, if we make all of the experimental assumptions plus locality. This constraint on C_i does require that the setting probabilities are calibrated to $\frac{1}{2}$, as is ensured by Experimental Assumption 2*. It should be noted however that this assumption is not vital to the current analysis – had we not made this specific calibration of the probabilities, we could still proceed by inserting constants into the definition of C_i to weight the different outcomes according to how often certain setting outcomes occur. In practical experiments, the setting probabilities usually *are* calibrated to 1/2, and we will avoid some extra clutter in the definition of C_i by assuming equiprobable settings.

Proposition 2.7.1. *Under assumptions* (2.48), (2.49), (2.50), *and the locality assumption* (2.52), *we have* $P(C_i = +1) \leq \frac{3}{4}$, *or equivalently,*

$$E(C_i) \leq 1/2. \tag{2.53}$$

Proof.

$$E(C_i) = E\big[E(C_i \mid (A_i, B_i))\big],$$

where $E(C_i \mid (A_i, B_i))$ will be a random variable with four outputs, corresponding to the four outputs of (A_i, B_i). Applying (2.49) and (2.48), we have

$$\begin{aligned} &E\big[E(C_i \mid (A_i, B_i))\big] \\ &= \tfrac{1}{4}\left[\begin{smallmatrix} E[C_i|(A_i,B_i)=(a,b)]+E[C_i|(A_i,B_i)=(a',b)] \\ +E[C_i|(A_i,B_i)=(a,b')]+E[C_i|(A_i,B_i)=(a',b')] \end{smallmatrix}\right] \\ &= \tfrac{1}{4}\left[E_{ab}(D_{1i}D_{2i}) - E_{a'b}(D_{1i}D_{2i}) + E_{ab'}(D_{1i}D_{2i}) + E_{a'b'}(D_{1i}D_{2i})\right]. \end{aligned}$$

Noting the similarity to (2.41), we obtain the following,

$$E(C_i) = \frac{1}{4} CHSH_i \tag{2.54}$$

where we take $CHSH_i$ to be as defined in (2.41), after substituting in the indexed variables from (2.47). Assumptions (2.48), (2.49), (2.50), and (2.52) are equivalent to the assumptions of Proposition 2.5.2 if applied to $CHSH_i$, so the proposition holds, giving us (2.53). □

For each i, C_i is a Bernoulli trial, taking outputs in the set $\{+1, -1\}$, so let us define

$$p_i := P(C_i = +1).$$

It is straightforward to compute

$$E(C_i) = 2p_i - 1, \quad \mathrm{Var}(C_i) = 4p_i(1 - p_i). \tag{2.55}$$

Under locality, (2.53) and (2.55) imply that p_i must be at most 75%. On the other hand, Quantum Mechanics predicts that $E(C_i) = \frac{\sqrt{2}}{2}$, which yields p_i of roughly 85.4%. This will allow us to discern a difference over many trials.

We can now formulate the hypothesis test in mathematical terms:

$$\begin{aligned} H_0 &: \forall i, p_i \le \frac{3}{4} && \text{(Locality)} \\ H_A &: \forall i, p_i = \frac{1+\sqrt{2}}{2\sqrt{2}} = .854... && \text{(Quantum)} \end{aligned}$$

Over n trials, the natural choice for a sample statistic is $\overline{C_n}$, defined as follows:

$$\overline{C_n} = \frac{\sum_{i=1}^{n} C_i}{n}.$$

and so under the assumption of H_0, we expect the sample statistic $\overline{C_n}$ to satisfy

$$E(\overline{C_n}) \leq 1/2. \tag{2.56}$$

We will reject the null hypothesis in favor of the alternative hypothesis if $\overline{C_n} > z$, where z will be some cut-point exceeding 1/2.

Let $p_n(\_)$ denote a probability mass function for the first n outputs of C_i, and let Θ_0 be the collection of $p_n(\_)$ that satisfy the assumptions (2.48)-(2.52), so Θ_0 denotes the collection of allowable distributions under the null hypothesis. Then the significance level of the hypothesis test – the probability of Type I error – is defined to be

$$\alpha = \sup_{p_n(\_) \in \Theta_0} P[\overline{C_n} > z \mid p_n(\_)]. \tag{2.57}$$

Calculating α is somewhat involved, because the null hypothesis does not assert that the various C_i are independent and/or independently distributed, and equation (2.56) alone does not provide us with an asymptotic distribution of $\overline{C_n}$. In the absence of the assumption of i.i.d. trials, we cannot *a priori* rule out trivialities such as

$$C_1 = C_2 = \cdots = C_{n-1} = C_n \tag{2.58}$$

(total dependence), for which we would have $\alpha = p_i$, independent of n!

Luckily, we can derive a certain degree of independence between the C_i for a system satisfying the conditions assumed by H_0: (2.48), (2.49), (2.50), (2.51), and the locality assumption (2.52). The following lemma rules out possibilities like (2.58), and it will allow us to demonstrate that α decreases as n increases.

Lemma 2.7.1. *Let $\vec{v}$ be any vector in $\{-1, +1\}^{i-1}$ for which $P(C_1, ..., C_{i-1} = \vec{v}) > 0$. Then, under the null hypothesis – which subsumes assumptions* (2.48)-(2.52) *– we have*

$$P\big(C_i = +1 \mid (C_1, ..., C_{i-1}) = \vec{v}\big) \leq \frac{3}{4}. \tag{2.59}$$

Proof. Let $\mathcal{C}$ denote the event $(C_1, ..., C_{i-1}) = \vec{v}$. Let C_i be a shorthand for the event $C_i = +1$. Then we have

$$\begin{aligned} P(C_i \mid \mathcal{C}) &= \frac{P(C_i \cap \mathcal{C})}{P(\mathcal{C})} \\ &= \frac{1}{P(\mathcal{C})} E(I_{C_i \cap \mathcal{C}}) \\ &= \frac{1}{P(\mathcal{C})} E\big[E(I_{C_i \cap \mathcal{C}} \mid \lambda_i)\big] \\ &= \frac{1}{P(\mathcal{C})} \int E(I_{C_i \cap \mathcal{C}} \mid \lambda_i) dP \\ &= \frac{1}{P(\mathcal{C})} \int P(C_i \cap \mathcal{C} \mid \lambda_i) dP. \end{aligned}$$

In the last integral above, note that $\mathcal{C}$ is in $\sigma(\lambda_i)$ by the time-sequential nature of the experiment, encapsulated in equation (2.51). This implies that

$$\int P(C_i \cap \mathcal{C} \mid \lambda_i) dP = \int P(C_i \mid \lambda_i) I_{\mathcal{C}} dP,$$

which is a consequence of Theorem 9.1.3 in [10] (which can also be found in the section "Properties of Conditional Expectation" in [9]). So the integral becomes

$$\int_{\mathcal{C}} P(C_i \mid \lambda_i) dP.$$

Using an i-indexed version of the "$+_1$" notation introduced following Definition 2.1.1, we apply Lemma 1.3.1 to decompose the integrand into the eight constituent sub-events of C_i, obtaining

$$\begin{aligned} \int_{\mathcal{C}} & P(+_{1i} \cap +_{2i} \cap a_i \cap b_i \mid \lambda_i) + P(-_{1i} \cap -_{2i} \cap a_i \cap b_i \mid \lambda_i) \\ & + P(+_{1i} \cap +_{2i} \cap a_i \cap b_i' \mid \lambda_i) + P(-_{1i} \cap -_{2i} \cap a_i \cap b_i' \mid \lambda_i) \\ & + P(+_{1i} \cap -_{2i} \cap a_i' \cap b_i \mid \lambda_i) + P(-_{1i} \cap +_{2i} \cap a_i' \cap b_i \mid \lambda_i) \\ & + P(+_{1i} \cap +_{2i} \cap a_i' \cap b_i' \mid \lambda_i) + P(-_{1i} \cap -_{2i} \cap a_i' \cap b_i' \mid \lambda_i) \quad dP. \end{aligned} \tag{2.60}$$

We apply (2.52) to the first term of (2.60) to get

$$P(+_{1i} \cap +_{2i} \cap a_i \cap b_i \mid \lambda_i) = P(+_{1i} \cap a_i \mid \lambda_i) P(+_{2i} \cap b_i \mid \lambda_i),$$

and multiplying right-hand side above by $P(a_i \cap b_i)/P(a_i \cap b_i)$ yields, via (2.48) and (2.49),

$$P(+_{1i} \cap +_{2i} \cap a_i \cap b_i \mid \lambda_i) = P(a_i \cap b_i)\big[\mu_{a_i}(+_{1i} \mid \lambda_i)\mu_{b_i}(+_{2i} \mid \lambda_i)\big] = \frac{1}{4}\big[\mu_{a_i}(+_{1i} \mid \lambda_i)\mu_{b_i}(+_{2i} \mid \lambda_i)\big].$$

The other seven terms simplify the same way, so (2.60) becomes

$$\begin{aligned}\frac{1}{4}\int_{\mathcal{C}} &\mu_{a_i}(+_{1i} \mid \lambda_i)\mu_{b_i}(+_{2i} \mid \lambda_i) + \mu_{a_i}(-_{1i} \mid \lambda_i)\mu_{b_i}(-_{2i} \mid \lambda_i) \\ &+\mu_{a_i}(+_{1i} \mid \lambda_i)\mu_{b_i'}(+_{2i} \mid \lambda_i) + \mu_{a_i}(-_{1i} \mid \lambda_i)\mu_{b_i'}(-_{2i} \mid \lambda_i) \\ &+\mu_{a_i'}(+_{1i} \mid \lambda_i)\mu_{b_i}(-_{2i} \mid \lambda_i) + \mu_{a_i'}(-_{1i} \mid \lambda_i)\mu_{b_i}(+_{2i} \mid \lambda_i) \\ &+\mu_{a_i'}(+_{1i} \mid \lambda_i)\mu_{b_i'}(+_{2i} \mid \lambda_i) + \mu_{a_i'}(-_{1i} \mid \lambda_i)\mu_{b_i'}(-_{2i} \mid \lambda_i)dP. \end{aligned} \tag{2.61}$$

If we define

$$t = \mu_{a_i}(+_{1i} \mid \lambda_i) \quad s = \mu_{a_i'}(+_{1i} \mid \lambda_i) \quad u = \mu_{b_i}(+_{2i} \mid \lambda_i) \quad v = \mu_{b_i'}(+_{2i} \mid \lambda_i),$$

we can factor the integrand in (2.61) and again apply Lemma 1.3.1 to obtain

$$\frac{1}{4}\int_{\mathcal{C}} t[u+v] + (1-t)[(1-u)+(1-v)] + s[v+(1-u)] + (1-s)[u+(1-v)]dP. \tag{2.62}$$

By Lemma 2.5.2, which applies by (2.50), we have s, t, u, and v in $[0,1]$. Using this constraint, the integrand in (2.60) is always bounded by 3. To see why this holds, notice that for fixed values of u and v, the integrand will be maximal when t is either 0 or 1, and similarly for s. Hence if we analyze all four cases for s and t, the value of the integrand is given as follows:

$$\begin{aligned}(s,t) = (0,0) \quad &\Rightarrow \quad \text{integrand} = 3 - 2v \leq 3 \\ (s,t) = (0,1) \quad &\Rightarrow \quad \text{integrand} = 1 + 2u \leq 3 \\ (s,t) = (1,0) \quad &\Rightarrow \quad \text{integrand} = 3 - 2u \leq 3 \\ (s,t) = (1,1) \quad &\Rightarrow \quad \text{integrand} = 1 + 2v \leq 3.\end{aligned}$$

Now, returning to the original expression, we have

$$P(C_i \mid \mathcal{C}) \leq \frac{1}{P(\mathcal{C})} \cdot \frac{1}{4}\int_{\mathcal{C}} 3dP = \frac{3}{4}. \qquad \square$$

Lemma 2.7.1 allows us to formulate a limiting distribution for $\overline{C_n}$, as shown in the following proposition. Note that the event "$\overline{C_n} > z$" is equivalent to the event "at least k of the C_i equal $+1$", where k is an integer determined by the particular value of z. We prove a slightly more general result than what is needed here, as this general form will also be useful in Chapter 3. A similar result was tacitly assumed without proof in [14], but as we are about to see, the proof does require a little bit of work.

Proposition 2.7.2. *Let $\{C_i\}_{i=1}^{\infty}$ be a sequence of random variables taking values in the set $\{-1, +1\}$. Suppose that there exists a number $p \in (0, 1)$ such that for any choice of i and any vector $\vec{v} \in \{-1, +1\}^{i-1}$ for which $P((C_1, ..., C_{i-1}) = \vec{v}) > 0$, the following relation holds:*

$$P(C_i = +1 \mid (C_1, ..., C_{i-1}) = \vec{v}) \leq p. \tag{2.63}$$

Then, if we fix a positive integer n and let B_n be the Binomial random variable corresponding to n trials with probability of success p, the following holds:

$$P(\text{at least } k \text{ of the } C_i \text{ equal } +1) \leq P(B_n \geq k), \tag{2.64}$$

where k is any nonnegative integer less than or equal to n.

Proof. To show this holds for any fixed positive integer n, we use mathematical induction.

Case 1: $n = 1$.

There are two possibilities for k: 0 and 1. For $k = 1$, $P(\text{at least } k \text{ of the } C_i \text{ equal } +1) = P(C_1 = +1) \leq p = P(B_1 \geq 1)$, and for $k = 0$, $P(\text{at least } k \text{ of the } C_i \text{ equal } +1) = 1 = P(B_1 \geq 0)$.

Case 2: Assume the claim is true for n, and derive that it is true for $n + 1$.

Now, k can range from 0 to $n + 1$. First, let us prove it for k between 1 and n, and later we will prove the boundary cases of $k = 0$ and $k = n + 1$.

Introduce a shorthand,

$$\begin{aligned} P_{n,k}(C) &:= P(\text{for } 1 \leq i \leq n, \text{ at least } k \text{ of the } C_i \text{ equal } +1), \\ P_{n,k}(B) &:= P(B_n \geq k), \end{aligned}$$

so what we are trying to prove can now be written as $P_{n+1,k}(C) \leq P_{n+1,k}(B)$. By conditioning,

$$P_{n+1,k}(C) = P_{n,k}(C) + p_{C_{n+1}} \cdot \big[P_{n,k-1}(C) - P_{n,k}(C)\big], \tag{2.65}$$

where we note that $\big[P_{n,k-1}(C) - P_{n,k}(C)\big]$ is the probability that we have *exactly* $k-1$ successes after n trials, and $p_{C_{n+1}}$ denotes the probability that $C_{n+1} = +1$, given exactly $k-1$ successes after n trials. Note that as we are temporarily omitting the possibility that $k = n+1$ or $k = 0$, it follows that $P_{n,k}(C)$ and $P_{n,k-1}(C)$ are well-defined and included in the scope of the inductive hypothesis.

Let S be the subset of $\{-1,+1\}^n$ where exactly $k-1$ of the entries are $+1$ and $\vec{v} \in S \Rightarrow P\big[(C_1,...,C_n) = \vec{v}\big] > 0$. We have

$$\begin{aligned}
p_{C_{n+1}}\big[P_{n,k-1}(C) - P_{n,k}(C)\big] &= \sum_{\vec{v}\in S} P\big[C_{n+1} = +1 \mid (C_1,...,C_n) = \vec{v}\big] P\big[(C_1,...,C_n) = \vec{v}\big] \\
&\leq \sum_{\vec{v}\in S} p \cdot P\big[(C_1,...,C_n) = \vec{v}\big] \\
&= p \sum_{\vec{v}\in S} P\big[(C_1,...,C_n) = \vec{v}\big] \\
&= p\big[P_{n,k-1}(C) - P_{n,k}(C)\big],
\end{aligned}$$

where the inequality above follows from the assumption (2.63). From this it follows that

$$p_{C_{n+1}} \leq p. \tag{2.66}$$

Returning to (2.65), we have

$$P_{n+1,k}(C) = (1 - p_{C_{n+1}})P_{n,k}(C) + p_{C_{n+1}} P_{n,k-1}(C)$$

and by (2.66), and the fact that $P_{n,k-1}(C) \geq P_{n,k}(C)$, we have

$$\leq (1-p)P_{n,k}(C) + p \cdot P_{n,k-1}(C).$$

By the inductive hypothesis, $P_{n,k}(C) \leq P_{n,k}(B)$ and $P_{n,k-1}(C) \leq P_{n,k-1}(B)$, so we have

$$\leq (1-p)P_{n,k}(B) + p \cdot P_{n,k-1}(B) = P_{n,k}(B) + p\,[P_{n,k-1}(B) - P_{n,k}(B)] = P_{n+1,k}(B).$$

Hence, $P_{n+1,k}(C) \leq P_{n+1,k}(B)$.

This leaves only the boundary cases unproven. For $k = 0$, we clearly have

$$P_{n+1,k}(C) = 1 = P_{n+1,k}(B),$$

so the inequality holds easily. For $k = n + 1$, we have

$$P_{n+1,k}(C) = P_{n,k-1}(C) \cdot P(C_{n+1} = +1 \mid C_i = +1 \text{ for } i = 1, ..., n).$$

As $P_{n,k-1}(C) \leq P_{n,k-1}(B)$ by the inductive hypothesis, and $P(C_{n+1} = +1 \mid C_i = +1 \text{ for } i = 1, ..., n)$ is less than or equal to p by assumption (2.63), we have

$$P_{n+1,k}(C) \leq p \cdot P_{n,k-1}(B) = P_{n+1,k}(B).$$

□

Proposition 2.7.2 tells us that under the null hypothesis, the probability of getting at least k "$C_i = +1$" results over the course of n trials is bounded above by the probability of getting at least k "successes" over the course of n Bernoulli trials with probability of success $\frac{3}{4}$. The bound is sharp: the i.i.d. case with $p_i = \frac{3}{4}$ is allowed (just not *implied*) by assumptions (2.48)-(2.52).

With these results, we have

$$\alpha = \sup_{p_n(\_) \in \Theta_0} P\big(\overline{C_n} > z \mid p_n(\_)\big) = P\left(\overline{C_n} > z \;\middle|\; C_i \text{ i.i.d. and } \forall i, p_i = \frac{3}{4}\right). \tag{2.67}$$

As α is bounded by the i.i.d case, which is achieved at the boundary of the null parameter space, we can calculate the p-value exactly by analyzing the relevant binomial distribution. We can use an asymptotic normal distribution to accurately estimate α for n sufficiently large.

Corollary 2.7.1. *For n sufficiently large,*

$$\alpha \cong 1 - \Phi\left(\frac{2z - 1}{\sqrt{3}/\sqrt{n}}\right), \tag{2.68}$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution.

Proof. From (2.67), α can be calculated by assuming the C_i are i.i.d., equaling +1 with probability

$\frac{3}{4}$, and (2.55) tells us that $\mu = \frac{1}{2}$ and $\sigma^2 = \frac{3}{4}$. So we can write

$$\alpha = P(\overline{C_n} > z) = P\left(\frac{\overline{C_n} - \frac{1}{2}}{\sqrt{\frac{3}{4}}/\sqrt{n}} > \frac{z - \frac{1}{2}}{\sqrt{\frac{3}{4}}/\sqrt{n}}\right). \tag{2.69}$$

By the assumption of i.i.d., the Central Limit Theorem tells us that for large n,

$$\frac{\overline{C_n} - \frac{1}{2}}{\sqrt{\frac{3}{4}}/\sqrt{n}} \sim N(0,1) \tag{2.70}$$

where $N(0,1)$ denotes the standard normal random variable. Then (2.68) follows. □

According to a common rule of thumb about Binomial random variables, the approximation should be sufficiently good for $n \geq 20$. Corollary 2.7.1 thus gives us a useful tool to quickly calculate the significance level of the hypothesis test.

The alternative hypothesis, H_A, specifies the distribution of the C_i exactly. The probabilities p_i are all equal to $\sim .854$, and the quantum mechanical description of the experiment asserts that successive trials are independent (as is intuitive). Hence it is possible to calculate the power of the test. The power is $1 - \beta$, where β is defined as

$$\beta = P(\overline{C_n} \leq z \mid H_A). \tag{2.71}$$

From (2.55), we calculate $E(C_i) = \frac{\sqrt{2}}{2}$ and $\mathrm{Var}(C_i) = \frac{1}{2}$. Then the Central Limit Theorem tells us that $\overline{C_n}$ is approximately normally distributed with mean $\frac{\sqrt{2}}{2}$ and variance $\frac{1}{2n}$, so

$$\beta \cong \Phi\left(\frac{2z - \sqrt{2}}{\sqrt{2}/\sqrt{n}}\right).$$

To obtain statistical significance, the needed number of trials is not exceedingly high. If the quantum prediction is correct, then $\overline{C_n}$ should tend to $\frac{\sqrt{2}}{2}$. Hence, if after n trials, C_n is about $\frac{\sqrt{2}}{2}$, we can calculate a p-value, using $z = \frac{\sqrt{2}}{2}$ in (2.68):

$$p\text{-value } = 1 - \Phi\left(\frac{\sqrt{2} - 1}{\sqrt{3}/\sqrt{n}}\right).$$

Then to get a p-value of $\alpha < .05$, it will suffice to have about $n = 50$ trials. This is about how many trials the experimenter will need to run to begin to see statistically significant results, if the

predictions of quantum mechanics are correct.

Remark 2.7.2. To calculate the power of the test, we used our knowledge of the quantum predictions. Of course, H_A could be extended to include *any* violation of locality – from a hypothesis test standpoint, our knowledge of the precise quantum predictions is not necessary. However, this knowledge did allow us to calculate the power of the test, and therefore estimate how many experimental trials one would need to witness statistical significance.

2.8 Conclusion

Experiments such as [16, 17, 18] have attempted to test the CHSH inequality. The results give some evidence that the quantum mechanical predictions are correct, but plausible local explanations have not been ruled out. The reason for this state of affairs is that the experiments have not precisely met all of the experimental parameters. For instance, the photonic experiment [16] was subject to many null measurements due to low detector efficiency. Hence the real outcome space of D_1 and D_2 would have to include a third possibility, such as $\{-1, +1, 0\}$, where 0 represents no outcome. To use the binary-outcome analysis of this chapter, one would have to make the (untestable) assumption that the signals that are not dropped constitute an unbiased sample of all signals. To properly analyze the data without such assumptions, one needs a new analysis that incorporates the third outcome, and even then, detector efficiency must exceed a certain threshold (that was not met in [16]) in order to rule out local theories. This type of analysis will be discussed in Chapter 3.

The experiments [17, 18] did not suffer from dropped signals, but they failed to meet another experimental parameter by not enforcing space-like separation between the detectors. Hence when a violation of the CHSH inequality was observed, this meant that (2.52) may have not applied, but the *interpretation* of this is not very interesting. Namely, the two systems exhibited correlations with each other beyond what could be accounted for in their shared history, but this could have been achieved in a non-exotic way, such as the exchange of sub-luminal signals of some sort. This would not violate our intuitive notions of locality. Hence the question of whether nature is local remains open.

We should also consider a little more carefully what it would mean if an experiment that met all of the necessary parameters generated data indicating that H_0 should be rejected. Since the CHSH inequality was derived as a consequence of the four assumptions (2.48)-(2.52), only one of these four would be required not to hold in reality. Generally, it is assumed that this would be the

locality assumption, (2.52), that is violated. Indeed, this is what Quantum Mechanics suggests, and Quantum Mechanics is a successful theory, upheld by countless experiments in the last 100 years.

However, the formulation of H_0, and the derivation of the CHSH inequality (2.43) also rest on three other assumptions; the "experimental assumptions," (2.48), (2.49), and (2.50). A physical theory could violate H_0, but still satisfy (2.52) so long as one of the experimental assumptions turned out to not hold.

What might this entail? The three experimental assumptions all concern the choice of detector setting, which is supposed to be random. All three assumptions are satisfiable, so long as we assume that the experimenter is capable of generating true randomness, uncorrelated from other processes. If this were not actually possible, how would we interpret this?

One way this could occur is if the system of photons sends out some mysterious "pre-signal" that somehow affects the random process governing the choice of detector setting; that way, the photons control the detector setting, (2.50) is violated, and the experiment can violate the CHSH inequality without violating (2.52). Notice that this is not experimentally testable; there is no way to definitively prove that something like this undetected "pre-signal" is *not* occurring.

There is another related possibility. All matter had the opportunity to interact locally in the distant past: according to the Big Bang theory, all the particles in the universe were in local contact in the beginning. So, a particle could have taken note of where all the other matter was, and where it was going, and if the universe then unfolded in a deterministic way, the particle could know that, 13 billion years in the future, it would be subjected to a Bell test experiment, and the particle would know in advance the detector setting. In this scenario, the particle could modify its behavior based on the detector setting, and the state of the particle λ_i would then not be independent of the detector setting A_i, so (2.50) would be violated. There is no way to experimentally rule out this possibility.

It appears that the only way to avoid some sort of non-local correlation would be to invoke a highly implausible scenario such as the ones described in the previous two paragraphs. However, as the wait for a definitive Bell test experiment continues, the possibility of a less exotic local theory of nature is not yet ruled out.

Chapter 3

Closing the Detection Loophole

3.1 Introduction

In this chapter, we explore how to deal with the dropped-signal issue that was raised in the last section of Chapter 2, as well as other issues that can arise in a real-world experiment. We will continue to work in a fully general measure-theoretic setting.

The dropped-signal issue is sometimes referred to as the *detection loophole*, and it is frequently present in Bell inequality experiments. This loophole presents itself through the real-world constraint in which particle detectors are not 100% efficient. Photons, which are otherwise ideal candidates for a Bell experiment, are notoriously hard to detect. When we allow for null results, a spin or polarization measurement really has *three* possible outcomes: $+1$ (up), -1 (down), and "undetected." A local hidden variable model can exploit this in such a way that the subset of particles actually detected obey quantum statistics, even though the collection of *all* the particles is governed by a local hidden variable theory – for an example of this, see [19]. Hence, it is best to derive a different Bell inequality that directly incorporates the "undetected" outcome.

In 1974, Clauser and Horne derived such an inequality, now known as the "CH74 inequality" [13]. In the CH74 inequality, "spin down" and "undetected" are collapsed into a single outcome, which in this paper we will refer to as "0". The other outcome, "spin up", is not modified. Then under the assumption of locality, the following constraint can be obtained:

$$P(+_1,+_2 \mid a,b)+P(+_1,+_2 \mid a,b')+P(+_1,+_2 \mid a',b)-P(+_1,+_2 \mid a',b')-P(+_1 \mid a)-P(+_2 \mid b)\leq 0. \tag{3.1}$$

Following the notational conventions established in the last chapter, $+_i$ means a spin-up detection in the ith detector (there are still two detectors), a and a' are the settings for the first detector, and b and b' are the settings for the second detector. If the state of the photon pair is the singlet state, the bound is predicted to be violated by quantum mechanics when detection efficiency exceeds $\frac{2}{1+\sqrt{2}}$, which is roughly 83%. Interestingly, it was demonstrated in [20] that other quantum states can violate (3.1) with lower detection efficiencies, asymptotically approaching $\frac{2}{3}$ from the positive direction. The resulting conjecture – that $\frac{2}{3}$ is the lower bound for needed detector efficiency for a quantum violation of (3.1) – was proved in [21]. Combined with earlier work [22] that shows that any system satisfying every permutation of (3.1) can be modeled by a hidden variable, this indicates that, at least in a 2-detector, 2-setting, 2-outcome scenario, hidden variables can only be ruled out if detector efficiency exceeds the crucial $\frac{2}{3}$ threshold. Recent experiments [23, 24] indicate that this

may soon be realizable.

However, there are unresolved issues about the design and analysis of CH74-style experiments, which came to the fore in the analyses of [23, 24] and the subsequent discussion [24, 25, 26] about experimental loopholes and statistical analysis. The assumption of i.i.d. trials continues to be made, which is problematic because the possibility that a local hidden variable could exploit memory effects to violate (3.1) cannot be ruled out *a priori*. More generally, the importance of various experimental parameters would be greatly clarified by a precise mathematical exposition, as was done for the CHSH inequality in Chapter 2.

This chapter resolves these issues. In this chapter, we prove a version of (3.1) that allows us to experimentally test the inequality without having to make the i.i.d. assumption. By continuing to work in a general measure-theoretic setting, we will avoid the tacit assumption that the hidden variable is discrete or absolutely continuous, as is done in most physics papers. We will demonstrate a procedure to distill a test statistic from experimental data that can be used to discriminate between Quantum Mechanics and local theories. As will be seen, this requires some new machinery and techniques that were not seen in Chapter 2, due to different characteristics of the inequality (3.1). An experiment following the framework in this chapter will be immune to (legitimate) after-the-fact criticism such as was seen [24, 25, 26]: this framework clearly delineates all necessary experimental assumptions, along with an optimal method of statistical analysis.

In the second half of the chapter, we explore some other issues that arise in the CH74 setting. Certain statistical analyses require the use of *martingales*, and we will show in Section 3.7 how to derive exact p-values for such statistics, obtaining a meaningful improvement over the previously best-known bounds given in the 2013 paper [27]. We also will discuss the effect of changing the measurement setting probabilities (Section 3.8) and working with deterministic local hidden variables (Section 3.9).

3.2 The Mathematical Model

Following the strategy of the previous chapter, we model the CH74 experimental setup as a sequence of sets of random variables.

Definition 3.2.1. Let $(\Omega, \mathcal{F}, P)$ be a probability space. We define a sequence of random vectors:

$$\{D_{1i}, D_{2i}, A_i, B_i, \lambda_i\}_{i \in \mathbf{N}^+} \tag{3.2}$$

$$\lambda_i : \Omega \to \Lambda_i, \quad \Lambda_i \text{ is a measurable space,}$$

$$A_i : \Omega \to \{a, a'\}, \quad B_i : \Omega \to \{b, b'\},$$

$$D_{1i:}\Omega \to \{+1, 0\}, \quad D_{2i} : \Omega \to \{+1, 0\}.$$

For the ith trial, we call λ_i, A_i, B_i, D_{1i}, D_{2i} the joint state of the system, detector 1's setting, detector 2's setting, detector 1's output and detector 2's output, respectively.

We assume the random variables in Definition 3.2.1 satisfy the following properties:

Experimental Assumption 1:

$$\forall i, \quad A_i \perp\!\!\!\perp B_i. \tag{3.3}$$

Experimental Assumption 2:

$$\forall i, \quad P(A_i = a) = p_a, \quad P(B_i = b) = p_b, \quad p_a, p_b \in (0, 1). \tag{3.4}$$

This assumption has been weakened a bit. We assume that the measurement setting probabilities are fixed from trial to trial, but do not require that the probability is equal to $\frac{1}{2}$, even though this would be a natural choice. The weaker assumption (3.4) will allow for the results to apply more generally. We will also use the notation $q_a = 1 - p_a = P(A_i = a')$ and $q_b = 1 - p_b$.

Experimental Assumption 3:

$$A_i \perp\!\!\!\perp \lambda_i, \quad B_i \perp\!\!\!\perp \lambda_i. \tag{3.5}$$

As in the previous chapter, λ_i models the state of the system at the ith trial, when the previous $i-1$ trials have already taken place. Hence, we assume that the σ-algebras generated by the λ_i form a filtration – that is,

$$\text{For } i < j, \quad \sigma(\lambda_i) \subseteq \sigma(\lambda_j).$$

The more crucial point is given in the following assumption, which encapsulates the time-sequential ordering of the the experimental trials.

TIME SEQUENTIALITY: For any positive integer $n \geq 2$, let I be a subset of $\{1, 2, ..., n-1\}$ whose cardinality we denote with the letter m. Let $\vec{v}_1$, $\vec{v}_2$ be elements of $\{+1, 0\}^m$, let $\vec{w}_a$ be an element of $\{a, a'\}^m$, and let $\vec{w}_b$ be an element of $\{b, b'\}^m$. Then the following event is in $\sigma(\lambda_n)$:

$$\bigcap_{i \in I} \big[\{D_{1i} = \vec{v}_{1i}\} \cap \{D_{2i} = \vec{v}_{2i}\} \cap \{A_i = \vec{w}_{ai}\} \cap \{B_i = \vec{w}_{bi}\}\big]. \tag{3.6}$$

It should be noted that this σ-algebra will necessarily contain many other sets, such as single event sets like $\{A_3 = a'\}$.

So far, all of the assumptions are compatible with Quantum Mechanics. The next one, locality, is not. It is necessary to derive Bell-style inequalities like (3.1).

LOCALITY ASSUMPTION: Let $V_{i1} = (D_{i1}, A_i)$ and $V_{i2} = (D_{i2}, B_i)$. Then

$$(V_{i1} \perp\!\!\!\perp V_{i2}) \mid \lambda_i. \tag{3.7}$$

We introduce a useful shorthand, closely related to one used in the previous chapter:

$$+_{i1} := \{D_{i1} = +1\}, \quad +_{i2} := \{D_{i2} = +1\}, \quad 0_{i1} := \{D_{i1} = 0\}, \quad 0_{i2} := \{D_{i2} = 0\},$$
$$a_i := \{A_i = a\}, \quad a'_i := \{A_i = a'\}, \quad b_i := \{B_i = b\}, \quad b'_i := \{B_i = b'\}.$$

We also understand $P(a_i b_i)$ to be a shorthand for $P(a_i \cap b_i)$, and will write things like $P(+0 \mid a'_i b_i)$ instead of $P(+_{i1} \cap 0_{i2} \mid a'_i \cap b_i)$, if the order of terms can be unambiguously interpreted as a proxy for the subscripts of $+_{i1}$ and 0_{i2}.

Now, the ith trial of the experiment results in an output of each of the four observed random variables D_{i1}, D_{i2}, A_i, and B_i. Each of these takes an output in a binary set, so there are 16 different observable outcomes of the ith trial, such as $++ab$, or $0+ab'$. Sometimes it will be convenient to refer to a random variable T_i which we define to be the outcome of the ith trial, so T_i takes one of 16 different possible values. Hence the two expressions $P(T_i = +0ab)$ and $P(+0 a_i b_i)$ are two ways of writing the same thing.

3.3 Consequences of the Locality Assumption

In this section, we develop constraints that the T_i random variables must obey when we supplement the various experimental assumptions with the locality assumption, (3.7). As our goal

is a statistical test that is robust to the memory loophole, we desire a Bell-style inequality for T_i that includes conditioning on the outcomes of past trials. Thus we have the following proposition:

Proposition 3.3.1. *Fix a positive integer i. Let $\mathcal{D}$ be any event consisting of various outcomes of the random variables in a subset of $\cup_{j=1}^{i-1}\{D_{1j}, D_{2j}, A_j, B_j\}$, and so under* (3.6)*, $\mathcal{D} \in \sigma(\lambda_i)$. Then, under assumptions* (3.3)*,* (3.4)*,* (3.5)*,* (3.6)*, and* (3.7)*, the following inequalities hold:*

$$\frac{P\left(++a_ib_i \mid \mathcal{D}\right)}{p_ap_b} - \frac{P\left(+0a_ib_i' \mid \mathcal{D}\right)}{p_aq_b} - \frac{P\left(0+a_i'b_i \mid \mathcal{D}\right)}{q_ap_b} - \frac{P\left(++a_i'b_i' \mid \mathcal{D}\right)}{q_aq_b} \leq 0 \tag{3.8}$$

$$\frac{P\left(++a_ib_i' \mid \mathcal{D}\right)}{p_aq_b} - \frac{P\left(+0a_ib_i \mid \mathcal{D}\right)}{p_ap_b} - \frac{P\left(0+a_i'b_i' \mid \mathcal{D}\right)}{q_aq_b} - \frac{P\left(++a_i'b_i \mid \mathcal{D}\right)}{q_ap_b} \leq 0 \tag{3.9}$$

$$\frac{P\left(++a_i'b_i \mid \mathcal{D}\right)}{q_ap_b} - \frac{P\left(+0a_i'b_i' \mid \mathcal{D}\right)}{q_aq_b} - \frac{P\left(0+a_ib_i \mid \mathcal{D}\right)}{p_ap_b} - \frac{P\left(++a_ib_i' \mid \mathcal{D}\right)}{p_aq_b} \leq 0 \tag{3.10}$$

$$\frac{P\left(++a_i'b_i' \mid \mathcal{D}\right)}{q_aq_b} - \frac{P\left(+0a_i'b_i \mid \mathcal{D}\right)}{q_ap_b} - \frac{P\left(0+a_ib_i' \mid \mathcal{D}\right)}{p_aq_b} - \frac{P\left(++a_ib_i \mid \mathcal{D}\right)}{p_ap_b} \leq 0. \tag{3.11}$$

Proof. We will prove only the first of these inequalities, as the other three inequalities are equivalent to the first inequality up to permutation of $\{a, a'\}$ and $\{b, b'\}$, and the proof will not depend on a particular permutation of the settings.

Our strategy is similar to the one employed in the proof of Lemma 2.7.1. We start by showing that we can re-write the conditional probabilities as integrals. For a fixed outcome m in the range of T_i, we have:

$$\begin{aligned} P(T_i = m \mid \mathcal{D}) &= \frac{P(\{T_i = m\} \cap \mathcal{D})}{P(\mathcal{D})} \\ &= \frac{1}{P(\mathcal{D})} E(I_{\{T_i=m\}\cap\mathcal{D}}) \\ &= \frac{1}{P(\mathcal{D})} E\big[E(I_{\{T_i=m\}\cap\mathcal{D}} \mid \lambda_i)\big] \\ &= \frac{1}{P(\mathcal{D})} \int E(I_{\{T_i=m\}\cap\mathcal{D}} \mid \lambda_i) dP \\ &= \frac{1}{P(\mathcal{D})} \int P(\{T_i = m\} \cap \mathcal{D} \mid \lambda_i) dP. \end{aligned}$$

Now, $\mathcal{D}$ is in $\sigma(\lambda_i)$ by the time-sequential assumption (3.6). So,

$$\int P(\{T_i = m\} \cap \mathcal{D} \mid \lambda_i) dP = \int P(T_i = m \mid \lambda_i) I_{\mathcal{D}} dP,$$

which is a consequence of Theorem 9.1.3 in [10]. So the integral becomes

$$\int_{\mathcal{D}} P(T_i = m \mid \lambda_i) dP,$$

and we can rewrite the left side of (3.8) as

$$[P(a_i b_i)P(\mathcal{D})]^{-1} \int_{\mathcal{D}} P(++a_i b_i \mid \lambda_i) dP - [P(a_i b_i')P(\mathcal{D})]^{-1} \int_{\mathcal{D}} P(+0 a_i b_i' \mid \lambda_i) dP$$
$$- [P(a_i' b_i)P(\mathcal{D})]^{-1} \int_{\mathcal{D}} P(0 + a_i' b_i \mid \lambda_i) dP - [P(a_i' b_i')P(\mathcal{D})]^{-1} \int_{\mathcal{D}} P(++a_i' b_i' \mid \lambda_i) dP.$$

Recalling the μ notation from (2.33) in the last chapter, and applying (3.7) and (3.3), we can rewrite the above expression as

$$\frac{1}{P(\mathcal{D})} \int_{\mathcal{D}} \mu_{a_i}(+_{i1} \mid \lambda)\mu_{b_i}(+_{i1} \mid \lambda) - \mu_{a_i}(+_{i1} \mid \lambda)\mu_{b_i'}(0_{i2} \mid \lambda)$$
$$-\mu_{a_i'}(0_{i1} \mid \lambda)\mu_{b_i}(+_{i2} \mid \lambda) - \mu_{a_i'}(+_{i1} \mid \lambda)\mu_{b_i'}(+_{i2} \mid \lambda) dP. \tag{3.12}$$

For readability, let $p = \mu_{a_i}(+_{i1} \mid \lambda)$, $q = \mu_{b_i}(+_{i1} \mid \lambda)$, $r = \mu_{b_i'}(0_{i2} \mid \lambda)$, and $s = \mu_{a_i'}(0_{i1} \mid \lambda)$. By Lemma 1.3.1, we can also say that $1 - s = \mu_{a_i'}(+_{i1} \mid \lambda)$ and $1 - r = \mu_{b_i'}(+_{i2} \mid \lambda)$. Thus we rewrite (3.12) as

$$\frac{1}{P(\mathcal{C})} \int_{\mathcal{C}} pq - pr - sq - (1-s)(1-r) dP. \tag{3.13}$$

Lemma 2.5.2 tells us that p, q, r, and s are in $[0, 1]$ almost surely. With this fact, we now assert that the integrand in (3.13) is bounded above by 0, almost surely, which suffices to prove (3.8). We use a case analysis to prove the assertion.

Case 1: $q \leq r$.

$$\begin{aligned} pq - pr - sq - (1-s)(1-r) &= p(q-r) - sq - 1 + s + r - sr \\ &\leq -1 + s + r - sr \\ &= s(1-r) - 1 + r \\ &\leq 1 - r - 1 + r \\ &= 0. \end{aligned}$$

Case 2: $q > r$.

$$\begin{aligned} pq - pr - sq - (1-s)(1-r) &= p(q-r) - sq - 1 + s + r - sr \\ &\leq (q-r) - sq - 1 + s + r \\ &= q - sq - 1 + s \\ &= q(1-s) - 1 + s \\ &\leq 1 - s - 1 + s \\ &= 0. \end{aligned}$$

□

Proposition 3.3.1 is a necessary step towards our immediate goal of constructing a hypothesis test that distinguishes Quantum Mechanics from local theories. This is also a good time to prove that the distribution of experimental outcomes satisfies a certain condition known as *no-signaling*, as the proof technique is similar to that of Proposition 3.3.1. While the no-signaling result is not necessary for a hypothesis test of locality, it will help with some of our deeper analyses of local theories in later sections of this chapter.

Proposition 3.3.2. *(No-Signaling) Under the assumptions of Proposition 3.3.1, the following equations hold:*

$$P(++a_ib_i \mid \mathcal{D})/p_ap_b + P(+0a_ib_i \mid \mathcal{D})/p_ap_b = P(++a_ib'_i \mid \mathcal{D})/p_aq_b + P(+0a_ib'_i \mid \mathcal{D})/p_aq_b \quad (3.14)$$

$$P(++a'_ib_i \mid \mathcal{D})/q_ap_b + P(+0a'_ib_i \mid \mathcal{D})/q_ap_b = P(++a'_ib'_i \mid \mathcal{D})/q_aq_b + P(+0a'_ib'_i \mid \mathcal{D})/q_aq_b \quad (3.15)$$

$$P(++a_ib_i \mid \mathcal{D})/p_ap_b + P(0+a_ib_i \mid \mathcal{D})/p_ap_b = P(++a'_ib_i \mid \mathcal{D})/q_ap_b + P(0+a'_ib_i \mid \mathcal{D})/q_ap_b \quad (3.16)$$

$$P(++a_ib'_i \mid \mathcal{D})/p_aq_b + P(0+a_ib'_i \mid \mathcal{D})/p_aq_b = P(++a'_ib'_i \mid \mathcal{D})/q_aq_b + P(0+a'_ib'_i \mid \mathcal{D})/q_aq_b \quad (3.17)$$

Proof. We prove (3.14) only, as the other equalities are equivalent to (3.14) up to permutation of detector setting, and the proof of (3.14) does not depend on the particular setting permutation.

Using the justifications similar those employed in the proof of Proposition 3.3.1, we have

$$\begin{aligned}
&P(++a_ib_i \mid \mathcal{D})/p_ap_b + P(+0a_ib_i \mid \mathcal{D})/p_ap_b - P(++a_ib_i' \mid \mathcal{D})/p_aq_b - P(+0a_ib_i' \mid \mathcal{D})/p_aq_b \\
&= \frac{1}{P(\mathcal{D})}\int_{\mathcal{D}} P(++a_ib_i \mid \lambda_i)/p_ap_b + P(+0a_ib_i \mid \lambda_i)/p_ap_b - P(++a_ib_i' \mid \lambda_i)/p_aq_b - P(+0a_ib_i' \mid \lambda_i)/p_aq_b dP \\
&\qquad = \frac{1}{P(\mathcal{D})}\int_{\mathcal{D}} P(+a_i \mid \lambda_i)P(+b_i \mid \lambda_i)/p_ap_b + P(+a_i \mid \lambda_i)P(0b_i \mid \lambda_i)/p_ap_b \\
&\qquad\qquad - P(+a_i \mid \lambda_i)P(+b_i' \mid \lambda_i)/p_aq_b - P(+a_i \mid \lambda_i)P(0b_i' \mid \lambda_i)/p_aq_b. \quad (3.18)
\end{aligned}$$

By Lemma 1.3.1 and (3.5), $P(+b_i \mid \lambda_i) + P(0b_i \mid \lambda_i) = P(b_i \mid \lambda_i) = P(b_i)$, so we can rewrite (3.18) as

$$\begin{aligned}
&\frac{1}{P(\mathcal{D})}\int_{\mathcal{D}} P(+a_i \mid \lambda_i)P(b_i)/p_ap_b - P(+a_i \mid \lambda_i)P(b_i')/p_aq_b dP \\
&\qquad = \frac{1}{P(\mathcal{D})}\int_{\mathcal{D}} P(+a_i \mid \lambda_i)/p_a - P(+a_i \mid \lambda_i)/p_a dP = 0.
\end{aligned}$$

□

Remark 3.3.1. The conditioner $\mathcal{D}$ can be removed from the relations in Propositions 3.3.1 and 3.3.2, and the results still hold. To see this, merely let $\mathcal{D}$ be the set Ω, and the conditioner drops out of the expression. For the proofs of the propositions to work, the only necessary property of $\mathcal{D}$ was that it be in $\sigma(\lambda_i)$, which also of course holds of Ω. Furthermore, it can be noted that there are other no-signaling constraints that can be obtained by permuting the other events in the expressions (3.14)-(3.17), but the four equations listed here will be the most relevant to our work.

No-signaling constraints are a more general type of constraint than locality constraints such as the CHSH or CH74 inequalities. The no-signaling constraints assert that an observer at detector 1 cannot analyze the outcomes at his detector to gain information about the setting choice b or b' occurring at detector 2 (and symmetrically for an observer at detector 2). To illustrate this notion, consider (3.14) if $p_a = p_b = \frac{1}{2}$ and we take $\mathcal{D} = \Omega$. Then the equation reduces to

$$\begin{aligned}
P(++a_ib_i) + P(+0a_ib_i) &= P(++a_ib_i') - P(+0a_ib_i') \\
\Rightarrow \qquad P(+_1a_ib_i) &= P(+_1a_ib_i'), \quad (3.19)
\end{aligned}$$

and since $P(a) = P(a_i) = P(b_i) = P(b'_i) = \frac{1}{2}$, (3.19) can be re-written as

$$P(+_1 \mid a_i b_i) = P(+_1 \mid a_i b'_i).$$

The above expression makes the notion clear: whether detector 2 is set to b or b', the probability of a specific outcome at detector 1 does not change. What if this probability did change? In this circumstance, it is possible that an agent at detector 2 could send a faster-than-light signal to an observer at detector 1 by choosing a particular setting b or b'. In the most extreme circumstance, where $P(+_1 \mid a_i b_i) = 1$ and $P(+_1 \mid a_i b'_i) = 0$, the observer could instantaneously determine with certainty which setting had been chosen by the agent. Even if there is a smaller difference between $P(+_1 \mid a_i b_i)$ and $P(+_1 \mid a_i b'_i)$, the observer still gains some information. Hence in a no-signaling theory, the probabilities must be equal.

Lastly, we note that Quantum Mechanics can violate the constraints (3.8)-(3.9), but it cannot violate the no-signaling constraints (3.14)-(3.17). Hence Quantum Mechanics predicts nonlocal behavior, but this behavior cannot be utilized to send signals faster than the speed of light.

3.4 The Quantum Prediction

Quantum mechanics can violate the constraints given in Proposition 3.3.1 in a way that is robust to a certain amount of dropped signal. If λ is the singlet state λ_s and a, a', b, and b' are the angles given by the following assignment,

$$\theta_a = 135 \quad \theta_{a'} = 0 \quad \theta_b = 67.5 \quad \theta_{b'} = 202.5, \tag{3.20}$$

then we can compute the probabilities appearing in (3.8) using the methods of Section 1.2. However, we also want to take in to account detector inefficiency, which will require a slight tweak to the method. We define the *quantum efficiency* of the detector to be a number $\eta \in [0, 1]$, which we will interpret as the probability that a photon is detected. According to the standard quantum models of the experiment, this probability of detection at a single detector is independent of all the other components of the experiment, such as the outcome of the projective measurement and/or whether a detection occurs at the other detector. Now the probabilities of various outcomes can be computed by applying the standard projection measurement procedure, and then switching any $+1$ or -1 polarization outcomes to a "null" count with probability $1 - \eta$. Then the final "+1" results for D_1

and D_2 will consist of all the polarization $+1$ outcomes that weren't flipped to null counts, and the final "0" results will be obtained as the sum of all -1 polarization counts and null counts (recalling that -1 polarizations and null counts can be collapsed into a single outcome for counting purposes as was discussed in Section 3.1). We now assert that the probability of seeing various experimental results are as follows:

$$\begin{aligned} P(++ \mid \mathbf{ab}) &= \eta^2 P_{\text{quantum}}(++ \mid \mathbf{ab}\lambda_s) && (3.21)\\ P(+0 \mid \mathbf{ab}) &= \eta P_{\text{quantum}}(+- \mid \mathbf{ab}\lambda_s) + \eta(1-\eta) P_{\text{quantum}}(++ \mid \mathbf{ab}\lambda_s) && (3.22)\\ P(0+ \mid \mathbf{ab}) &= \eta P_{\text{quantum}}(-+ \mid \mathbf{ab}\lambda_s) + \eta(1-\eta) P_{\text{quantum}}(++ \mid \mathbf{ab}\lambda_s), && (3.23) \end{aligned}$$

where $\mathbf{a} \in \{a, a'\}$, $\mathbf{b} \in \{b, b'\}$, and $P_{\text{quantum}}(\cdot \mid \mathbf{ab}\lambda_s)$ is the standard quantum probability obtained via projective measurements with the angles given in (3.20). To see where (3.21) comes from, note that the only way to get a $(+1, +1)$ count for (D_1, D_2) is for both detectors to detect the polarization $+1$ *and* for the signal not to have been dropped at either detector, which occurs with probability η^2. In (3.22), a $+_1 -_2$ polarization result will be counted as a $+0$ outcome as long as the $+_1$ signal isn't dropped, and the probability of this is η. But a $+0$ outcome will also be recorded if measurement result is $+_1 +_2$, the first signal is not dropped, and the second signal *is* dropped, a scenario that occurs with probability $\eta(1-\eta)$. A symmetrical argument implies (3.23).

According to the quantum model of the experiment, outcomes of previous trials do not have an effect on probabilities, so the probabilities appearing in (3.8) can be computed by ignoring the conditioner $\mathcal{D}$. Then the left side of (3.8) can be calculated with the help of the following conditional probabilities, computed using expressions (3.21)-(3.23):

$$\begin{aligned} P(++ \mid ab) &= \eta^2 \left(\frac{2+\sqrt{2}}{8}\right)\\ P(+0 \mid ab') &= \eta \left(\frac{2-\sqrt{2}}{8}\right) + \eta(1-\eta)\left(\frac{2+\sqrt{2}}{8}\right)\\ P(0+ \mid a'b) &= \eta \left(\frac{2-\sqrt{2}}{8}\right) + \eta(1-\eta)\left(\frac{2+\sqrt{2}}{8}\right)\\ P(++ \mid a'b') &= \eta^2 \left(\frac{2-\sqrt{2}}{8}\right). && (3.24) \end{aligned}$$

As the left side of (3.8) can now be written

$$P(++ \mid ab) - P(+0 \mid ab') - P(0+ \mid a'b) - P(++ \mid a'b'), \qquad (3.25)$$

we can calculate the quantum prediction. If $\eta = 1$ (perfect detection efficiency), the quantity (3.25) is equal to $\frac{4\sqrt{2}-4}{8}$, so (3.8) is violated. Clearly if we decrease η by just a little bit, the inequality (3.8) will still be violated, as (3.25) is a continuous function of η. The natural question ask is how low η can go while still violating (3.8). As (3.25) is quadratic in η, it is straightforward to analyze the behavior of (3.25) with respect to this parameter, and so we find that for $\eta \in [0, 1]$,

$$(3.25) > 0 \qquad \Leftrightarrow \qquad \eta > \frac{2\sqrt{2}}{2+\sqrt{2}} \approx .83 \tag{3.26}$$

So 83% is the crucial cut-point for efficiency required for an experiment attempting to violate a Bell inequality with the singlet state.

As mentioned in Section 3.1, other quantum states can violate Bell inequalities while tolerating even lower detection efficiencies. One such family of states are known as the Eberhard states, parameterized by a fixed choice of $r \in (0, 1)$:

$$\lambda_r^e = \frac{1}{\sqrt{1+r^2}} \left(|01\rangle + r|10\rangle \right). \tag{3.27}$$

It was explicitly shown in [20] that for certain judicious choices of r and corresponding choices for the angles θ_a, $\theta_{a'}$, θ_b, and $\theta_{b'}$, quantum probabilities can be obtained so that (3.25) is positive for $\eta = \frac{2}{3} + \epsilon$, for values of ϵ as low as 10^{-3}. There are no quantum states that can violate (3.8) for values of η below $\frac{2}{3}$ [22, 21].

This completes the analysis of detector efficiency, but there is another experimental issue that we have not yet mentioned that can create problems. This is the fact that quantum states cannot be generated on demand. To create photon states, one can attenuate the power of a laser until it is almost completely off. If this is done correctly, during any finite time window, photon states will be emitted according to a *Poisson distribution*. A Poisson random variable is a random variable $\mathrm{Poi} : \Omega \to \mathbb{N}$ for which

$$\forall k \in \{0, 1, ..., n, ...\}, \quad P(\mathrm{Poi} = k) = \frac{\lambda_{\mathrm{poi}}^k}{k!} e^{-\lambda_{\mathrm{poi}}}, \tag{3.28}$$

where $\lambda_{\mathrm{poi}} > 0$ is a parameter depending on the power of the laser and the length of the time window. (Unfortunately, the symbol "λ" is generally used for Poisson parameters, so here we denote it λ_{poi} in order to distinguish it from our preexisting definition of λ.) Hence if one is trying to generate a photon state in a given time window, there is a probability $e^{-\lambda_{\mathrm{poi}}}$ of not generating a photon, a

probability $\lambda_{\text{poi}} e^{-\lambda_{\text{poi}}}$ of seeing one photon (the desired outcome), and various probabilities of seeing two, three, or more photons, so-called "multi-photon states."

Referring back to the scheme of Figure 2.1, if the "Photon Generator" is an attenuated laser, the "Beam-splitter" (which in reality may be a complex system of multiple apparatuses, but the details are not relevant to our discussion) can be tuned to reliably produce the entangled state λ_s (or alternatively, λ_r^e) when the photon generator emits a single photon. However, if the photon generator emits a multi-photon state, the beam-splitter will not generate the desired state λ_s, but will instead produce a state that will produce a different distribution over polarization outcomes, and cannot be counted on to violate the Bell inequality (3.8). To get around this problem, the laser power can be turned down, decreasing the value of λ_{poi}. As this is done, the probability of a single photon emission goes down, but the probability of multi-photon states decreases at a far faster rate – this can be seen by inspecting (3.28). Hence the ratio of single photons to multi photons can be made arbitrarily large, and so the effect of multi-photon states on the probabilities (3.25) can be made as small as desired. The price of this tactic is that there will be a large number of time windows in which no photon is produced, but if these states always result in $D_1 = D_2 = 0$ counts, this will not adversely affect the violation of (3.8). (In absolute terms, the probabilities appearing in (3.25) will all grow smaller if more and more no-photon states are produced, but the violation of (3.8) depends only on the size of these probabilities relative to one another.)

Unfortunately, $P(0_1 0_2 \mid \text{no photon produced})$ is not exactly one in practice, but very slightly less than one. Photon detectors tend to have a tiny probability of occasionally producing a $+1$ or -1 count when there is no photon present; this phenomenon is known as a *dark count*. In many experimental scenarios, the dark counts have a low enough probability to not have a large effect on the overall value of (3.25), even if the laser is very attenuated so that λ_{poi} is a small number. However, for certain Eberhard states, the dark counts do begin to become a problem. It turns out that the more λ_r^e is optimized to lower the needed efficiency η to violate (3.8), the more the dark count issue starts to adversely affect the value of (3.25). Thus for given experimental constraints on detector efficiency and dark counts, it takes some effort to find the optimal state λ_r^e for violating (3.8).

The paper [20] outlines an approach for finding the optimal Eberhard state to violate (3.8) given fixed experimental parameters for η and the dark count rate. This procedure was followed by two recent experiments [23] and [24] taking place in 2013, and violations of (3.8) were found using Eberhard states. Now, neither of these experiments enforced strict space-like separation of

the photon detectors, so the implied non-satisfaction of (3.7) cannot be interpreted as irrefutable evidence of non-local behavior – correlation between the systems could theoretically be achieved with sub-luminal signals. Nonetheless, these experiments do represent important progress towards a loophole-free test of a Bell inequality, as they are the first Bell experiment with photons that do not make any fair-sampling assumptions in analyzing the experimental data. Furthermore, they provide us with a great example of the kind of raw data one might expect to see in a loophole-free Bell test, and therefore can provide us with a good reference point when we are assessing the efficacy of various statistical tests that we explore in this chapter.

Using the data from these experiments, we can form an empirical distribution for the twelve non-00 outcomes of the Bell experiment, given in Table 3.1. Due the Poissonian nature of the photon source, most of the experimental outcomes were actually 00, so the probabilities in Table 3.1 are conditioned on the occurrence of a non-00 outcome.

Table 3.1: Empirical distributions for the experiments [23] (left) and [24] (right).

	++	+0	0+
ab	.050	.021	.029
ab'	.054	.017	.157
$a'b$	.056	.165	.023
$a'b'$	.003	.217	.207

	++	+0	0+
ab	.044	.026	.026
ab'	.049	.020	.162
$a'b$	.051	.172	.019
$a'b'$	.003	.219	.209

Note that both probability distributions in Table 3.1 violate (3.8) and satisfy the other three inequalities (3.9)-(3.11). While an eyeball analysis indicates a violation of (3.8), this could be a statistical fluctuation, and so we desire a rigorous hypothesis test that can quantify the degree of violation of locality by assigning a p-value to the extremity of a test statistic, which we will do in the next section.

The experiment [24] also reported the number of 00 counts, illustrating how the multi-photon issue is controlled by weakening the laser power. For one particular setting configuration, the Christensen experiment saw about 150,000 non-00 counts in about 27,900,000 trials, and for the Poisson probability mass function given by (3.28), this suggests a parameter in the range of $\lambda_{\text{poi}} = .005$. For this choice of λ_{poi}, in any given trial window there is about a .995 chance of no photon, about a .005 chance of a single photon, and less than a .000015 chance of a multi-photon state (two or more photons). So with these numbers, if there is a photon, there is only about a .003 chance that it is an (undesirable) multi-photon state, and potentially adverse effects on (3.25) are minimal.

3.5 Defining a New Test Statistic

In Chapter 2, we proceeded directly from the constraint (2.43) to a statistic C_i whose expectation was equal to the quantity on the left side of (2.43). To follow this tactic here, we would start with the constraint (3.8), and define a statistic as follows:

$$D_i = \begin{cases} (p_a p_b)^{-1}, & \text{if } (A_i, B_i) = (a, b) \text{ and } (D_{1i}, D_{2i}) = (+1, +1) \\ -(p_a q_b)^{-1}, & \text{if } (A_i, B_i) = (a, b') \text{ and } (D_{1i}, D_{2i}) = (+1, 0) \\ -(q_a p_b)^{-1}, & \text{if } (A_i, B_i) = (a', b) \text{ and } (D_{1i}, D_{2i}) = (0, +1) \\ -(q_a q_b)^{-1}, & \text{if } (A_i, B_i) = (a', b') \text{ and } (D_{1i}, D_{2i}) = (+1, +1) \\ 0, & \text{in all other situations.} \end{cases}$$

The statistic D_i has expectation exactly equal to the left side of expression (3.8). Hence a local theory will predict long-term D_i averages that are less than or equal to 0, while a quantum theory with outcome probabilities like those in Table 3.1 will predict long term D_i averages converging to a positive quantity.

Thus it seems that the statistic $\sum_{i=1}^{n} D_i$ could be used to distinguish the two theories. However, there is a problem with this scenario, related to the possible "0" outcome for D_i. Indeed, we have seen in Section 3.4, according to quantum mechanics, the D_i output is *usually* zero: in addition to the majority of time windows when no photon pair is produced, there are also detection events, such as $++a'b$, for which the D_i output is still zero. Now, according to Quantum Mechanics $E(\overline{D_n})$ is positive and does increase slowly with time, even as most trials result in $D_i = 0$. So it still seems as though one could form a hypothesis test where one rejects the null hypothesis if after a sufficiently large number of trials n, $\overline{D_n} \geq \delta$, where $\delta > 0$ is some judiciously chosen positive real number.

Unfortunately, there would be no way to glean a useful p-value from such a test. For an illustration of what goes wrong, suppose $p_a = p_b = \frac{1}{2}$, so D_i always takes a value in the set $\{-4, 0, +4\}$. Now $\sum D_i$ is a random walk on a lattice consisting of integer multiples of 4. Under a quantum mechanical theory, one would predict that most D_i would be 0 – for instance, if the Poisson parameter were about $\lambda_{\text{poi}} = .005$ (such as was seen in Section 3.4), about 99.5% of the trials would not result in a photon state being produced, and thus automatically yield a 0 output for D_i. Among the less than .5% of the D_i assuming values in the subset $\{-4, +4\}$, the +4 values

would slightly predominate. However, a local theory obeying (3.8) could simulate D_i behavior where D_i *always* takes a nonzero value, with a 50-50 distribution on $\{-4, +4\}$. After say, $n = 2,000,000$ trials, assume that the quantum D_i is expected to take about $10,000$ values in the set $\{-4, +4\}$. (It would actually be even less than that, if we factor in the fact that many actual detection events still result in D_i equaling zero.) The expectation value of $E(\sum D_i)$ under a quantum theory, suggested by the probabilities on the right side of Table 3.1, would then be 232, so we could choose δ to be, say, 115. Under the local theory that never results in $D_i = 0$ trials, the expectation of $\sum D_i$ is 0 but the standard deviation is about 700. Then the probability of exceeding δ under a local theory – the p-value – is roughly .435, only slightly less than $\frac{1}{2}$!

It is theoretically possible to get around this bug by going out to very large n, because the quantum expectation $E(\sum D_i)$ grows linearly with n, whereas the local theory standard deviation grows linearly with $\sqrt{n}$. However, there are more efficient approaches. The best solution is to condition on a nonzero outcome, and thus analyze only the experimental trials where D_i moves.

To proceed in this manner, we first need to show that if we focus only on the subcollection of experimental trials taking a value in a subset of the experimental outcomes, that constraints like $(3.8) - (3.11)$ and $(3.14) - (3.17)$ will still apply to the subcollection. This is intuitively plausible, but it requires proof. The result will be used in a few different situations in subsequent sections, so we prove it in a fairly general form.

Proposition 3.5.1. *Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of random variables taking values in a finite set S. Suppose the X_i satisfy the following condition: there is a collection of real constants $\{c_1, ..., c_n\}$ and a subset $W = \{s_1, ..., s_n\} \subseteq S$ such that for any fixed i,*

$$\sum_{j=1}^{n} c_j P(X_i = s_j \mid \mathcal{X}) \leq 0, \tag{3.29}$$

where $\mathcal{X}$ is any fixed event of the form $(X_1, ..., X_{i-1}) = \vec{v}$, $\vec{v} \in S^{i-1}$ for which $P(\mathcal{X}) > 0$.

Now suppose V is a subset of S that contains W, where we write $V = \{s_1, ..., s_n, t_1, ..., t_m\}$ (omitting the t_i if $V = W$), and we define the new random variable sequence $\{Y_k\}_{k=1}^{\infty}$ as follows:

$$Y_k = \text{ the value of the } k\text{th } X_i \text{ taking a value in the set } V \tag{3.30}$$

and we define Y_k to be 0 if fewer than k X_i take a value in the set V.

Then the following holds for any fixed value of k:

$$\sum_{j=1}^{n} c_j P(Y_k = s_j \mid \mathcal{Y}) \leq 0, \tag{3.31}$$

where $\mathcal{Y}$ is any fixed event of the form $(Y_1, ..., Y_{k-1}) = \vec{v}$, $\vec{v} \in V^{k-1}$ for which $P(\mathcal{Y}) > 0$.

Proof. It will help to examine the definition of Y_k a little more closely. For each $r \in V$, we have

$$\begin{aligned} &Y_k(\omega) = \\ &\quad r \text{ if } \omega \in \bigcup_{i=k}^{\infty} \left[\begin{array}{c} \{\text{Exactly } k-1 \text{ random variables in the set } \{X_j \mid j < i\} \text{ assume a value in V}\} \\ \cap \quad \{X_i = r\} \end{array} \right], \end{aligned} \tag{3.32}$$

while $Y_k = 0$ if none of these conditions apply. Note that the union in (3.32) is disjoint, so we can add up over probabilities of events like this (defined in terms of the X_j) to get the probabilities of Y_k-type events. Now consider an event $\mathcal{Y}$. $\mathcal{Y}$ can be broken down into disjoint union of sets of the form $\mathcal{X}_i = \{(X_1, ..., X_{m_i}) = \vec{w_i}\}$, where m_i is an integer greater than or equal to k (m_i may vary for different choices of $\mathcal{X}_i$), and $\vec{w_i}$ is an m_i-dimensional vector with exactly $k-1$ entries assuming a value in the set V such that the final entry is one of the entries in V. Let I be an indexing set for all such sets $\mathcal{X}_i$ of positive probability that comprise $\mathcal{Y}$. Then we can write $\mathcal{Y} = \cup_{i \in I} \mathcal{X}_i$. Now consider

$$\begin{aligned} \sum_{j=1}^{n} c_j P(Y_k = s_j \cap \mathcal{Y}) &= \sum_{j=1}^{n} c_j \sum_{i \in I} P(Y_k = s_j \cap \mathcal{X}_i) \\ &= \sum_{j=1}^{n} c_j \sum_{i \in I} \sum_{l=m_i+1}^{\infty} P\left(\mathcal{X}_i \cap \{\forall r(m_i < r < l), X_r \notin V\} \cap \{X_l = s_j\}\right) \\ &= \sum_{j=1}^{n} \sum_{i \in I} \sum_{l=m_i+1}^{\infty} c_j P\left(\{X_l = s_j\} \cap \mathcal{X}_i \cap \{\forall r(m_i < r < l), X_r \notin V\}\right) \\ &= \sum_{i \in I} \sum_{l=m_i+1}^{\infty} \sum_{j=1}^{n} c_j P\left(\{X_l = s_j\} \cap \mathcal{X}_i \cap \{\forall r(m_i < r < l), X_r \notin V\}\right). \end{aligned}$$

In the last step, it was possible to switch the order of summation because the infinite sum converges absolutely, due to our working in a probability space. Recalling (3.29), the expression $\mathcal{X}_i \cap \{\forall r(m_i < r < l), X_r \notin V\}$ can serve as a "$\mathcal{X}$" event for $\{X_l = s_j\}$. As we can multiply both

sides of (3.29) by $P(\mathcal{X})$ to turn the conditional probability into a joint probability, we see that

$$\sum_{j=1}^{n} c_j P\left(\{X_l = s_j\} \cap \mathcal{X}_i \cap \{\forall r(m_i < r < l), X_r \notin V\}\right) \leq 0,$$

which, in light of what we have already derived, implies that

$$\sum_{j=1}^{n} c_j P(Y_k = s_j \cap \mathcal{Y}) \leq 0 \quad \Rightarrow \quad \sum_{j=1}^{n} c_j P(Y_k = s_j \mid \mathcal{Y}) \leq 0.$$

□

Recall the sequence of T_i random variables, which take outputs in a sixteen element set. This will be our "X_i" sequence as defined in Proposition 3.5.1. Let K be the subset of 12 outcomes that do not result in a 00, so for example, $+0ab \in K$, $00a'b' \notin K$. Then we can define a dependent sequence U_j in the following manner:

$$U_j := \text{ the } j\text{th } T_i \text{ taking a value in } K. \tag{3.33}$$

Table 3.1 actually gives distributions for the U_j random variables, as opposed to the T_i random variables. Proposition 3.5.1 now allows us to say that the U_j random variables obey the constraints (3.8)-(3.11) and (3.14)-(3.17) under locality.

To come up with a test statistic, we want to restrict the range even further, and define the set $W = V$ in Proposition 3.5.1 to be $\{++ab, +0ab', 0+a'b, ++a'b'\}$. Then we will have a random variable sequence Y_k that takes a value only when the underlying X_i sequence (here, T_i) assumes one of these four values. To turn this into a test statistic, we need a random variable taking values in the real numbers, as opposed to a collection of symbols. We also need to address the degenerate situation that occurs when $Y_k = 0$, which occurs when fewer than k X_i random variables take a value in the set K. The following proposition allows us to do these things.

Proposition 3.5.2. *Under the assumptions of Proposition 3.5.1 with $V = W$, define a random*

variable sequence $\{H_k\}_{k=1}^{\infty}$ *as follows:*

$$H_k = \begin{cases} c_1 & \text{if } Y_k = s_1 \\ \vdots & \\ c_n & \text{if } Y_k = s_n \\ d_1 & \text{if } Y_k = 0, \end{cases}$$

where d_1 *is a number chosen to equal to one of the* c_i*. Re-label the outcome space of* H_k *as* $\{d_1, ..., d_m\}$*, so* m *is at most equal to* n*, and is strictly less than* n *if there is are choices of* $i \neq j$ *for which* $c_i = c_j$*.*

Then if d_1 *is not positive, the following holds for any fixed value of* k*:*

$$\sum_{j=1}^{m} d_j P(H_k = d_j \mid \mathcal{H}) \leq 0, \tag{3.34}$$

where $\mathcal{H}$ *is any fixed event of the form* $(H_1, ..., H_{k-1}) = \vec{v}, \vec{v} \in \{d_1, ..., d_m\}^{k-1}$ *for which* $P(\mathcal{H}) > 0$*.*

Proof. The event $\mathcal{H}$ can be written as a disjoint union of events in the following way:

$$\mathcal{H} = \left(\cup_{a \in A} \mathcal{Y}_a^0\right) \bigcup \left(\cup_{b \in B} \mathcal{Y}_b\right), \tag{3.35}$$

where A indexes the (possibly empty) collection of positive-probability events $\mathcal{Y}_a^0$ of type $(Y_1, ..., Y_{k-1}) = \vec{v}$ where $\vec{v}$ contains at least one 0 entry, and B indexes the (possibly empty) collection of positive-probability events $\mathcal{Y}_b$ of type $(Y_1, ..., Y_{k-1}) = \vec{v}$ where $\vec{v}$ does not contain a 0 entry. At least one of A or B must be nonempty.

Note that if an event of type $\mathcal{Y}_a^0$ occurs, then $H_k = d_1$ with probability 1, so $P(H_k = d_i \cap Y_a^0) = \delta_{i1}$, where δ_{i1} is 1 if $i = 1$ and 0 otherwise. This is because if any particular Y_j equals 0, then all subsequent Y_j must equal 0 as well. This implies that

$$\sum_{j=1}^{m} d_j P(H_k = d_j \cap \mathcal{Y}_a^0) = d_1 \leq 0.$$

As for events of type $\mathcal{Y}_b$, we have

$$\begin{aligned}\sum_{j=1}^{m} d_j P(H_k = d_j \cap \mathcal{Y}_b) &= d_1 P(Y_k = 0 \cap \mathcal{Y}_b) + \sum_{j=1}^{n} c_j P(Y_j = c_j \mid \mathcal{Y}_b) \\ &\leq \sum_{j=1}^{n} c_j P(Y_j = c_j \mid \mathcal{Y}_b) \\ &\leq 0\end{aligned}$$

where the first inequality follows from the fact that $d_1 \leq 0$, and the second inequality follows from (3.31).

Summing over all the constituent events of $\mathcal{H}$ as given in (3.35), this implies that

$$\sum_{j=1}^{m} d_j P(H_k = d_j \cap \mathcal{H}) \leq 0,$$

and dividing both sides by $P(\mathcal{H})$ yields (3.34). □

Remark 3.5.1. Note that in the proofs of Propositions 3.5.1 and 3.5.2, it was critical that the upper bound in the inequality was 0, as this allowed us to freely move between conditional and joint probabilities by multiplying or dividing both sides of the expression by the conditioner. This is an important characteristic of CH-style inequalities that can be used for statistical tests.

These results will allow us to define a statistic that differentiates quantum mechanics from local theories. To see this, let us first make the simplifying assumption that $p_a = p_b = \frac{1}{2}$. Then if we take V to be $\{++ab, +0ab', 0+a'b, ++a'b'\}$, the derived random variable H_k will take values in the set $\{-4, 4\}$. To simplify things, let us take a scaled version $J_k = \frac{1}{4}H_k$, so now we have

$$J_k = \begin{cases} +1 & \text{if the } k\text{th } T_i \text{ taking a value in } V \text{ is } ++ab \\ -1 & \text{if the } k\text{th } T_i \text{ taking a value in } V \text{ is not } ++ab \\ -1 & \text{if fewer than } k \text{ of the } T_i \text{ take a value in the set } V. \end{cases} \tag{3.36}$$

It should be pointed out that assigning the -1 result to sequences that take fewer than k values in the set of interest is just a mathematical convenience. In practice, statistical analysis of the results will only be performed on the J_i values that were set by actual V outcomes, so we should not be too distracted by this. To perform this statistical analysis, we note that in light of Proposition 3.5.2,

expression (3.8) implies

$$P\left(J_i = +1 \mid \mathcal{J}\right) \leq P(J_i = -1 \mid \mathcal{J}). \tag{3.37}$$

This means that $P\left(J_i = +1 \mid \mathcal{J}\right) \leq \frac{1}{2}$. Recalling a result from the previous chapter, Proposition 2.7.2, under a local theory J_k can do no better at accumulating more $+1$ than -1 counts than a binomial random variable with a $\frac{1}{2}$ chance of resulting in $+1$ each trial. If we think of J_i as a random walk on the integers that starts at 0, then $J := \sum_{i=1}^{n} J_i$ cannot do any better at attaining large positive results than the simple random walk that at each step moves up or down independently with probability $\frac{1}{2}$.

The data of the experiment [24] suggests that a J_i random variable described by quantum theory could violate (3.37) with an i.i.d. probability distribution for which $P(J_i = +1)$ is roughly .5116. (This figure comes from combining the appropriate probabilities in Table 3.1.) This is not a large deviation above 50%, but any difference will be statistically detectable with only a moderately large number of trials, now that we are discarding any trial that doesn't give us one of the 4 outcomes of interest. For instance, after 20,000 trials resulting in an output in V, a sequence of random variables J_i satisfying (3.37) would have (at best) no more than a 0.24% chance of seeing 51% or more of the trials resulting in "+1."

3.6 Martingale Statistics

We have just seen that the statistic J can be used to distinguish Quantum Mechanics from local theories in a CH-style experiment. However, other statistics can do this as well. In this section, we explore another such statistic Ch (to be defined) that is more closely related to the original CH74 inequality (3.1). Unfortunately, the analysis of this statistic turns out to be significantly more complicated than the analysis of J. One motivation for delving into the study of this more complicated statistic is that different statistics can demonstrate significance with fewer trials, and Ch turns out have this property in some experimental settings. However, this effect is small, and could possibly also be achieved with minor modifications to the J statistic that do not complicate the statistical analysis so severely. A stronger motivation for studying Ch is that the methods introduced are widely applicable to other scenarios in which J-style statistics are not possible. This is the situation, for instance, in a CH-style experiment for which the setting probabilities are anything *other* than 50-50 at both ends. Also, there are other Bell locality scenarios with different numbers of detectors and/or measurement settings and/or outcomes. In some of these scenarios, only the

more-complicated statistics like Ch will be available. Hence the new techniques introduced in this section will also find uses outside of the analysis of the particular statistic Ch.

For the rest of this section, we continue to make the simplifying assumption that $p_a = p_b = \frac{1}{2}$. Then the inequality (3.8) becomes

$$P\left(++a_ib_i \mid \mathcal{D}\right) - P\left(+0a_ib_i' \mid \mathcal{D}\right) - P\left(0+a_i'b_i \mid \mathcal{D}\right) - P\left(++a_i'b_i' \mid \mathcal{D}\right) \leq 0, \tag{3.38}$$

where we recall that $\mathcal{D}$ can be any particular event consisting of various outcomes of experimental trials prior to the ith trial. This assumption also simplifies the no-signaling relations of Proposition 3.3.2. In particular, (3.14) and (3.16) can now be written as

$$P(++a_ib_i \mid \mathcal{D}) + P(+0a_ib_i \mid \mathcal{D}) = P(++a_ib_i' \mid \mathcal{D}) + P(+0a_ib_i' \mid \mathcal{D}) \tag{3.39}$$

$$P(++a_ib_i \mid \mathcal{D}) + P(0+\mathrm{a}_ib_i \mid \mathcal{D}) = P(++a_i'b_i \mid \mathcal{D}) + P(0+a_i'b_i \mid \mathcal{D}). \tag{3.40}$$

Subtracting (3.39) from (3.38) and subtracting (3.40) from (3.38) yield

$$P(++a_ib_i' \mid \mathcal{D}) - P(+0a_ib_i \mid \mathcal{D}) - P\left(0+a_i'b_i \mid \mathcal{D}\right) - P\left(++a_i'b_i' \mid \mathcal{D}\right) \leq 0 \tag{3.41}$$

$$P(++a_i'b_i \mid \mathcal{D}) - P\left(+0a_ib_i' \mid \mathcal{D}\right) - P(0+\mathrm{a}_ib_i \mid \mathcal{D}) - P\left(++a_i'b_i' \mid \mathcal{D}\right) \leq 0, \tag{3.42}$$

and the sum of (3.41) and (3.42) is

$$\begin{aligned} &P(++a_ib_i' \mid \mathcal{D}) + P(++a_i'b_i \mid \mathcal{D}) - 2P\left(++a_i'b_i' \mid \mathcal{D}\right) \\ &\qquad - P(+0a_ib_i \mid \mathcal{D}) - P\left(0+a_i'b_i \mid \mathcal{D}\right) - P\left(+0a_ib_i' \mid \mathcal{D}\right) - P(0+\mathrm{a}_ib_i \mid \mathcal{D}) \leq 0. \end{aligned} \tag{3.43}$$

It turns out that (3.43) is equivalent to the original CH74 inequality (3.1). To see this, add and subtract the quantity $P(++a_ib_i' \mid \mathcal{D}) + P(++a_i'b_i \mid \mathcal{D}) + 2P(++a_ib_i' \mid \mathcal{D})$ from (3.43), and then use the facts that

$$P(+_1a_i \mid \mathcal{D}) = P(++a_ib_i \mid \mathcal{D}) + P(+0a_ib_i \mid \mathcal{D}) + P(++a_ib_i' \mid \mathcal{D}) + P(+0a_ib_i' \mid \mathcal{D}),$$

$$P(+_2b_i \mid \mathcal{D}) = P(++a_ib_i \mid \mathcal{D}) + P(0+a_ib_i \mid \mathcal{D}) + P(++a_i'b_i \mid \mathcal{D}) + P(0+a_i'b_i \mid \mathcal{D}),$$

to obtain

$$2P(++a_ib_i \mid \mathcal{D})+2P(++a_ib_i' \mid \mathcal{D})+2P(++a_i'b_i \mid \mathcal{D})-2P(++a_i'b_i' \mid \mathcal{D})-P(+_1a_i \mid \mathcal{D})-P(+_2b_i \mid \mathcal{D}) \leq 0. \tag{3.44}$$

As noted in Remark 3.3.1, the conditioner $\mathcal{D}$ can be removed without changing the validity of (3.44). Then, using the fact that $P(a_ib_i)$ and similar terms are equal to $\frac{1}{4}$, and $P(a_i) = P(b_i) = \frac{1}{2}$, (3.44) is equivalent to

$$P(++ \mid a_ib_i) + P(++ \mid a_ib_i') + P(++ \mid a_i'b_i) - P(++ \mid a_i'b_i') - P(+_1 \mid a_i) - P(+_2 \mid b_i) \leq 0,$$

which is the original CH74 inequality.

Now we return to (3.43), as this is the most useful form of the CH74 inequality for our current purpose. We can define a statistic Ch whose expectation is equal to the left side of (3.43), and this statistic can thusly be taken as a measure of violation of the original CH74 inequality.

Now let V be the collection of seven outcomes appearing in the left side of inequality (3.43). Then define

$$Ch_k = \begin{cases} +1 & \text{if the } k\text{th } T_i \text{ taking a value in } V \text{ is } ++ab' \text{ or } ++a'b \\ -1 & \text{if the } k\text{th } T_i \text{ taking a value in } V \text{ is } +0ab,\ 0{+}ab,{+}0ab' \text{ or } 0{+}a'b \\ -2 & \text{if the } k\text{th } T_i \text{ taking a value in } V \text{ is } ++a'b' \\ -2, & \text{in all other cases.} \end{cases}$$

Then by applying Proposition 3.5.2 with the constraint (3.43), we have

$$P(Ch_k = +1 \mid \mathcal{CH}) \leq P(Ch_k = -1 \mid \mathcal{CH}) + 2P(Ch_k = -2 \mid \mathcal{CH}) \tag{3.45}$$

where $\mathcal{CH}$ is any positive-probability event of the form $(Ch_1, ..., Ch_{k-1}) = \vec{v}$, $\vec{v} \in \{-2, -1, +1\}^{k-1}$. Once again, Ch_k is a random walk on the integers, whose expected value is 0 or less. However, the outcome of Ch_k is not binary and so Proposition 2.7.2 cannot be applied. A different method of statistical analysis must be used.

To use this method, define W_n to be the sum of the first n instances of Ch_i: $W_n = \sum_{i=1}^{n} Ch_i$. Now, expression (3.45) allows us to demonstrate that the sequence $\{W_i\}_{i \in \mathbb{N}}$ is a *supermartingale*, as defined in Chapter 9 of [10]:

Definition 3.6.1. A sequence of random variables $\{X_i\}_{i\in\mathbb{N}^+}$ and σ-algebras $\{\mathcal{F}_i\}_{i\in\mathbb{N}^+}$ is called a *supermartingale* iff we have for each i,

1. $\mathcal{F}_i \subseteq \mathcal{F}_{i+1}$ and $X_i \in \mathcal{F}_i$
2. $E(|X_i|) < \infty$
3. $E(X_{i+1} \mid \mathcal{F}_i) \leq X_i, \quad a.s.$

Proposition 3.6.1. *The sequence $\{W_i\}_{i\in\mathbb{N}^+}$ is a supermartingale with respect to the sequence $\{\mathcal{G}_i\}_{i\in\mathbb{N}^+}$, where $\mathcal{G}_i = \sigma(Ch_1, ..., Ch_i)$. That is to say, $\forall n \in \mathbb{N}^+$, we have $\mathcal{G}_n \subseteq \mathcal{G}_{n+1}$, $W_n \in \mathcal{G}_n$, $E|W_n| < \infty$, and*

$$E(W_{n+1} \mid \mathcal{G}_n) \leq W_n. \tag{3.46}$$

Proof. Clearly, $\mathcal{G}_n \subseteq \mathcal{G}_{n+1}$, and $W_n \in \mathcal{G}_n$ as W_n is the sum of the random variables $\{Ch_1, ..., Ch_n\}$ that generate $\mathcal{G}_n$. Furthermore, $E|W_n| < \infty$, as $E|W_n| = E\big|\sum_{i=1}^n Ch_i\big| \leq E\big(\sum_{i=1}^n |Ch_i|\big)$, and $E\big(\sum_{i=1}^n |Ch_i|\big)$ is finite because for each Ch_i, $E(|Ch_i|) \leq 2$ as Ch_i only takes values in the set $\{-2, -1, +1\}$.

This leaves only (3.46) to be demonstrated. Consider now $E\big(W_{n+1} \mid (Ch_1, ..., Ch_n) = \vec{v}\big)$, where $\vec{v}$ is a fixed element of the set $\{-2, -1, +1\}^n$. We have

$$\begin{aligned} &E\big(W_{n+1} \mid (Ch_1, ..., Ch_n) = \vec{v}\big) = \\ &\big(1 + \sum_i \vec{v}_i\big) P\big(Ch_{n+1} = +1 \mid (Ch_1, ..., Ch_n) = \vec{v}\big) \\ &+\big(-1 + \sum_i \vec{v}_i\big) P\big(Ch_{n+1} = -1 \mid (Ch_1, ..., Ch_n) = \vec{v}\big) \\ &+\big(-2 + \sum_i \vec{v}_i\big) P\big(Ch_{n+1} = -2 \mid (Ch_1, ..., Ch_n) = \vec{v}\big), \end{aligned} \tag{3.47}$$

where $\sum \vec{v}_i$ denotes the sum of the components of $\vec{v}$; that is, the value of W_n. Now, rewrite (3.47) in the following manner,

$$\begin{aligned} E\big(W_{n+1} \mid (Ch_1, ..., Ch_n) = \vec{v}\big) = \sum_i \vec{v}_i + \Big[&P\big(Ch_{n+1} = +1 \mid (Ch_1, ..., Ch_n) = \vec{v}\big) \\ &- P\big(Ch_{n+1} = -1 \mid (Ch_1, ..., Ch_n) = \vec{v}\big) - 2P\big(Ch_{n+1} = -2 \mid (Ch_1, ..., Ch_n) = \vec{v}\big)\Big] \end{aligned}$$

and (3.45) tells us that the bracketed expression above is less than or equal to 0, so we have

$$E\big(W_{n+1} \mid (Ch_1, ..., Ch_n) = \vec{v}\big) \leq \sum_i \vec{v}_i = W_n.$$

Returning to (3.46), recall that $E(W_n \mid \mathcal{G}_n)$ is a random variable. Since $\mathcal{G}_n$ is generated by a finite collection of events – that is, all possible outcomes of $Ch_1, ..., Ch_n$ – $\mathcal{G}_n$ will consist of a finite collection of atomic sets, and $E(W_n \mid \mathcal{G}_n)$ will be a simple function taking on the value $E\big(W_{n+1} \mid (Ch_1, ..., Ch_n) = \vec{v}\big)$ on the set $\{(Ch_1, ..., Ch_n) = \vec{v}\}$ for any given $\vec{v}$. As we have shown, for every possible outcome of $(Ch_1, ..., Ch_n)$, $E\big(W_{n+1} \mid (Ch_1, ..., Ch_n) = \vec{v}\big)$ is less than or equal to W_n. So, (3.46) is proven. □

Martingale theory is a well-studied field of mathematics. Knowing that $\{W_i\}_{i\in\mathbb{N}^+}$ is a supermartingale allows us to appeal to results from this field. The following theorem is particularly useful.

Theorem 3.6.1. *(Azuma-Hoeffding Inequality) Suppose $\{W_i\}_{i\in\mathbb{N}^+\cup\{0\}}$ is a martingale or supermartingale, and for all $i \in \mathbb{N}^+$,*

$$|W_i - W_{i-1}| \leq c_i \quad a.s.,$$

where $\{c_i\}_{i\in\mathbb{N}^+}$ is a sequence of non-negative constants. Then for any fixed positive integer n and any positive real t,

$$P(W_n - W_0 \geq t) \leq \exp\left(\frac{-t^2}{2\sum_{i=1}^n c_i^2}\right). \tag{3.48}$$

Proof. See, for example, Grimmett and Strizaker [28]. □

If we take the sequence $\{W_i\}_{i\in\mathbb{N}^+}$ that we have defined and supplement it with the random variable $W_0 := 0$, we can apply the Azuma-Hoeffding inequality to $\{W_i\}_{i\in\mathbb{N}^+\cup\{0\}}$. The c_i constants can all be taken to be 2, as $|W_i - W_{i-1}| = |Ch_i|$, and $|Ch_i|$ takes values in $\{1, 2\}$. We thereby obtain

$$P(W_n \geq \delta n) \leq e^{-\delta^2 n/8}, \tag{3.49}$$

where δ is any fixed positive real number. Equation (3.49) can be used to distinguish quantum behavior of W_n from local predictions. Under Quantum Mechanics, $E(W_n) = cn$ where c is a positive constant. No matter how small c is, there will be a sufficiently large n for which (3.49) will indicate that under locality, W_n cannot take on values near (or exceeding) cn, except with

vanishingly small probability. This can be phrased more precisely, employing p-value language. Suppose that n trials of an experiment have been run and the observed value of W_n is m, with $m \simeq cn$. Then we have

$$\text{p-value} = \sup P\left(W_n \geq m \mid H_0 \text{ (locality) }\right) \simeq P\left(W_n \geq cn \mid H_0\right) \leq e^{-c^2 n/8},$$

and $e^{-c^2 n/8}$ is asymptotically zero as $n \to \infty$.

3.7 Optimal Analysis of Martingale Statistics

As demonstrated in the previous section, we have a statistic W_n that can be used to distinguish Quantum Mechanics from local theories, via an application of the Azuma-Hoeffding inequality. However, this inequality is not tight. This means that there might be better bounds on the probability that would give better (i.e., smaller) p-values for the same experimental outcomes. This would lessen the number of experimental trials needed to witness statistical significance, which could have real value in an experiment with high per-trial costs (e.g., an experiment with expensive components that wear out quickly). In this section, we introduce a new method for calculating *exact* p-values.

We first note that (3.48) is not the best known bound for our situation. A recent paper [27] discussing memory effects in other Bell scenarios (not CH74-style experiments) noted that tighter Azuma-Hoeffding-style bounds can be obtained from Theorem 6.1 in [29]. This can be applied to our situation to yield

$$P(W_n \geq \delta n) \leq \left[\left(\frac{2}{2+\delta}\right)^{\frac{2+\delta}{3}} \left(\frac{1}{1-\delta}\right)^{\frac{1-\delta}{3}}\right]^n, \tag{3.50}$$

which can outperform (3.48); i.e., the upper bound is smaller. Equation (3.50) is the standard for comparison that we will use when using the method for computing exact p-values, which we now describe.

W_n is a random walk on the integers starting at zero, with the kth step given by Ch_k, resulting in a move either up 1, down 1, or down 2. The "goal" for the local theory is to finish at or above a specified cut-point $K = n\delta$ after n steps, and the p-value is the probability of attaining this goal. The trick for determining the best local strategy for finishing at or above K is to trace backwards from the end.

We start by giving an informal description of the procedure, and then we will follow with a formal proof that the procedure is valid. First, note the following two distributions for Ch_k that

saturate (3.45):

$$P(Ch_k = +1) = \frac{1}{2} \quad \text{and} \quad P(Ch_k = -1) = \frac{1}{2}, \tag{3.51}$$

$$P(Ch_k = +1) = \frac{2}{3} \quad \text{and} \quad P(Ch_k = -2) = \frac{1}{3}, \tag{3.52}$$

There are other distributions that saturate the inequality, but these two turn out to be the most important. The "goal" for the local theory is to end up with $W_n \geq K$, and we will soon see that at any point in the random walk, one of (3.51) or (3.52) will always be the optimal local distribution for Ch_k to achieve this goal. Accepting this claim without proof for now, one can see in Table 3.2 how to calculate the eventual probability of success for various locations of the random walk, by starting at the end and tracing backwards.

Table 3.2: Optimal Strategies and Probabilities of Eventual Success at Various Locations

Location	$n-2$ step		$n-1$ step		n step	
of W_n	O.S.	$P(S)$	O.S.	$P(S)$	O.S.	$P(S)$
$K+3$	any	1	any	1	-	1
$K+2$	(3.51)	1	any	1	-	1
$K+1$	(3.52)	8/9	(3.51)	1	-	1
K	(3.51)	5/6	(3.52)	2/3	-	1
$K-1$	(3.52)	4/9	(3.52)	2/3	-	0
$K-2$	(3.52)	4/9	any	0	-	0
$K-3$	any	0	any	0	-	0

To understand Table 3.2, recognize that at the final, mth step, the random walk W_n has "succeeded" if it is at or above K, and "failed" otherwise. Hence the probability of eventual success, denoted $P(S)$, is either 1 or 0, respectively. Now, looking to the previous $(n-1)$ step of the random walk, the two potential optimal strategies (O.S.) (3.51) and (3.52) can be compared to see which maximizes the chance of $W_n \geq K$. For instance, we see the random walk that finds itself at location $K+1$ at the penultimate step should opt for the 50-50 strategy (3.51) that puts equal weight on moving up 1 and down 1, because this strategy ensures eventual success. On the other hand, the random walk that finds itself at location K will want to select strategy (3.52), which gives a best-possible $\frac{2}{3}$ chance of ending at/above K. After the entire $m-1$ column is thus filled, the process can be repeated for step $m-2$, then $m-3$, etc. Table 3.2 only calculates the first few iterates of this process, but a computer can quickly extrapolate back to the first step, and the probability of eventual success for the random walk at location 0 on step 0 will be the *exact* p-value.

We can already notice something interesting just from the first few steps of the process in Table 3.2: there is something to be gained by employing different strategies at different steps of the process (i.e., exploiting the memory loophole). This is in contrast to the situation with the binary variable J, where a local variable cannot do any better than a particular i.i.d. strategy.

Now, we provide a formal proof for the validity of the algorithm. The following notation will help: we use $\mathcal{CH}_a^m$ to refer to any particular event, consisting of the outcomes of the variables $(Ch_1, ..., Ch_m)$, for which $\sum_{i=1}^m Ch_i = a$.

Proposition 3.7.1. *Fix a positive integer n, any nonnegative integer $K < n$ (the "cut point for success"), and a positive integer $m \leq n$ (the "current step"). Suppose that there exist constants c_a, $a \in \{-2m, -2m+1, ..., m-1, m\}$, for which $P(W_n \geq 0 \mid \mathcal{CH}_a^m) \leq c_a$ always holds (for any choice of $\mathcal{CH}_a^m$) and the c_a are monotone in a: $c_{-2m} \leq c_{-2m+1} \leq ... \leq c_m$.*

Then "the same is true for the previous step": there exist constants d_a, with $a \in \{-2(m-1), -2(m-1)+1, ..., m-1\}$, for which $P(W_n \geq 0 \mid \mathcal{CH}_a^{m-1}) \leq d_a$, and the d_a are monotone in a.

Furthermore, for all $a \in \{-2(m-1), -2(m-1)+1, ..., m-1\}$, we can take d_a to be equal to $\max\{\frac{c_{a+1}}{2} + \frac{c_{a-1}}{2}, \frac{2c_{a+1}}{3} + \frac{c_{a-2}}{3}\}$.

Proof. We condition on the outcome of Ch_m to obtain:

$$\begin{aligned} P\left(W_n \geq 0 \mid \mathcal{CH}_a^{m-1}\right) &= P\left(W_n \geq 0 \mid \mathcal{CH}_a^{m-1} \cap \{Ch_m = +1\}\right) P\left(Ch_m = +1 \mid \mathcal{CH}_a^{m-1}\right) \\ &+ P\left(W_n \geq 0 \mid \mathcal{CH}_a^{m-1} \cap \{Ch_m = -1\}\right) P\left(Ch_m = -1 \mid \mathcal{CH}_a^{m-1}\right) \\ &+ P\left(W_n \geq 0 \mid \mathcal{CH}_a^{m-1} \cap \{Ch_m = -2\}\right) P\left(Ch_m = -2 \mid \mathcal{CH}_a^{m-1}\right). \end{aligned}$$

Note that $\mathcal{CH}_a^{m-1} \cap \{Ch_m = -1\}$ is an event of the form $\mathcal{CH}_{a+1}^m$, and a similar relation holds for two other terms, so we can say

$$\begin{aligned} P\left(W_n \geq 0 \mid \mathcal{CH}_a^{m-1}\right) \leq c_{a+1} P\left(Ch_m = +1 \mid \mathcal{CH}_a^{m-1}\right) \\ + c_{a-1} P\left(Ch_m = -1 \mid \mathcal{CH}_a^{m-1}\right) + c_{a-2} P\left(Ch_m = -2 \mid \mathcal{CH}_a^{m-1}\right). \end{aligned} \quad (3.53)$$

For readability, we re-write the right side of equation (3.53) as follows:

$$c_{a+1}x + c_{a-1}y + c_{a-2}z. \quad (3.54)$$

Note that by the laws of probability, $x, y, z \in [0,1]$ and $x+y+z=1$, and by (3.45), $x \leq y+2z$. Of interest is the maximum possible value of (3.54). First, we assert that it suffices to maximize this expression for values x, y, and z for which $x = y+2z$. To prove the assertion, suppose inequality (3.45) were strict; i.e., $x+\epsilon = y+2z$, $1 > \epsilon > 0$. We will show that there are values $x_0, y_0, z_0 \in [0,1]$ for which $x_0+y_0+z_0 = 1$ and $x_0 = y_0+2z_0$, and

$$c_{a+1}x + c_{a-1}y + c_{a-2}z \leq c_{a+1}x_0 + c_{a-1}y_0 + c_{a-2}z_0. \tag{3.55}$$

Case 1: $\frac{\epsilon}{2} \leq y$.

First, note that $\frac{\epsilon}{2} \leq 1-x$: recalling that $x+\epsilon = y+2z$, we have

$$\frac{\epsilon}{2} = \frac{y}{2} + z - \frac{x}{2} = \frac{1}{2}(x+y+z) + \frac{z}{2} - x = \frac{1}{2} + \frac{z}{2} - x \leq 1-x.$$

Now, define $x_0 = x+\frac{\epsilon}{2}$, $y_0 = y-\frac{\epsilon}{2}$, and $z_0 = z$. Because $\frac{\epsilon}{2} \leq y$ and $\frac{\epsilon}{2} \leq x$, we have $x_0, y_0, z_0 \in [0,1]$, and the equations $x_0+y_0+z_0 = 1$ and $x_0 = y_0+2z_0$ clearly hold. Then because $c_{a+1} \geq c_{a-1}$, (3.55) holds.

Case 2: $\frac{\epsilon}{2} > y$.

Note that $\frac{\epsilon}{2} \leq 1-x$ still holds (the proof did not depend on the case assumption) so we have $y < 1-x$. Hence we can define an intermediate distribution $x_0^* = x+y$, $y_0^* = y-y = 0$, and $z_0^* = z$, for which $x_0^*, y_0^*, z_0^* \in [0,1]$ and $x_0^*+y_0^*+z_0^* = 1$, and $c_{a+1} \geq c_{a-1}$ implies

$$c_{a+1}x + c_{a-1}y + c_{a-2}z \leq c_{a+1}x_0^* + c_{a-1}y_0^* + c_{a-2}z_0^*.$$

Now, this isn't quite yet what we want, because we still have $x_0^*+\epsilon^* = y_0^*+2z_0^*$, where $0 < \epsilon^* = \epsilon-2y$. To get our desired distribution, note that because $y_0^* = 0$, $z_0^* = 1-x_0^*$, we have

$$x_0^* < x_0^* + \epsilon^* = 2 - 2x_0^*,$$

which implies that $x_0^* < \frac{2}{3}$ (and consequently $z_0^* > \frac{1}{3}$). Thus if we take $x_0 = \frac{2}{2}$, $y_0 = 0$, and $z_0 = \frac{1}{3}$, the fact that $c_{a+1} \geq c_{a-2}$ implies

$$c_{a+1}x_0^* + c_{a-1}y_0^* + c_{a-2}z_0^* \leq c_{a+1}x_0 + c_{a-1}y_0 + c_{a-2}z_0, \tag{3.56}$$

and thus (3.55) holds.

We have demonstrated that if we want to maximize (3.54), it will suffice to consider $x, y, z \in [0,1]$ for which $x + y + z = 1$ and $x = y + 2z$. Using algebraic manipulations, it is found that this set consists of the parameterized family

$$\begin{aligned} x &= \frac{1}{2} + \frac{t}{6} \\ y &= \frac{1}{2} - \frac{t}{2} \\ z &= \frac{t}{3}, \end{aligned}$$

where t takes a value in the set $[0,1]$. Then if we look back at the expression (3.54), we see that it is linear in t and therefore will take on its maximum value at one of the endpoints; either $t = 0$ or $t = 1$ (or possibly both). Note that these two cases correspond to the distributions (3.51) and (3.52). Plugging in zero and one for t, the maximum value of (3.54) will be equal to $\max\{\frac{c_{a+1}}{2} + \frac{c_{a-1}}{2}, \frac{2c_{a+1}}{3} + \frac{c_{a-2}}{3}\}$. If we take d_a to be this value, we thus have $P(W_n \geq 0 \mid \mathcal{CH}_a^{m-1}) \leq d_a$.

Finally, the d_a are monotone, as the monotonicity of the c_a allows us to write

$$d_{a+1} = \max\left\{\frac{c_{a+2}}{2} + \frac{c_a}{2}, \frac{2c_a}{3} + \frac{c_{a-1}}{3}\right\} \geq \max\left\{\frac{c_{a+1}}{2} + \frac{c_{a-1}}{2}, \frac{2c_{a+1}}{3} + \frac{c_{a-2}}{3}\right\} = d_a.$$

□

Proposition 3.7.1 gives a theoretical foundation for implementing the algorithm outlined in Table 3.2: the c_a constants for the first step, $m = n$, are all 1 for $a \geq K$ and 0 for $a < K$, and this sequence readily satisfies the conditions of the proposition. After repeated application of the algorithm of Proposition 3.7.1, the last step is moving from $m = 1$ to $m - 1 = 0$. For this step, the degenerate event "$\mathcal{CH}_0^0$" is just Ω, and we are left with a straightforward bound:

$$P(W_n \geq K) \leq d_0,$$

where d_0 is obtained from the algorithm.

Importantly, this bound is *tight*. Looking back at the proof of Proposition 3.7.1, the equation (3.54) can be maximized when either

$$P\left(Ch_m = +1 \mid \mathcal{CH}_a^{m-1}\right) = P\left(Ch_m = -1 \mid \mathcal{CH}_a^{m-1}\right) = \frac{1}{2} \tag{3.57}$$

or

$$P\left(Ch_m = +1 \mid \mathcal{CH}_a^{m-1}\right) = 2P\left(Ch_m = -2 \mid \mathcal{CH}_a^{m-1}\right) = \frac{2}{3}. \tag{3.58}$$

Both of these distributions can be achieved by a local hidden variable. For (3.57), this can be achieved with a constant local state variable λ_i for which

$$P(+a \mid \lambda_i) = P(0a' \mid \lambda_i) = 1 \quad P(0b \mid \lambda_i) = P(+b' \mid \lambda_i) = 1,$$

and for (3.58), this can be achieved with a constant local state variable λ_i for which

$$P(+a \mid \lambda_i) = P(+a' \mid \lambda_i) = 1 \quad P(+b \mid \lambda_i) = P(+b' \mid \lambda_i) = 1.$$

In such models, joint probabilities of events like $P(++ab \mid \lambda_i)$ are given by the product of $P(+a \mid \lambda_i)$ and $P(+b \mid \lambda_i)$ so that (3.7) is satisfied, and (3.5) – the only other assumption referring to λ_i – is satisfied by virtue of λ_i being a constant.

One can simulate a CH-style experiment according to the data of [23] or [24], by using the empirical distributions in Table 3.1 to generate simulated data. The quantum predictions after $100,000$ non-00 trial outcomes are given in Table 3.3, which demonstrates that the back-tracing method of Proposition 3.7.1 yields meaningful improvements over the previously-known bounds (3.48) and (3.50).

Table 3.3: Comparison of exact p-values to previously known upper bounds for simulated experiment of 100,000 trials (only a subset of the trials contribute to W_n)

	Results		p-values		
	W_n	n	Exact	(3.50) Bound	(3.48) Bound
Christensen *et al.* [24]	447	19,359	.0136	.0750	.5636
Giustina *et al.* [23]	1,135	20,395	9.90×10^{-9}	1.14×10^{-7}	.0299

3.8 Changing the Measurement Setting Probabilities

In this section, we explore the effects of changing the measurement setting probabilities. At a minimum, one should at least consider small deviations from equiprobable setting probabilities, in order to make the experimental model robust. This is desirable because the statement "the probability that the setting will be a is *exactly* $\frac{1}{2}$" is empirically problematic: how can such an

assertion be made about a real-world process? We may be sure that it is $\frac{1}{2}$ with a very high degree of precision, but what we really mean by this is "the probability that the setting will be a is within some ϵ of $\frac{1}{2}$." We start by examining this situation. Later in the section, we will also discuss larger deviations from equiprobability.

For our examination of epsilon tolerance, we work with the J statistic defined earlier, in (3.36). Under our earlier assumption that $p_a = p_b = \frac{1}{2}$, we were able to show that $E(J_i) \leq 0$. However, we now have a weaker assumption: we assume that there is a fixed $\epsilon > 0$ for which

$$\begin{aligned} \tfrac{1}{2} - \epsilon \leq p_a \leq \tfrac{1}{2} + \epsilon \\ \tfrac{1}{2} - \epsilon \leq p_b \leq \tfrac{1}{2} + \epsilon. \end{aligned} \tag{3.59}$$

Our goal is a constraint of the form $E(J_i) \leq f(\epsilon)$, where ideally $f(\epsilon)$ is still a small number. It will help to define an intermediate statistic J_k^*. For $V = \{++ab, +0ab', 0{+}a'b, ++a'b'\}$, define J^* as follows:

$$J_k^* = \begin{cases} +1 & \text{if the } k\text{th outcome of } T_i \text{ taking a value in } V \text{ is } ++ab \\ -1 & \text{if the } k\text{th outcome of } T_i \text{ taking a value in } V \text{ is not } ++ab \\ 0 & \text{if fewer than } k \text{ outcomes of } T_i \text{ are in the set } V \end{cases}$$

The following lemma gives a constraint on J^*.

Proposition 3.8.1. *Under* (3.3)-(3.7) *and* (3.59)*,*

$$P(J_k^* = +1 \mid \mathcal{J}^*) - P(J_k^* = -1 \mid \mathcal{J}^*) \leq \frac{4\epsilon}{1 + 4\epsilon^2} P(J_k^* \neq 0 \mid \mathcal{J}^*), \tag{3.60}$$

where $\mathcal{J}^*$ is any fixed event of the form $(J_1^*, ..., J_{k-1}^*) = \vec{v}, \vec{v} \in \{-1, +1\}^{k-1}$ for which $P(\mathcal{J}^*) > 0$.

Proof. We have

$$\begin{aligned} &P(J_k = +1 \mid \mathcal{J}^*) - P(J_k = -1 \mid \mathcal{J}^*) \\ &\qquad = \frac{1}{P(\mathcal{J}^*)} \left\{ P\left(J_k^* = +1 \cap \mathcal{J}^*\right) - P\left(J_k^* = -1 \cap \mathcal{J}^*\right) \right\} \\ &\qquad\qquad = \frac{1}{P(\mathcal{J}^*)} \left\{ \sum_{\mathcal{T}} \left[P\left(J_k^* = +1 \cap \mathcal{T}\right) - P\left(J_k^* = -1 \cap \mathcal{T}\right) \right] \right\}, \end{aligned} \tag{3.61}$$

where $\mathcal{T}$ is an event consisting of the first m outcomes of T_i (m may vary for different choices of $\mathcal{T}$),

of the following form:

$$\{(T_1, ..., T_{m-1}) = \vec{v}\} \cup \{T_m \in V\}, \tag{3.62}$$

where $\vec{v}$ is a fixed outcome vector (dependent on choice of $\mathcal{T}$) containing exactly $k-1$ outcomes in V, consistent with the occurrence of the event $\mathcal{J}^*$. We will use $\mathcal{T}_{m-1} \cup \{T_m \in V\}$ as a shorthand for the expression (3.62). Now, the outcome of J_k^* is based on whether T_m is in $V^+ := \{++ab\}$ or $V^- := V \backslash V^+$, so the expression $[P(J_k^* = +1 \cap \mathcal{T}) - P(J_k^* = -1 \cap \mathcal{T})]$ in (3.61) can be rewritten as

$$\begin{aligned} P(\mathcal{T}_{m-1} \cap \{T_m \in V^+\}) &- P(\mathcal{T}_{m-1} \cap \{T_m \in V^-\}) \\ &= \int P(\mathcal{T}_{m-1} \cap \{T_m \in V^+\} \mid \lambda_m) - P(\mathcal{T}_{m-1} \cap \{T_m \in V^-\} \mid \lambda_m) dP \\ &= \int_{\mathcal{T}_{m-1}} P(T_m \in V^+ \mid \lambda_m) - P(T_m \in V^- \mid \lambda_m) dP, \end{aligned} \tag{3.63}$$

where we used the fact that $\sigma(\lambda_m) \perp\!\!\!\perp \mathcal{T}_{m-1}$ to obtain the second equality by way of Theorem 9.1.3 in [10].

Now note that if $P(T_m \in V \mid \lambda_m) = 0$, then $P(T_m \in V^+ \mid \lambda_m) = P(T_m \in V^- \mid \lambda_m) = 0$ almost surely. Thus the integration set can be changed from $\mathcal{T}_{m-1}$ to a set $\mathcal{S}$, which we define as $\mathcal{S} := \mathcal{T}_{m-1} \cap \{P(T_m \in V \mid \lambda_m) > 0\}$, without changing the value of expression (3.63). Doing this allows us to multiply the integrand by quantity $1 = P(T_m \in V \mid \lambda_m)/P(T_m \in V \mid \lambda_m)$ without worrying that this could be $0/0$, so everything that we have derived so far implies that

$$\begin{aligned} &P(J_k^* = +1 \mid \mathcal{J}^*) - P(J_k^* = -1 \mid \mathcal{J}^*) = \\ &\qquad \frac{1}{P(\mathcal{J}^*)} \sum_{\mathcal{T}} \left\{ \int_{\mathcal{S}} \left[\frac{P(T_m \in V^+ \mid \lambda_m) - P(T_m \in V^- \mid \lambda_m)}{P(T_m \in V \mid \lambda_m)} \right] P(T_m \in V \mid \lambda_m) dP \right\}. \end{aligned} \tag{3.64}$$

The strategy for the rest of the proof is to first provide an upper bound for the expression in brackets in (3.64), and once this is done, perform the integration followed by the summation.

First recall that $P(T_m \in V^+ \mid \lambda_m) = P(++ab \mid \lambda_m)$ and $P(T_m \in V^- \mid \lambda_m)$ is equal to the sum $P(+0ab' \mid \lambda_m) + P(0+a'b \mid \lambda_m) + P(++a'b' \mid \lambda_m)$. The locality assumption (3.7) can be applied to break these events into $P(+a \mid \lambda_m)P(+b \mid \lambda_m)$, etc. Then using the μ notation and the fact that, for instance, $\mu_a(+ \mid \lambda) = \frac{P(+a|\lambda_m)}{p_a}$, the expression in brackets in (3.63) can be re-writtten as

$$\frac{\begin{array}{c} p_a p_b \mu_a(+|\lambda_m)\mu_b(+|\lambda_m) - p_a q_b \mu_a(+|\lambda_m)\mu_{b'}(0|\lambda_m) \\ -q_a p_b \mu_{a'}(0|\lambda_m)\mu_b(+|\lambda_m) - q_a q_b \mu_{a'}(+|\lambda_m)\mu_{b'}(+|\lambda_m) \end{array}}{\begin{array}{c} p_a p_b \mu_a(+|\lambda_m)\mu_b(+|\lambda_m) + p_a q_b \mu_a(+|\lambda_m)\mu_{b'}(0|\lambda_m) \\ +q_a p_b \mu_{a'}(0|\lambda_m)\mu_b(+|\lambda_m) + q_a q_b \mu_{a'}(+|\lambda_m)\mu_{b'}(+|\lambda_m) \end{array}} \tag{3.65}$$

and Lemma 2.5.2 applies to the μ terms above, so each μ must take a value between 0 and 1. For readability, we re-write (3.65) in the following form:

$$\frac{(p_a p_b)ab - (p_a q_b)ac - (q_a p_b)bd - (q_a q_b)(1-c)(1-d)}{(p_a p_b)ab + (p_a q_b)ac + (q_a p_b)bd + (q_a q_b)(1-c)(1-d)} \tag{3.66}$$

The task is now to find the maximum possible value for (3.66) subject to the constraints (3.59) and $a, b, c, d \in [0, 1]$. Note that a few particular values could result in the denominator of (3.66) being zero; however, any of those values would imply that $P(T_m \neq 0 \mid \lambda_m) = 0$ for that particular point in Ω, which we know does not occur on our set of integration $S \subseteq \Omega$. Therefore we only need to worry about possible values for which the expression (3.66) is defined.

It is possible to show that (3.66) is bounded above by $\frac{4\epsilon}{1+4\epsilon^2}$. This maximization proof is rather lengthy and technical, and so we defer it to Lemma 3.8.1. Once we have this bound, expression (3.64) is bounded above by

$$\begin{aligned}
& \frac{1}{P(\mathcal{J}^*)} \sum_{\mathcal{T}} \left\{ \int_{\mathcal{S}} \left[\frac{4\epsilon}{1+4\epsilon^2} \right] P(T_m \in V \mid \lambda_m) dP \right\} \\
= \; & \frac{1}{P(\mathcal{J}^*)} \sum_{\mathcal{T}} \left\{ \int_{\mathcal{T}_{m-1}} \left[\frac{4\epsilon}{1+4\epsilon^2} \right] P(T_m \in V \mid \lambda_m) dP \right\} \\
= \; & \frac{1}{P(\mathcal{J}^*)} \sum_{\mathcal{T}} \left\{ \int \left[\frac{4\epsilon}{1+4\epsilon^2} \right] P(\{T_m \in V\} \cap \mathcal{T}_{m-1} \mid \lambda_m) dP \right\} \\
= \; & \frac{1}{P(\mathcal{J}^*)} \sum_{\mathcal{T}} \left\{ \frac{4\epsilon}{1+4\epsilon^2} P(\{T_m \in V\} \cap \mathcal{T}_{m-1}) \right\} \\
= \; & \frac{1}{P(\mathcal{J}^*)} \sum_{\mathcal{T}} \left\{ \frac{4\epsilon}{1+4\epsilon^2} P(\mathcal{T}) \right\} \\
= \; & \frac{1}{P(\mathcal{J}^*)} \left\{ \frac{4\epsilon}{1+4\epsilon^2} P\left(\{J_k^* \neq 0\} \cap \mathcal{J}^*\right) \right\} \\
= \; & \frac{4\epsilon}{1+4\epsilon^2} P(J_k^* \neq 0 \mid \mathcal{J}^*).
\end{aligned}$$

□

Lemma 3.8.1. *For any six numbers p_a, p_b, a, b, c, d satisfying* (3.59) *and $a, b, c, d \in [0, 1]$ for which the expression* (3.66) *(where $q_a = 1 - p_a$, $q_b = 1 - p_b$) is defined (the denominator is nonzero), the maximum possible value of* (3.66) *is $\frac{4\epsilon}{1+4\epsilon^2}$.*

Proof. We start by taking p_a and p_b to be fixed. We will show that the maximum value of (3.66) is

then

$$\max\left\{\frac{p_ap_b - q_aq_b}{p_ap_b + q_aq_b}, \frac{p_ap_b - q_ap_b}{p_ap_b + q_ap_b}, \frac{p_ap_b - p_aq_b}{p_ap_b + p_aq_b}, 0\right\}. \tag{3.67}$$

First, we demonstrate that for every interior point $a, b, c, d \in (0, 1)$, there is a boundary point (at least one a, b, c, d is in $\{0, 1\}$) for which the value of (3.66) is at least as large. To see this, consider $\frac{\partial}{\partial a}$ of (3.66), which is

$$\frac{\partial}{\partial a}(3.66) = \frac{2(p_ap_b)b\left[(q_ap_b)bd + (q_aq_b)(1-c)(1-d)\right]}{\left[(p_ap_b)ab + (p_aq_b)ac + (q_ap_b)bd + (q_aq_b)(1-c)(1-d)\right]^2}. \tag{3.68}$$

By the assumption that all of the variables a, b, c, d are in $(0, 1)$, expression (3.68) must be positive. Hence if we fix b, c, d and vary a, the expression (3.66) is strictly increasing as $a \to 1$. Hence for any interior point, we get a larger value for (3.66) if we replace the a coordinate with 1.

Hence we can restrict our attention to the boundary cases, of which there are 8.

Case 1: $a = 0$

As we have assumed the expression (3.66) is defined, that means that the denominator is nonzero and therefore positive, and the numerator is nonpositive if $a = 0$, so expression (3.66) is bounded above by 0.

Case 2: $a = 1$

In this case, the expression (3.66) reduces to

$$\frac{(p_ap_b)b - (p_aq_b)c - (q_ap_b)bd - (q_aq_b)(1-c)(1-d)}{(p_ap_b)b + (p_aq_b)c + (q_ap_b)bd + (q_aq_b)(1-c)(1-d)}. \tag{3.69}$$

If $b = 0$, this expression is bounded above by 0 (when it is defined). For the case when $b \neq 0$, consider the partial derivative of (3.69) with respect to d:

$$\frac{\partial}{\partial d}(3.69) = \frac{2(p_ap_b)b\left[q_aq_b(1-c) - (q_ap_b)b\right]}{\left[(p_ap_b)b + (p_aq_b)c + (q_ap_b)bd + (q_aq_b)(1-c)(1-d)\right]^2}. \tag{3.70}$$

For any fixed choice of b and c, $b \neq 0$, the sign of (3.70) does not change as d varies. Hence we can get the largest possible value for the expression (3.69) by replacing d with either 0 or 1. We examine these two possibilities:

Subcase 2.1: $a = 1$, $b \neq 0$, $d = 0$

In this scenario, the expression is

$$\frac{(p_a p_b)b - (p_a q_b)c - (q_a q_b)(1-c)}{(p_a p_b)b + (p_a q_b)c + (q_a q_b)(1-c)}, \tag{3.71}$$

and the partial derivative of (3.71) with respect to b is strictly positive, so for fixed c, this expression is largest when $b = 1$. Thus the expression simplifies to

$$\frac{(p_a p_b) - (p_a q_b)c - (q_a q_b)(1-c)}{(p_a p_b) + (p_a q_b)c + (q_a q_b)(1-c)}, \tag{3.72}$$

and the partial derivative of (3.72) with respect to c is

$$\frac{p_a p_b(q_a q_b - p_q p_b)}{\left[(p_a p_b) + (p_a q_b)c + (q_a q_b)(1-c)\right]^2},$$

which is either uniformly positive, uniformly negative, or uniformly 0, as c varies. Hence the maximum possible value of (3.72) is achieved either when $c = 0$ or when $c = 1$, and is thus

$$\max\left\{\frac{p_a p_b - q_a q_b}{p_a p_b + q_a q_b}, \frac{p_a p_b - p_a q_b}{p_a p_b + p_a q_b}\right\}.$$

Subcase 2.2: $a = 1$, $b \neq 0$, $d = 1$

Now the expression is

$$\frac{(p_a p_b)b - (p_a q_b)c - (q_a p_b)b}{(p_a p_b)b + (p_a q_b)c + (q_a p_b)b},$$

which is clearly maximized as $c \to 0$, thus taking on a maximum possible value of $\frac{p_a p_b - q_a p_b}{p_a p_b + q_a p_b}$.

Case 3: $d = 0$

(3.66) reduces to

$$\frac{(p_a p_b)ab - (p_a q_b)ac - (q_a q_b)(1-c)}{(p_a p_b)ab + (p_a q_b)ac + (q_a q_b)(1-c)}. \tag{3.73}$$

If $a = 0$, all (defined) values of this expression are nonpositive, and therefore it is bounded above by 0. If $a \neq 0$, then the partial derivative of (3.73) with respect to b is strictly positive, so we can replace b with 1 and we are left with

$$\frac{(p_a p_b)a - (p_a q_b)ac - (q_a q_b)(1-c)}{(p_a p_b)a + (p_a q_b)ac + (q_a q_b)(1-c)},$$

and the partial derivative of the above expression with respect to c is either uniformly positive,

uniformly negative, or uniformly zero as c varies, so the maximum is assumed either when $c = 0$ or $c = 1$. If $c = 0$, then the expression is $\frac{(p_a p_b)a-(q_a q_b)}{(p_a p_b)a+(q_a q_b)}$ which assumes a maximum possible value of $\frac{p_a p_b - q_a q_b}{p_a p_b + q_a q_b}$.

If $c = 1$, the term a cancels and the maximum possible value is $\frac{p_a p_b - p_a q_b}{p_a p_b + p_a q_b}$.

Case 4: $d = 1$

(3.66) reduces to

$$\frac{(p_a p_b)ab - (p_a q_b)ac - (q_a p_b)b}{(p_a p_b)ab + (p_a q_b)ac + (q_a p_b)b} \leq \frac{(p_a p_b)ab - (q_a p_b)b}{(p_a p_b)ab + (q_a p_b)b} = \frac{(p_a p_b)a - (q_a p_b)}{(p_a p_b)a + (q_a p_b)} \leq \frac{p_a p_b - q_a p_b}{p_a p_b + q_a p_b}.$$

Case 5, 6, 7, 8: $b = 0$, $b = 1$, $c = 0$, $c = 1$

These cases follow by symmetry. The expression (3.66) is unchanged if we switch a and b, c and d, p_a and p_b, and q_a and q_b. Hence the arguments that we employed for a and d will apply for b and c.

We have now shown that every case results in a quantity that is bounded by the expression (3.67). All that remains is to determine the largest possible value of (3.67). We have

$$\frac{p_a p_b - q_a p_b}{p_a p_b + q_a p_b} = \frac{p_a - q_a}{p_a + q_a} = \frac{2p_a - 1}{1} \leq 2\left(\frac{1}{2} + \epsilon\right) - 1 = 2\epsilon,$$

and so similarly, $\frac{p_a p_b - p_a q_b}{p_a p_b + p_a q_b} \leq 2\epsilon$. The final remaining nonzero term in (3.67) can be re-written

$$\frac{p_a p_b - (1 - p_a)(1 - p_b)}{p_a p_b + (1 - p_a)(1 - p_b)}. \tag{3.74}$$

To maximize this, first note that the partial derivative of the above expression with respect to p_a is

$$\frac{2p_b(1 - p_b)}{\left[p_a p_b + (1 - p_a)(1 - p_b)\right]^2}. \tag{3.75}$$

Recall that $p_a, p_b \in [\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon]$ with $\epsilon \in (0, \frac{1}{2})$, which implies that (3.75) is uniformly positive. Hence (3.74) reaches its maximum when p_a is largest. By symmetry, the same reasoning applies to p_b, so (3.74) is maximized when $p_a = p_b = \epsilon + \frac{1}{2}$. The value of the expression at this point is the now-familiar $\frac{4\epsilon}{1+4\epsilon^2}$. As this quantity exceeds 2ϵ for all values of $\epsilon \in (0, \frac{1}{2})$, this is the largest possible value for (3.66). □

Note that the bound $\frac{4\epsilon}{1+4\epsilon^2}$ is tight. It can be achieved with a constant local state variable

λ_i for which

$$P(+a \mid \lambda_i) = P(+b \mid \lambda_i) = P(+a' \mid \lambda_i) = P(+b' \mid \lambda_i) = 1.$$

For simplicity, just consider the first J^*-type variable J_1^*, which will take a value when the measurement setting first registers ab or $a'b'$. Note that this will eventually occur with probability 1, and so $P(J_1^* \neq 0) = 1$ in this scenario. Now let $\mathcal{T}_i$ be the event that the first occurrence of either configuration ab or $a'b'$ happens on the ith trial. Then we can write

$$\begin{aligned}
P(J_1^* = +1) - P(J_1^* = -1) &= \sum_{i=1}^{\infty} \left[P(J_1^* = +1 \mid \mathcal{T}_i) - P(J_1^* = +1 \mid \mathcal{T}_i)\right] P(\mathcal{T}_i) \\
&= \sum_{i=1}^{\infty} \left[\frac{p_a p_b - q_a q_b}{p_a p_b + q_a q_b}\right] \left(1 - (p_a p_b + q_a q_b)\right)^{i-1} (p_a p_b + q_a q_b) \\
&= \sum_{i=1}^{\infty} \left[\frac{4\epsilon}{1 + 4\epsilon^2}\right] \left(1 - (p_a p_b + q_a q_b)\right)^{i-1} (p_a p_b + q_a q_b) \\
&= \frac{4\epsilon}{1 + 4\epsilon^2} \sum_{i=1}^{\infty} \left(1 - (p_a p_b + q_a q_b)\right)^{i-1} (p_a p_b + q_a q_b) \\
&= \frac{4\epsilon}{1 + 4\epsilon^2},
\end{aligned}$$

which is the saturation of the bound (3.60) with $\mathcal{J}^* = \Omega$, as $P(J_1^* \neq 0) = 1$.

The bound of Proposition 3.8.1 allows the statistical analysis to tolerate small deviations from equiprobability. For instance, the number of trials in each setting configuration in the raw data of the experiment [24] suggest a value for p_a that is close to .5058. With the trials numbering in the millions, one could be highly confident that the true probability was within $\pm.0002$ of this figure. Hence the methods of this section could be used for statistical analysis with $\epsilon = .006$, which will be done following the proof of Lemma 3.8.2. For such an analysis, one can use the binary statistic J_k as defined in Section 3.5 so that all the 0 outcomes of J_k^* are collapsed into the -1 outcome. This aligns with the following definition:

$$J_k = \begin{cases} +1 & \text{if } J_k^* = +1 \\ -1 & \text{if } J_k^* = -1 \\ -1 & \text{if } J_k^* = 0. \end{cases}$$

The following lemma shows that a version of the bound of (3.60) would continue to hold for J_k.

Lemma 3.8.2. *Under the assumptions of Proposition 3.8.1,*

$$P(J_k = +1 \mid \mathcal{J}) - P(J_k = -1 \mid \mathcal{J}) \leq \frac{4\epsilon}{1+4\epsilon^2}, \tag{3.76}$$

where $\mathcal{J}$ is any event of the form $(J_1, ..., J_{k-1}) = \vec{v}$, $\vec{v} \in \{-1, +1\}^{k-1}$ for which $P(\mathcal{J}) > 0$.

Proof. We have

$$P(J_k = +1 \mid \mathcal{J}) - P(J_k = -1 \mid \mathcal{J}) = \frac{1}{P(\mathcal{J})} \left[P(J_k = +1 \cap \mathcal{J}) - P(J_k = -1 \cap \mathcal{J})\right]. \tag{3.77}$$

Now, we can break the event $\mathcal{J}$ into a disjoint union of constituent events as so:

$$\mathcal{J} = \left(\cup_{a \in A} \mathcal{J}_a^{*0}\right) \bigcup \left(\cup_{b \in B} \mathcal{J}_b^*\right),$$

where $\mathcal{J}_a^{*0}$ is a positive-probability event of the form $(J_1, ..., J_{k-1}) = \vec{v}$ where $\vec{v}$ contains a zero, and $\mathcal{J}_b^*$ is a similarly defined except that $\vec{v}$ contains no zeros. Hence we have

$$\begin{aligned}
&[P(J_k = +1 \cap \mathcal{J}) - P(J_k = -1 \cap \mathcal{J})] \\
&= \sum_{a \in A} \left[P(J_k = +1 \cap \mathcal{J}_a^{*0}) - P(J_k = -1 \cap \mathcal{J}_a^{*0})\right] + \sum_{b \in B} \left[P(J_k = +1 \cap \mathcal{J}_b^*) - P(J_k = -1 \cap \mathcal{J}_b^*)\right].
\end{aligned}$$

Now, if a $\mathcal{J}_a^{*0}$ event occurs, then J_k is necessarily -1, so the $\sum_{a \in A}$ term is nonpositive. As for the $\sum_{b \in B}$ term, we have

$$\begin{aligned}
& \sum_{b \in B} \left[P(J_k = +1 \cap \mathcal{J}_b^*) - P(J_k = -1 \cap \mathcal{J}_b^*)\right] \\
= & \sum_{b \in B} \left[P(J_k^* = +1 \cap \mathcal{J}_b^*) - P(J_k^* = -1 \cap \mathcal{J}_b^*) - P(J_k^* = 0 \cap \mathcal{J}_b^*)\right] \\
\leq & \sum_{b \in B} \left[P(J_k^* = +1 \cap \mathcal{J}_b^*) - P(J_k^* = -1 \cap \mathcal{J}_b^*)\right] \\
\leq & \sum_{b \in B} \left[\frac{4\epsilon}{1+4\epsilon^2} P(J_k^* \neq 0 \cap \mathcal{J}_b^*)\right] \\
\leq & \sum_{b \in B} \left[\frac{4\epsilon}{1+4\epsilon^2} P(\mathcal{J}_b^*)\right] \\
\leq & \frac{4\epsilon}{1+4\epsilon^2} P(\mathcal{J}).
\end{aligned}$$

Hence

$$P(J_k = +1 \mid \mathcal{J}) - P(J_k = -1 \mid \mathcal{J}) \leq \frac{1}{P(\mathcal{J})} \left[\frac{4\epsilon}{1+4\epsilon^2} P(\mathcal{J}) \right] \leq \frac{4\epsilon}{1+4\epsilon^2}.$$

□

Lemma 3.8.2 implies

$$P(J_k = +1 \mid \mathcal{J}) \leq \frac{1}{2} + \frac{2\epsilon}{1+4\epsilon^2}, \tag{3.78}$$

and hence Proposition 2.7.2 can be used to upper-bound the probability of the binary variable J_k achieving large numbers of $+1$ counts in proportion to -1 counts. The data of [24] witnessed $34,145$ $+1$ counts and $31,731$ -1 counts. While a distribution obeying (3.78) can consistently generate more $+1$ counts than -1 counts, the p-value for a large $+1$ margin of $2,414$ after $65,876$ total J_k counts would be .00058, which is highly significant.

Lemma 3.8.2 can be used for small deviations from equiprobability. But if the deviation from equiprobability is too large, the bound (3.78) becomes useless. Instead, a test statistic should be chosen that reflects the true setting probabilities. The generalized version of J_k would be

$$J_k^{\text{gen}} = \begin{cases} (p_a p_b)^{-1} & \text{if the } k\text{th } T_i \text{ taking a value in } V \text{ is } ++ab \\ -(p_a q_b)^{-1} & \text{if the } k\text{th } T_i \text{ taking a value in } V \text{ is } +0ab' \\ -(q_a p_b)^{-1} & \text{if the } k\text{th } T_i \text{ taking a value in } V \text{ is } 0{+}a'b \\ -(q_a q_b)^{-1} & \text{if the } k\text{th } T_i \text{ taking a value in } V \text{ is } ++a'b' \\ -(q_a q_b)^{-1} & \text{if fewer than } k \ T_i \text{ take values in } V, \end{cases} \tag{3.79}$$

and (3.8), along with the results of Section 3.5, tell us that $E(J_k^{\text{gen}} \mid \mathcal{J}^{\text{gen}}) \leq 0$. However, we now appreciate the allure of equiprobable measurement settings, which made the three outcomes $\left\{-(p_a q_b)^{-1}, -(q_a p_b)^{-1}, -(q_a q_b)^{-1}\right\}$ collapse into one outcome. In the absence of this coincidence, one cannot use the binary analysis of the style of Proposition 2.7.2. The more-complicated martingale techniques of Sections 3.6 and 3.7 must be used.

While changing the setting probabilities can complicate the statistical analysis, there is a possibile payoff. The minimum detector efficiency required to demonstrate nonlocality cannot be lowered, but a different distribution over the setting probabilities may actually lower the number of trials needed to demonstrate nonlocality. Fully exploring this possibility would be a project in and of itself: as noted in [30], "to find the (joint) setting distribution that optimizes the [statistical]

strength of a nonlocality proof is a highly nontrivial computation," and our current scenario here is no exception to this rule. But if one wanted to explore these possibilities, a good place to start would be with a more general version of (3.50), which would apply to J_k^{gen}, where $b = \frac{1}{p_a p_b}$ and $a = \min\left\{-(p_a q_b)^{-1}, -(q_a p_b)^{-1}, -(q_a q_b)^{-1}\right\}$:

$$P\left(\sum_{n=1}^{m} S_n \geq mt\right) = \left[\left(\frac{a}{a-t}\right)^{\frac{t-a}{b-a}}\left(\frac{b}{b-t}\right)^{\frac{b-t}{b-a}}\right]^m . \tag{3.80}$$

Analyzing the bound (3.80) and the corresponding quantum predictions for different values of p_a, p_b would be a useful starting point for exploring the (possibly beneficial) statistical effects of varied setting probabilities. Of course, to get the exact p-values, a version of the back-tracing method of Section 3.7 would have to be employed, and making the analysis ϵ-tolerant to small deviations in setting probabilities, as was done with the J-statistic in this section, would add an additional layer of complication to the analysis.

3.9 Expressivity of Deterministic Hidden Variable Models

Our work thus far in this chapter has been very general, proving all of the results for a completely general state variable λ. In this section, we relax the assumptions somewhat and explore how the situation might look if only simpler hidden variable theories are allowed, such as finite-λ-models (as were previously discussed in Chapter 2, Section 2.3). We find that a particular class of finite-λ-theories known as the *local deterministic* hidden variable theories are indeed quite expressive: in the equiprobable-measurement-setting scenario that we have examined this chapter, local deterministic theories can generate any probability distribution that arises from a more general λ description. This result indicates that, at least in some settings, working with the (easier, more intuitive) finite-λ-models does not restrict the types of observable behavior that a local hidden variable theory can model.

Recall the constraints derived in Propositions 3.3.1 and 3.3.2. In the equiprobable setting

regime, the $p_a p_b$-style terms can be cancelled out, and these constraints become

$$\begin{aligned}
P\left(++a_ib_i \mid \mathcal{D}\right) - P\left(+0a_ib_i' \mid \mathcal{D}\right) - P\left(0+a_i'b_i \mid \mathcal{D}\right) - P\left(++a_i'b_i' \mid \mathcal{D}\right) &\leq 0 \\
P\left(++a_ib_i' \mid \mathcal{D}\right) - P\left(+0a_ib_i \mid \mathcal{D}\right) - P\left(0+a_i'b_i' \mid \mathcal{D}\right) - P\left(++a_i'b_i \mid \mathcal{D}\right) &\leq 0 \\
P\left(++a_i'b_i \mid \mathcal{D}\right) - P\left(+0a_i'b_i' \mid \mathcal{D}\right) - P\left(0+a_ib_i \mid \mathcal{D}\right) - P\left(++a_ib_i' \mid \mathcal{D}\right) &\leq 0 \\
P\left(++a_i'b_i' \mid \mathcal{D}\right) - P\left(+0a_i'b_i \mid \mathcal{D}\right) - P\left(0+a_ib_i' \mid \mathcal{D}\right) - P\left(++a_ib_i \mid \mathcal{D}\right) &\leq 0,
\end{aligned}$$

$$\begin{aligned}
P(++a_ib_i \mid \mathcal{D}) + P(+0a_ib_i \mid \mathcal{D}) &= P(++a_ib_i' \mid \mathcal{D}) + P(+0a_ib_i' \mid \mathcal{D}) \\
P(++a_i'b_i \mid \mathcal{D}) + P(+0a_i'b_i \mid \mathcal{D}) &= P(++a_i'b_i' \mid \mathcal{D}) + P(+0a_i'b_i' \mid \mathcal{D}) \\
P(++a_ib_i \mid \mathcal{D}) + P(0+a_ib_i \mid \mathcal{D}) &= P(++a_i'b_i \mid \mathcal{D}) + P(0+a_i'b_i \mid \mathcal{D}) \\
P(++a_ib_i' \mid \mathcal{D}) + P(0+a_ib_i' \mid \mathcal{D}) &= P(++a_i'b_i' \mid \mathcal{D}) + P(0+a_i'b_i' \mid \mathcal{D}). \qquad (3.81)
\end{aligned}$$

These constraints are obeyed by every trial of the experiment, which results in one of 16 outcomes (4 setting configurations, 4 possible pairs of outcomes). By Proposition 3.5.1, these constraints can also be applied to the 12-outcome random variable U_j as defined in (3.33). This is true for a completely general measure-theoretic local hidden variable theory.

Now for the rest of the section, we assume that the variables T_i are defined by a *local deterministic* hidden variable theory, and see what effect (if any) this has on the constraints (3.81). Since we are restricting the class of available local theories, one might think that we might get more restrictive constraints than those of (3.81). Interestingly, we will see that this is not the case.

The intuition for the notion of a local deterministic hidden variable theory is that the value that λ assumes is essentially an instruction set that determines the value of D_1 and D_2, conditioned only on whatever measurement setting is chosen. We will use the following formal definition.

Definition 3.9.1. A *deterministic local hidden variable* λ is a random variable taking values in the finite set $\{+1, 0\}^4$. A *deterministic local hidden variable theory* is a mathematical model satisfying assumptions (3.3)-(3.7) for which each λ_i is a deterministic local hidden variable, and for $x \in \{+1, 0\}$

and $y \in \{a, a', b, b'\}$,

$$P(xy \mid \lambda_i) = \begin{cases} 1 & \text{if } y = a \text{ and the first component of } \lambda_i \text{ is } x, \\ 1 & \text{if } y = a' \text{ and the second component of } \lambda_i \text{ is } x, \\ 1 & \text{if } y = b \text{ and the third component of } \lambda_i \text{ is } x, \\ 1 & \text{if } y = b' \text{ and the fourth component of } \lambda_i \text{ is } x, \\ 0 & \text{otherwise.} \end{cases}$$

For an example to illustrate the above definition, if λ took on the value $(0, +1, 0, +1)$, we would see $D_1 = 0$ if $A = a$, $D_1 = +1$ if $A = a'$, $D_2 = 0$ if $B = b$, $D_2 = +1$ if $B = b'$. Table 3.4 lists all such scenarios, referring to the 16 elements of $\{+1, 0\}^4$ as the set $\{v_k\}_{k=1}^{16}$.

Table 3.4: Deterministic strategies and resultant observable outcomes

		ab				ab'				$a'b$				$a'b'$			
	a, a', b, b'	++	+0	0+	00	++	+0	0+	00	++	+0	0+	00	++	+0	0+	00
v_1	+ + + +	X				X				X				X			
v_2	0 + + +			X				X		X				X			
v_3	+ 0 + +	X				X						X				X	
v_4	+ + 0 +		X			X					X			X			
v_5	+ + + 0	X					X			X					X		
v_6	0 0 + +			X				X				X				X	
v_7	+ + 0 0		X				X				X				X		
v_8	0 + + 0			X					X	X					X		
v_9	+ 0 0 +		X			X							X			X	
v_{10}	+ 0 + 0	X					X					X					X
v_{11}	0 + 0 +				X			X			X			X			
v_{12}	0 0 0 +				X			X					X			X	
v_{13}	0 0 + 0			X					X			X					X
v_{14}	0 + 0 0				X				X		X				X		
v_{15}	+ 0 0 0		X				X						X				X
v_{16}	0 0 0 0				X				X				X				X

We can consider the probability distributions that will arise if λ_i uniformly takes a particular value v_k for each value of i. As an example, Table 3.5 gives the probability distributions generated by the assignment v_{11}=(0, +, 0, +). We call these induced distributions $\vec{v_{11}}$ and $\vec{v_{11}}^{**}$ for the two variables T_i and U_j, respectively.

Table 3.5: Probability distributions $\vec{v_{11}}$ and $\vec{v_{11}}^{**}$ induced by deterministic strategy v_{11}=(0, +, 0, +) for the random variables T_i (left) and U_j (right)

	++	+0	0+	00
ab	0	0	0	1/4
ab'	0	0	1/4	0
$a'b$	0	1/4	0	0
$a'b'$	1/4	0	0	0

	++	+0	0+
ab	0	0	0
ab'	0	0	1/3
$a'b$	0	1/3	0
$a'b'$	1/3	0	0

Under the most general local deterministic theory, λ does not have to always pick one particular v_k, but can instead implement a mixed strategy of assigning different probabilities to the different v_k. Then the distribution for the random variable T_i will be given by $\sum_{k=1}^{16} c_k \vec{v_k}$, where c_k represents $P(\lambda_i = v_k)$. As for the random variables U_j, the general situation is a little less clear, but one might conjecture that a similar result holds: i.e., that the probability distribution must be a convex combination of the distributions $\vec{v_k}^{**}$. This conjecture turns out to be true, with one caveat. The particular strategy $v_{16} = (0, 0, 0, 0)$ is problematic: if repeated, v_{16} will result in U_j being undefined. (This would yield an experiment that never produces any detection results.) So we want to set aside the T_i sequences that end with an infinite sequence of v_{16}. While it may be possible to prove everything that follows under the conditional assumption that U_j is defined, we instead make the simplifying assumption that the distribution of the T_i variables is such that each U_j takes a value with probability one. This rules out, for instance, T_i sequences that end with an infinite chain of $\vec{v_{16}}$ distributions. This is natural, as we are trying to model experiments that would continue to indefinitely produce non-00 outcomes if the apparatus were left running. With this restriction, we can state the following result:

Proposition 3.9.1. *Suppose that the sequence T_i is modeled by a local deterministic theory, for which U_1 is defined with probability 1. Then the probability distribution for U_1 can be expressed as a convex combination of the 15 induced local deterministic strategies $\{\vec{v_k}^{**}\}_{k=1}^{15}$.*

Proof. Let $\vec{T_i^*}$ denote the restriction of $\vec{T_i}$ to the (sub)probability distribution on the elements of $\mathcal{K}$ (the 12 non-00 outcomes), and let $q_i = P(T_i \notin \mathcal{K})$ with $q_0 = 1$. Then the distribution of U_1 can be expressed as

$$\vec{U_1} = \sum_{i=1}^{\infty} \left[\prod_{j=0}^{i-1} q_j \right] \vec{T_i^*}. \tag{3.82}$$

Now, in this infinite sum, some of the $\vec{T_i^*}$ distributions may be the degenerate distribution induced by v_{16}, corresponding to $c_{i_{16}} = 1$. We would like to remove these terms from the expression (3.82). Define $S \subseteq \mathbb{N}$ to be $S = \{i \mid \vec{T_i^*} \neq \vec{0}\}$. Note that if $\vec{T_k^*}$ is the nth occurrence of a non-$\vec{0}$ distribution, and $\vec{T_l^*}$ comes some trials later with the $(n+1)$th non-$\vec{0}$ distribution, then q_k will be less than 1, but $q_{k+1} = q_{k+2} = ... = q_{l-2} = q_{l-1} = 1$, so the intervening q terms in the product in (3.82) can be dropped without changing the value of the product. Hence we can rewrite (3.82) as

$$\vec{U_1} = \sum_{i \in S} \left[\prod_{(S \cup \{0\}) \cap \{j \mid j < i\}} q_j \right] \vec{T_i^*}. \tag{3.83}$$

As (3.83) is a little unwieldy, we continue to work from expression (3.82), taking the indices to have been re-enumerated so as to only to refer to the $\vec{T_i}^*$ for which $i \in S$. If we define

$$p_{v_k} = \text{the sum of the 12 CH components of } \vec{v_k},$$

we can see that $\vec{T_i}^*$ equals $\sum_{k=1}^{15} c_{i_k}(p_{v_k}\vec{v_k}^{**})$, so we can write

$$\vec{U_1} = \sum_{i=1}^{\infty}\left[\prod_{j=0}^{i-1} q_j\right]\sum_{k=1}^{15} c_{i_k}p_{v_k}\vec{v_k}^{**} = \sum_{k=1}^{15}\left\{\sum_{i=1}^{\infty}\left[\prod_{j=0}^{i-1} q_j\right]c_{i_k}p_{v_k}\right\}\vec{v_k}^{**}. \tag{3.84}$$

To prove the proposition, we must show that the expressions in the curly braces in (3.84) are nonnegative and sum to 1. The nonnegativity is clear. As for summing to 1, we have

$$\sum_{k=1}^{15}\left\{\sum_{i=1}^{\infty}\left[\prod_{j=0}^{i-1} q_j\right]c_{i_k}p_{v_k}\right\} = \sum_{i=1}^{\infty}\left[\prod_{j=0}^{i-1} q_j\right]\sum_{k=1}^{15} c_{i_k}p_{v_k} = \sum_{i=1}^{\infty}\left[\prod_{j=0}^{i-1} q_j\right]p_i, \tag{3.85}$$

where $p_i = P(T_i \in \mathcal{K}) = 1 - q_i$. But the expression in (3.85) is equal to exactly the probability that U_1 eventually does assumes a value. Since it is taken as a postulate that this probability equals one, the proof is complete. □

We only proved the above result for U_1, but it applies to all U_j by the intuitive principle that once U_1 has been recorded, we can consider ourselves to be re-starting the experiment at $i = 1$ with a new sequence of T_i's, for which U_2 is now "the U_1" of this sequence.

Proposition 3.9.1 tells us that any allowable distribution for U_1 must be a convex combination of the $\vec{v_k}^{**}$ distributions. This raises the question of whether the converse is true: for any given convex combination of the $\vec{v_k}^{**}$ distributions, can this distribution be realized for U_1 by a sequence of local deterministic T_i variables? The following proposition answers this in the affirmative.

Proposition 3.9.2. *If the probability distribution $\vec{U_1}$ is in the convex hull of the 15 deterministic strategies $\{\vec{v_k}^{**}\}_{k=1}^{15}$, then there exists a sequence of local deterministic distributions $\{\vec{T_i}\}$ inducing $\vec{U_j}$; furthermore, the $\vec{T_i}$ can be taken to be identically distributed and independent.*

Proof. By assumption,

$$\vec{U_1} = \sum_{k=1}^{15} d_k\vec{v_k}^{**}, \tag{3.86}$$

where $\{d_k\}$ is a set of nonnegative real numbers whose sum is 1. To come up with a similar expression

for $\vec{T}_i$, first define the following constants:

$$x = \sum_{k=1}^{7} d_k \qquad y = \sum_{k=8}^{11} d_k \qquad z = \sum_{k=12}^{15} d_k$$

$$s = \frac{3}{3x+4y+6z} \qquad t = \frac{4}{3x+4y+6z} \qquad u = \frac{6}{3x+4y+6z}.$$

The constants x, y, and z are chosen to sort and add the $\vec{v_i}^{**}$ coefficients for which the corresponding p_{v_k} values are 1, $\frac{3}{4}$, and $\frac{1}{2}$, respectively. Note that $x+y+z=1$. It turns out that $\{\vec{T}_i\}$ realizes $\vec{U}_1$ if the T_i are independent and identically distributed with the distribution

$$\vec{T}_i = \sum_{k=1}^{7} s d_k \vec{v_k} + \sum_{k=8}^{11} t d_k \vec{v_k} + \sum_{k=12}^{15} u d_k \vec{v_k}. \tag{3.87}$$

To check that this is true, note that

$$\vec{U}_{1\,(\text{induced})} = \vec{T_1}^{\,*} + q\vec{T_2}^{\,*} + q^2\vec{T_3}^{\,*} + q^3\vec{T_4}^{\,*} + \cdots \tag{3.88}$$

where $q = \frac{1}{4}yt + \frac{1}{2}zu$ is the probability that a given T_i random variable yields one of the four "00" outcomes. As all of the $\vec{T_i}^{\,*}$ are identically distributed with the distribution (3.87), the expression (3.88) can be rewritten as

$$\begin{aligned}
\vec{U}_{1\,(\text{induced})} &= \left(\frac{1}{1-q}\right)\vec{T_i}^{\,*} \\
&= \left(\frac{3x+4y+6z}{3}\right)\left[\sum_{k=1}^{7} s d_k \cdot 1 \cdot \vec{v_k}^{\,**} + \sum_{k=8}^{11} t d_k \cdot \left(\frac{3}{4}\right) \cdot \vec{v_k}^{\,**} + \sum_{k=12}^{15} u d_k \cdot \left(\frac{1}{2}\right) \cdot \vec{v_k}^{\,**}\right] \\
&= \sum_{k=1}^{7} d_k \vec{v_k}^{\,**} + \sum_{k=8}^{11} d_k \vec{v_k}^{\,**} + \sum_{k=12}^{15} d_k \vec{v_k}^{\,**} = \sum_{k=1}^{15} d_k \vec{v_k}^{\,**}.
\end{aligned}$$

□

This result is interesting. Not only can any distribution $\sum_{k=1}^{15} d_k \vec{v_k}^{\,**}$ be modeled by a sequence of local deterministic T_i variables, but the T_i sequence can in fact be taken to be independent and identically distributed. In light of Proposition 3.9.1, this means that "nothing can be gained" for a distribution of U_1 by varying the underlying T_i distributions from trial to trial: such a U_1 distribution could still be modeled by an i.i.d. T_i sequence. Of course, the U_j distributions can still differ from each other.

Propositions 3.9.1 and 3.9.2 together imply the following theorem, which classifies the collection of local deterministic distributions for U_1:

Theorem 3.9.1. *The space of local deterministic distributions for U_1 is equivalent to the convex hull of the 15 distributions $\{\vec{v_k}^{**}\}_{k=1}^{15}$ induced by the 15 deterministic strategies $\{v_k\}_{k=1}^{15}$.*

The convex hull described in Theorem 3.9.1, which is a subset of $\mathbb{R}^{12}$, is a geometric object known as a *convex polytope*. The collection of points in a convex polytope can be characterized by a collection of inequalities that must be satisfied, and this particular polytope turns out to be exactly characterized by the eight constraints (3.81). That is to say, any distribution for U_1 that satisfies these constraints can be expressed as a convex combination of the $\vec{v_k}^{**}$ distributions, and therefore can be modeled by a local deterministic theory. Since our earlier work shows that any more general hidden variable theory must satisfy all of these constraints, this indicates that, at least in this case, no expressive power is lost by taking the hidden variables to be local deterministic. This result has some appeal as the local deterministic hidden variables are more intuitive than their more general measure-theoretic counterparts.

3.10 Conclusion

We have demonstrated how to analyze an experiment that would unequivocally test Bell's inequality. Should an experiment meet the requirements outlined in this chapter and produce statistically significant data, locality will have to be abandoned as a physical principle. Though this has yet to happen, certain experts in the field [3] assert that it is only a matter of a few years before there will be a fully loophole-free test of Bell's inequality. This claim is bolstered by the recent advances exhibited by the experiments [23] and [24].

The impact of a loophole-free test of Bell's inequality would not be limited to the profound implications for the nature of the physical world. There would be practical applications as well. Recently, new protocols for cryptography [31] and random number generation [32] ingeniously utilize the non-locality exhibited by Quantum Mechanics to certify the security and randomness, respectively, of the protocols. The complete explanation of how this works is complicated and subtle, but our work in the past two chapters can provide some intuitive feeling for how these protocols function. In violating a Bell inequality, Quantum Mechanics produces results that cannot be fully attributed to the state of the system, denoted λ. So if λ encodes everything that happens between the detection events prior to detection, it stands to reason that λ can be related to the information available to

an "eavesdropper" on the experiment. Thus if the violation of a Bell inequality implies that there is some sort of randomness going on beyond what can be accounted for by λ, this will yield some degree of true, new randomness that is not accessible to the eavesdropper. This randomness can be used for its own sake [32] or exploited for cryptographic purposes [31].

The ability to implement to the protocols [31] and [32] would be beneficial, but there is a possibility that there will never be a loophole-free test of Bell's inequality and non-local behavior will never be exhibited by nature. The failure thus far to implement a loophole-free test of Bell's inequality is most commonly interpreted as a technical engineering obstacle that has yet to be overcome. However, it is possible that in the future a more complete theory of nature, agreeing with Quantum Mechanics in its domain of validity, could be found to be compatible with locality. In such a scenario, it would be a yet-unknown principle of nature that is preventing the production of entangled quantum states that are stable enough to exhibit non-locality in the context of a loophole-free Bell experiment. As discussed in Chapter 1, the prevailing theory of gravity for over 200 years – Newton's law of universal gravitation – predicted non-local interactions. Although laboratory equipment in the 18th and 19th centuries was never sensitive enough to attempt to exploit this prediction to send faster-than-light signals, only the introduction of General Relativity in the early 20th century allows us to know that such an attempt would have been doomed to failure. By analogy, a 21st century physicist trying to violate Bell's inequality could be similarly impeded by the currently-unknown successor theory to Quantum Mechanics, or even a new principle derived from the postulates of Quantum Mechanics that has yet to be discovered.

Whether we see a loophole-free Bell test in the next few years, or whether the experimentalists continue to be stymied, it will be interesting to see. In any case, it is an exciting time for this rapidly developing field.

Chapter 4

Contextuality and Domain Theory

4.1 Introduction

In the last section of Chapter 3, we discussed two protocols [31, 32] that exploit quantum non-locality to generate randomness. These are just two examples from the rapidly expanding field of *Quantum Information Theory*, the study of the use of quantum effects to achieve practical goals in communication and information processing. Starting with the 2002 work of Coecke & Martin [6], efforts have been made to apply information-theoretic constructions from the field of Domain Theory (introduced in Chapter 1) to Quantum Information Theory. In this chapter, we will use the Kochen-Specker theorem (Theorem 1.1.1), a no-go theorem related to the Bell-CHSH-CH theorems of Chapters 2 and 3, to better understand some of the information-theoretic objects studied in [6].

In particular, we will be studying two *exact domains* (recall Definition 1.4.6) that were first introduced in [6]. These two are known as the *Bayesian order* on discrete, finite probability distributions and the *spectral order* on quantum states. These orders, along with the *majorization* relation (a preorder), have interesting information-theoretic properties that provide the motivation for their study. The main insight of this chapter is that the Kochen-Specker theorem can be used to prove a 2002 conjecture in [6] about the spectral order on quantum states. This result shows that the spectral order encodes enough quantum structure to witness the contextual nature of Quantum Mechanics. The nature of the result also raises some interesting new questions about classes of isomorphisms with respect to the Bayesian order, which are explored. This yields new insights on the information-theoretic properties of the Bayesian order, and how it compares to majorization in this respect.

4.2 Majorization and the Bayesian Order Compared

Majorization and the Bayesian order are relations on the *classical n-states*, which are probability distribution on n outcomes.

Definition 4.2.1. A classical n-state is a vector $\vec{v} \in \mathbb{R}^n$ satisfying $\vec{v}_i \geq 0$ for $i \in (1, ..., n)$, and $\sum_{i=1}^{n} \vec{v}_i = 1$. We denote the set of all classical n-states as Δ^n.

We start by defining the majorization relation, which has many applications and has been studied extensively.

Definition 4.2.2. For two classical n-states $\vec{a}, \vec{b} \in \Delta^n$, we say that $\vec{b}$ *majorizes* $\vec{a}$, denoted $\vec{a} \prec \vec{b}$, if

$$\forall k \in \{1, ..., n\}, \quad \sum_{i=1}^{k} (\text{sort}(\vec{a}))_i \leq \sum_{i=1}^{k} (\text{sort}(\vec{b}))_i,$$

where $(\text{sort}(\vec{x}))_i$ denotes the ith component of $\text{sort}(\vec{x})$, and $\text{sort}(\vec{x})$ is the vector consisting of the re-arrangement of the components of $\vec{x}$ into descending order.

It is easy to see that majorization is reflexive and transitive, and is therefore a preorder. However, it is not antisymmetric; for example, in Δ^2, $(\frac{1}{3}, \frac{2}{3}) \prec (\frac{2}{3}, \frac{1}{3})$ and $(\frac{1}{3}, \frac{2}{3}) \succ (\frac{2}{3}, \frac{1}{3})$, but $(\frac{1}{3}, \frac{2}{3}) \neq (\frac{2}{3}, \frac{1}{3})$. Hence majorization is not a true partial ordering. In the counterexample just given, a distribution is permuted to get a second distribution, and the two distributions majorize each other despite not being equal. It is possible to eliminate this issue so that majorization can satisfy antisymmetry. To do this, consider the following definition:

Definition 4.2.3. Define $\Lambda^n \subseteq \Delta^n$ to be the "descending order" classical n-states; that is,

$$\Lambda^n = \{\vec{v} \in \Delta^n \mid \text{sort}(\vec{v}) = \vec{v}\}.$$

When restricted to Λ^n, it turns out that majorization is antisymmetric, and therefore a partial order. But if we want to study the full set Δ^n, we have seen this is not true. Furthermore, as majorization is not a partial order on Δ^n, it has no hope of falling under the rubric of Domain Theory, with all of the attendant results in the field. And unfortunately, there is no small tweak to the definition of majorization that could allow it to be a partial order on all of Δ^n.

Majorization has many useful properties, as we will see later in the section. We would like to see if there is another order on all of Δ^n – a true partial order – that shares some of these properties, or has other interesting properties of its own. One candidate is the Bayesian order. This order was introduced by Coecke & Martin [6] in 2002, and one of a handful of equivalent definitions is as follows:

Definition 4.2.4. (Bayesian order) For $\vec{v}, \vec{w} \in \Delta^n$, we say that $\vec{v}$ is less than $\vec{w}$ with respect to the Bayesian order if the following hold:

1. There is a coordinate permutation, σ, that puts both $\vec{v}$ and $\vec{w}$ in descending order; i.e., $\text{sort}(\sigma\vec{v}) = \sigma\vec{v}$ and $\text{sort}(\sigma\vec{w}) = \sigma\vec{w}$.

2. For $i \in (1, ..., n-1)$, we have $(\sigma\vec{v})_i (\sigma\vec{w})_{i+1} \leq (\sigma\vec{v})_{i+1} (\sigma\vec{w})_i$.

We notate this relation as follows: $\vec{v} \sqsubseteq_\Delta \vec{w}$.

This definition is a little hard to parse, so an example will be clarifying. Consider, in Δ^3, $\vec{v} = \left(\frac{3}{10}, \frac{6}{10}, \frac{1}{10}\right)$ and $\vec{w} = \left(\frac{5}{10}, \frac{3}{10}, \frac{2}{10}\right)$. By looking to condition 1 of Definition 4.2.4, we can see that $\vec{v}$ and $\vec{w}$ do not compare, because they cannot be simultaneously sorted into descending order: $\vec{w}$ is already in descending order, but $\vec{v}$ is not; furthermore, the permutation σ that puts $\vec{v}$ into descending order will scramble $\vec{w}$ out of descending order. But if we take a third vector $\vec{x} = \left(\frac{3}{10}, \frac{5}{10}, \frac{2}{10}\right)$, it potentially compares to $\vec{v}$: if we take σ to be the permutation that switches the first two entries, we have $\text{sort}(\sigma\vec{v}) = \sigma\vec{v}$ and $\text{sort}(\sigma\vec{x}) = \sigma\vec{x}$. Looking now at condition 2 of Definition 4.2.4, we have

$$\begin{aligned} i = 1: \qquad & \frac{5}{10} \cdot \frac{3}{10} \leq \frac{3}{10} \cdot \frac{6}{10} \\ i = 2: \qquad & \frac{3}{10} \cdot \frac{1}{10} \leq \frac{2}{10} \cdot \frac{3}{10} \end{aligned}$$

so condition 2 is satisfied.

Condition 2 is a little clearer when we think of it in terms of what it is saying about ratios. If $\vec{v}$ and $\vec{w}$ in Definition 4.2.4 do not contain zeroes, condition 2 can be re-expressed as

$$\text{For } i \in (1, ..., n-1), \text{ we have } \frac{(\sigma\vec{v})_i}{(\sigma\vec{v})_{i+1}} \leq \frac{(\sigma\vec{w})_i}{(\sigma\vec{w})_{i+1}}. \tag{4.1}$$

Then it is clearer, visually, that this is satisfied by $\sigma\vec{v} = \left(\frac{6}{10}, \frac{3}{10}, \frac{1}{10}\right)$ and $\sigma\vec{x} = \left(\frac{5}{10}, \frac{3}{10}, \frac{2}{10}\right)$. It is demonstrated in Coecke & Martin [6] that the Bayesian order defines a partial order on Δ^n, for a fixed n. It was additionally shown that the Bayesian order satisfies the axioms of an *exact domain*, as was defined in Definition 1.4.6.

What additional properties does the Bayesian order have? Majorization can be used as a standard for comparison. Majorization has received much study and has many interesting applications. We give some illustrative examples from the field of information theory that demonstrate its utility.

To start, one nice algebraic result is given in the following fact, which uses the notion of a *doubly stochastic matrix*: a square matrix of nonnegative real numbers for which each row and column sums to one.

Fact 4.2.1. $\vec{v} \prec \vec{w} \Leftrightarrow$ *There is a doubly stochastic matrix* M *for which* $\vec{v} = M\vec{w}$.

Stochastic matrices can be used to model communication channels between classical n-states, so

Fact 4.2.1 indicates that majorization could have some utility in the study of channels. Another information-theoretic property of majorization relates to the information-theoretic notion of *Shannon entropy*:

Definition 4.2.5. *Shannon Entropy*, denoted H, is a function $H : \Delta^n \to \mathbb{R}$ defined as such:

$$H(\vec{v}) := -\sum_{i=1}^{n} \vec{v}_i \ln(\vec{v}_i). \tag{4.2}$$

In a sense, Shannon entropy measures the degree of uncertainty in a probability distribution. Equiprobable distributions have maximum entropy, while distributions that tilt most probability towards one outcome have low entropy. The relationship of majorization to Shannon entropy is encapsulated in the following fact, which states that Shannon entropy is (reverse) monotone with respect to majorization.

Fact 4.2.2. *For* $\vec{v}, \vec{w} \in \Delta^n$, $\vec{v} \prec \vec{w} \Rightarrow H(\vec{v}) \geq H(\vec{w})$.

Majorization also has applications in Quantum Information Theory. For instance, it was shown by Nielsen [33] that in the set of circumstances known as *local operations and classical communications*, one quantum state can be transformed into another if and only if the vector of eigenvalues of the one state majorizes the vector of eigenvalues of the other state.

These facts give some flavor of the utility of majorization. Interestingly, the Bayesian order shares one of these properties:

Fact 4.2.3. *(proved in [6]) For* $\vec{v}, \vec{w} \in \Delta^n$, $\vec{v} \sqsubseteq_\Delta \vec{w} \Rightarrow H(\vec{v}) \geq H(\vec{w})$.

So at least one interesting property, the reverse-monotonicity of entropy, is shared by majorization and the Bayesian order. On the other hand, the prospects for an analog of Fact 4.2.1 for the Bayesian order are not especially bright. It seems unlikely that there could be a characterization of the Bayesian order via matrices, due to the non-linearity in its definition – recall the ratios in (4.1) in the definition of the Bayesian order. (The definition of majorization, on the other hand, involves linear constraints.)

As for Quantum Information Theory, we saw that majorization has an application in this field. The Bayesian order can also be related to quantum behavior, as will be explored in the next section. This involves an extension of the definition of the Bayesian order to a new order on quantum states.

4.3 The Spectral Order and the Kochen-Specker Theorem

For the purpose of defining a new order, we will take a "quantum state" to be as follows:

Definition 4.3.1. A quantum n-state is a self-adjoint positive n-by-n complex matrix ρ satisfying $\mathrm{tr}(\rho) = 1$. For a fixed n, we denote the collection of quantum n-states as Ω^n.

Definition 4.3.1 coincides with the density operator concept introduced in the last paragraph of Section 1.2. The conditions satisfied by a quantum n-state ρ imply that ρ is diagonalizable in an orthogonal basis, and that in this basis, its diagonal entries (which are the eigenvalues) form a classical n-state. This is the sense in which the quantum n-states can be thought of as an extension of the classical n-states. This allows us to lift the Bayesian order to the quantum states in the following way:

Definition 4.3.2. For two quantum n-states ρ and τ, we say that ρ is less than τ in the *spectral order*, denoted $\rho \sqsubseteq_\Omega \tau$, if

1. There is an orthonormal basis $E = (e_1, ..., e_n)$ of $\mathbb{C}^n$ in which both ρ and τ are diagonal.
2. $\mathrm{spec}_E(\rho) \sqsubseteq_\Delta \mathrm{spec}_E(\tau)$, where $\mathrm{spec}_E(\rho)$ is the vector of eigenvalues of ρ corresponding to $(e_1, ..., e_n)$.

In Coecke & Martin [6], it was proved that the spectral order is well-defined, and that just like the Bayesian order, it is an exact domain. Furthermore, the map $q : \Omega^n \to \Lambda^n$ that takes a quantum n-state to its eigenvalues sorted in descending order was shown to be Scott continuous and monotone. The map q is essentially a quantum extension of the map sort:$\Delta^n \to \Lambda^n$ that was defined in the previous section.

In studying the structure of the spectral order, Section 3.3 of the paper [6] posited a conjecture which uses the notion of an "order isomorphism," which we define here:

Definition 4.3.3. Let "$\sqsubseteq_a$" denote a preorder on a set A, let "$\sqsubseteq_b$" denote a preorder on a set B, and let $\phi : A \to B$ be a map. We say that ϕ is an *order isomorphism* if ϕ is a bijection and for $x, y \in A$,

$$x \sqsubseteq_a y \Leftrightarrow \phi(x) \sqsubseteq_b \phi(y). \tag{4.3}$$

We say that ϕ is a *restriction order isomorphism* if ϕ is injective and satisfies (4.3).

The motivation for the term "restriction order isomorphism" is that when we restrict an injective map to its image, it is a bijection, so it is essentially an order isomorphism onto its image.

The notion of a restriction order isomorphism will be used in the next section. We now present the conjecture from [6]:

CONJECTURE: There is no map $M : \Omega^n \to \Delta^n$ for which

1. $q = \text{sort} \circ M$
2. For any basis E of $\mathbb{C}^n$, $M|_{\Omega|_E}$ is an order isomorphism from $\Omega^n|_E$ (with the inherited order) to Δ^n, where $\Omega^n|_E$ is the collection of quantum n-states diagonalizeable in E, and $M|_{\Omega|_E}$ is the restriction of M to $\Omega^n|_E$.

We show here that conjecture is true: there is no such map. The existence of a map satisfying condition (2) above would violate the Kochen-Specker theorem.

Proposition 4.3.1. *For $n \geq 3$, there is no map M satisfying condition (2) above.*

Proof. Recall that the Kochen-Specker theorem (Theorem 1.1.1) says that in $\mathbb{R}^3$, it is impossible to assign the numbers 1 or 0 to every unit vector in such a way that for every set of three orthogonal vectors, exactly one of them is labeled 1, and the other two are labeled 0. It turns out that if a map M existed which satisfied condition (2), we could generate a number-assignment map that does precisely this.

Suppose we had such an M for Ω^3. Every quantum 3-state ρ corresponds to an orthonormal basis of $\mathbb{C}^3$ with an associated classical 3-state $\vec{v}$, where the ith entry of $\vec{v}$ is the eigenvalue of ρ corresponding to the ith entry of the orthonormal basis. Hence the quantum 3-states include orthonormal bases of $\mathbb{R}^3$ with the associated $\vec{v}$ being $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, or $e_3 = (0, 0, 1)$. (Such quantum 3-states, where the classical n-state consists of one 1 and two 0s, are examples of so-called *pure states*.) Let E be such an orthonormal basis of $\mathbb{R}^3$, and denote the three associated "pure states" by ρ_1, ρ_2, and ρ_3. These are maximal elements in Ω^3, as shown in Theorem 3.26 of [6]. Thus by condition (2), M must map them bijectively to the maximal elements in Δ^3, which are $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, and $e_3 = (0, 0, 1)$. Now consider the simple map $P : \{e_1, e_2, e_3\} \to \{0, 1\}$ where $P(e_1) = 1$ and $P(e_2) = P(e_3) = 0$. Then if we restrict the domain of M to be quantum states diagonalizable by an orthonormal basis of $\mathbb{R}^3$, $P \circ M$ is a labeling map satisfying the Kochen-Specker condition, and therefore M cannot exist.

Suppose now that we had such an M for Ω^n for a fixed $n > 3$. By the argument above, this could be used to define a map $L : \mathbb{R}^n \to \{0, 1\}$ such that for any orthogonal basis of $\mathbb{R}^n$, L

maps exactly one basis vector to 1. Such a map implies the existence of a lower-dimensional map $L^* : \mathbb{R}^3 \to \{0, 1\}$ that contradicts Theorem 1.1.1, and therefore M cannot exist. □

The Kochen-Specker theorem implies that Quantum Mechanics exhibits non-contextual behavior in certain measurement settings. Thus Proposition 4.3.1 demonstrates that the spectral order is rich enough to capture some of the essential quantum features of Ω^n – for instance, if a map M *were* to exist, it may not be possible to prove that Quantum Mechanics is non-contextual. At the very least, certain low-dimensional quantum scenarios that are known to be contextual would have to be non-contextual for such a map M to exist.

This result raises further questions. One question is whether there could be any sort of converse to Proposition 4.3.1 – for instance, an order-theoretic proof of the Kochen-Specker theorem, or of the noncontextuality of quantum mechanics. This is an interesting idea, but it does not seem that such a proof could really be novel, and would likely be just a repackaging of a geometric argument similar to the ones employed in [4, 5] with order-theoretic language. Alternatively, one could pursue a narrower converse by attempting to show that the existence of a Kochen-Specker map from $\mathbb{R}^3$ to $\{0, 1\}$ would imply the constructability of an M-style map from Ω^n to Δ^n. Seeing as M exists in the higher-dimensional complex space Ω^n and has many order-theoretic requirements (while the Kochen-Specker map has no requirements beyond the orthogonal-triple criterion), this does not seem likely.

Another more fruitful line of inquiry concerns whether the conjecture can be modified somewhat. Specifically, Proposition 4.3.1 is an order-theoretic result, but condition (2) is not strictly order-theoretic: it refers to structures beyond $\sqsubseteq_\Omega$, the bases of $\mathbb{C}^n$. Could (2) be rephrased in purely order-theoretic terms? As the bases of $\mathbb{C}^n$ are used in the definition of the spectral order, it may be possible to recover these structures order-theoretically, and thus rephrase (2) as a purely order-theoretic statement. On the other hand, there is a more direct way to rephrase (2) using only order-theoretical terms, in the following manner:

$(2')$: For any subset S of Ω^n isomorphic to Δ^n in the inherited order,
M restricted to S is an order isomorphism.

This will do the trick: since any set of the form $\Omega|_E$ is isomorphic to Δ^n in the inherited order, we can replace condition (2) with $(2')$, and the proof of Proposition 4.3.1 will still hold. Furthermore, $(2')$

only makes references to the spectral order, and does not require reference to any outside geometric structures.

4.4 Restriction Order Isomorphisms and Channels

Proposition 4.3.1 with the substitution of $(2')$ demonstrates that the Spectral order retains enough of the quantum structure to witness the effects of Kochen-Specker contextuality, doing so in purely order-theoretic terms. This achieves the task, but one might also hope that $(2')$ is not an overly strong condition. For instance, it is possible that there are subsets of Ω^n that are isomorphic to Δ^n, but are not sets of the form $\Omega|_E$. If $(2')$ encompasses many more sets than we thought, then perhaps the violation of the Kochen-Specker theorem would not be the only obstruction to the existence of a map M. In such a situation, the presence of quantum contextual constraints in the spectral order would still be demonstrated by Proposition 4.3.1 with the $(2')$ replacement, but the hypothesized map M might not be the simplest witness of such constraints.

This issue is also related to the classes of communication channels between classical n-states. Of particular interest is the possibility that a proper subset of a particular set $\Omega|_E$ could be order isomorphic to Δ^n: because $\Omega|_E$ is isomorphic to Δ^n, this would have implications for the nature of both the spectral and Bayesian orders.

A proper subset of Δ^n can be isomorphic to Δ^n if and only if there is a restriction order isomorphism $\phi : \Delta^n \to \Delta^n$ whose range is a subset of Δ^n. What could such a map look like? One first candidate for such a map would be the *depolarization channel*. This is defined as follows:

Definition 4.4.1. For any given $t \in (0, 1)$, the corresponding *depolarization channel* is the map $\eta_t : \Delta^n \to \Delta^n$ defined as such:

$$\eta_t(\vec{v}) := t \cdot \perp + (1 - t) \cdot \vec{v}, \tag{4.4}$$

where $\perp$ is the n-state with $1/n$ as every entry.

This is perhaps the most basic map from Δ^n to a proper subset of Δ^n, and the simplest candidate for a restriction order isomorphism. Indeed, it is a restriction order isomorphism with respect to a familiar order:

Proposition 4.4.1. *The depolarization channel is a restriction order isomorphism on Δ^n with respect to majorization.*

Proof. It is clear that η_t is injective: if $\eta_t(\vec{v}) = \eta_t(\vec{w})$, then

$$t \cdot \perp + (1-t) \cdot \vec{v} = t \cdot \perp + (1-t) \cdot \vec{w} \qquad \Rightarrow \qquad \vec{v} = \vec{w},$$

which we obtain just by subtracting $t \cdot \perp$ from both sides, then dividing by $(1-t)$.

It requires a few more manipulations to show that (4.3) holds:

$$\begin{aligned}
\vec{v} \prec \vec{w} \quad &\Leftrightarrow \quad \forall k \in (1,...,n), \sum_{i=1}^{k} (\mathrm{sort}(\vec{v}))_i \leq \sum_{i=1}^{k} (\mathrm{sort}(\vec{w}))_i \\
&\Leftrightarrow \quad \forall k \in (1,...,n), \sum_{i=1}^{k} (1-t)(\mathrm{sort}(\vec{v}))_i \leq \sum_{i=1}^{k} (1-t)(\mathrm{sort}(\vec{w}))_i \\
&\Leftrightarrow \quad \forall k \in (1,...,n), \sum_{i=1}^{k} \left[(1-t)(\mathrm{sort}(\vec{v}))_i + t/n\right] \leq \sum_{i=1}^{k} \left[(1-t)(\mathrm{sort}(\vec{w}))_i + t/n\right] \\
&\Leftrightarrow \quad \forall k \in (1,...,n), \sum_{i=1}^{k} \mathrm{sort}\left[(1-t)\vec{v} + t \cdot \perp\right]_i \leq \sum_{i=1}^{k} \mathrm{sort}\left[(1-t)\vec{w} + t \cdot \perp\right]_i \\
&\Leftrightarrow \quad \eta_t(\vec{v}) \prec \eta_t(\vec{w}).
\end{aligned}$$

□

Proposition 4.4.1 suggests that the depolarization channel is a reasonable candidate for a restriction order isomorphism with respect to the Bayesian order on Δ^n, as well. However, this is *not* the case for $n \geq 3$, thus highlighting an important difference between these two orders. To see why, consider $\vec{v} = (1, 0, ..., 0)$ and $\vec{w} = (\frac{1}{2}, \frac{1}{2}, 0, ..., 0)$, so that $\vec{v} \sqsupseteq_\Delta \vec{w}$. We then have

$$\begin{aligned}
\eta_t(\vec{v}) &= \left(1 - \frac{(n-1)t}{n}, \frac{t}{n}, ..., \frac{t}{n}\right) \\
\eta_t(\vec{w}) &= \left(\frac{1}{2}\left(1 - \frac{n-1}{n}\right)t, \frac{1}{2}\left(1 - \frac{n-1}{n}\right)t, \frac{t}{n}, ..., \frac{t}{n}\right),
\end{aligned}$$

and so $\eta_t(\vec{v}) \not\sqsupseteq_\Delta \eta_1(\vec{w})$.

The failure of the depolarization channel to be a restriction order isomorphism for $n \geq 3$ is symptomatic of a more fundamental constraint on restriction order isomorphisms, which is given below as Proposition 4.4.2. The proposition appeals to the notion of Scott continuity (Definition 1.4.7) – an essential characteristic for a well-behaved map between domains. It requires the following technical lemma:

Lemma 4.4.1. *For $\vec{v} = (a, b, c) \in \Delta^3$ satisfying $a > b > c$, $\downarrow\vec{v}$ includes all vectors of the form*

(d, e, f) satisfying $d \geq e \geq f$ with $\frac{d}{e} < \frac{a}{b}$ and $ec < bf$.

Proof. Fix a vector (d, e, f) satisfying the above conditions. Let D be any directed set for which $\sup D = \vec{v} = (a, b, c)$. By Proposition 2.16 (ii) in Coecke & Martin [6], D contains an increasing sequence with the same supremum. Let us notate this sequence as $S = \{(x_i, y_i, z_i) \mid i \in \mathbb{N}\}$. By 2.16(i) *Ibid.*, $\sup S = (\lim x_i, \lim y_i, \lim z_i)$. Because S is an increasing sequence, $\frac{x_i}{y_i}$ is increasing with limit $\frac{a}{b}$. Since $\frac{a}{b} > \frac{d}{e}$, $\frac{x_i}{y_i}$ will eventually exceed $\frac{d}{e}$. A similar argument applies to y_i and z_i, so we see that eventually, S contains elements that are greater than (d, e, f) in the Bayesian order. Since $S \subseteq D$, this means that $(d, e, f) \ll_e \vec{v}$. □

Proposition 4.4.2. *Let $\phi : \Delta^3 \to \Delta^3$ be a Scott continuous restriction order isomorphism. Then ϕ must map maximal elements to maximal elements.*

Proof. Consider $(1, 0, 0)$, which we will refer to as e_1. Here is the strategy: we will construct a family of chains in Δ^3, all of which have supremum e_1, and we will show that these chains bear certain relations to each other, relations that ϕ must preserve. We will then see that no non-maximal point in Δ^3 can have a family of chains satisfying this condition. This will make it impossible for a restriction order isomorphism to map e_1 to a non-maximal point.

For a fixed $\alpha \in (0, 1)$, define the chain $C^\alpha \subseteq \Delta^3$ as follows:

$$C^\alpha := \{(1 - (1 + \alpha)y, y, \alpha y) \mid y \in (0, 1/2)\} \tag{4.5}$$

C^α is indeed a chain. To see this, let C^α_y be the element of C^α with parameter value y. Then it turns out that $C^\alpha_y \sqsupseteq_\Delta C^\alpha_w$ iff $y \leq w$. This is because every element of C^α is already sorted in decreasing order, and the ratios in (4.1) yield

$$\frac{1 - (1 + \alpha)y}{y} \geq \frac{1 - (1 + \alpha)w}{w} \quad \text{and} \quad \frac{y}{\alpha y} \geq \frac{w}{\alpha w},$$

which both hold iff $y \leq w$. Hence C^α is isomorphic to $(0, \frac{1}{2})$ in reverse order, and so is a chain.

It is clear from 2.16 (i) in Coecke & Martin [6] that e_1 is an upper bound for C^α. This result says that the supremum of an increasing sequence is equal to its componentwise limit; applying this to C^α yeilds a limit of $(1, 0, 0)$ as $y \to 0$.

The indexed family of chains C^α satisfies the following important condition:

$$\alpha_1 > \alpha_2 \Rightarrow \downarrow C^{\alpha_1} \cap C^{\alpha_2} = \emptyset. \tag{4.6}$$

We can see that this holds by looking at the second ratio in (4.1); if $\alpha_1 > \alpha_2$, then for any $y, w \in (0, 1/2)$, we have

$$\frac{y}{\alpha_1 y} \not\geq \frac{w}{\alpha_2 w}.$$

Equation (4.6) implies the complete disjointness of all the C^α's from $\downarrow e_1$: for any $\vec{v}$ in any C^α, we have $\vec{v} \not\ll e_1$, because we can pick a C^β chain with $\beta > \alpha$, and C^β is a directed set with $\sup C^\beta = e_1$, and $\vec{v}$ is not below any element of C^β. Succinctly,

$$\forall \alpha \in (0, 1), \quad C^\alpha \cap \downarrow e_1 = \emptyset. \tag{4.7}$$

Now, let's see what this says about $\phi(e_1)$. Let $\vec{v} = \phi(e_1)$, and let $\phi(C^\alpha)$ denote the image of the set C^α under ϕ. By the Scott continuity of ϕ, $\sup \phi(C^\alpha) = \vec{v}$; by the restriction order isomorphism property of ϕ, $\phi(C^\alpha)$ is a chain for each $\alpha \in (0, 1)$, and $\alpha_1 > \alpha_2 \Rightarrow \downarrow \phi(C^{\alpha_1}) \cap \phi(C^{\alpha_2}) = \emptyset$. So, $\vec{v}$ must be the supremum of an indexed, disjoint family of chains satisfying (4.6). We now show that no non-maximal $\vec{v}$ can have this property.

Write an arbitrary non-maximal element $\vec{v}$ in Δ^3 as (a, b, c), and assume without loss of generality that $a \geq b \geq c$. By assumption of non-maximality, this means that $b > 0$. We can quickly rule out a few degenerate cases: if $\vec{v}$ is of the form (a, b, b) with $b \neq 0$, or of form (a, a, c), it can be shown that $\downarrow \vec{v}$ is a chain; this is proved in proposition 4.2 of Coecke [34]. If $\downarrow \phi(e_1)$ is a chain, there is no way that $\downarrow \phi(e_1)$ can contain an indexed family of chains satisfying the incomparability condition (4.6). This leaves the only candidate for $\phi(e_1)$ as (a, b, c) with $a > b > c$ and $b \neq 0$. The characterization of $\downarrow \vec{v}$ is then given by Lemma 4.4.1, and this characterization allows us to complete the proof. Suppose $\vec{v} = (a, b, c)$ with $a > b > c$ and $b \neq 0$. Then $\vec{v}$ must be the supremum of an indexed disjoint family of chains satisfying (4.6), which means that the chains must satisfy (4.7) by the argument preceding (4.7).

That is, the images of the C^α chains must all lie in $\downarrow \vec{v} \cap (\downarrow \vec{v})^C$; by Lemma 4.4.1, every element (d, e, f) in this set has either $\frac{d}{e} = \frac{a}{b}$ or $ec = bf$. In Δ^3, sets of vectors with fixed ratios between two components are chains – this is essentially what was proved when we demonstrated that each C^α, as defined in (4.5), is a chain. Hence $\downarrow \phi(\vec{v}) \cap (\downarrow \vec{v})^C$ is a union of two chains, and it is clearly impossible to find an indexed family of chains satisfying (4.6) in such a set. (One could find an indexed family of two chains lying in such a set, but not more than two, and infinitely many are needed.) This rules out the last non-maximal possibility for $\phi(e_1)$. So, $\phi(e_1)$ must be a maximal

element. □

Note that the maximal elements in Δ^3 are $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$. Since ϕ is assumed to be injective, Proposition 4.4.2 thus says that ϕ acts as a permutation on this set of three elements. This result is additional demonstration that once we move beyond the trivial $n = 2$ case (majorization and the Bayesian order are in fact identical on Λ^2), the two orders have a very different characteristics.

The maximal elements in Δ^3 are a spanning set, which gives the following:

Corollary 4.4.1. *The only linear Scott continuous restriction order isomorphisms of Δ^3 are permutations of the coordinates.*

Corollary 4.4.1 has implications for the relationship between the Bayesian order and channels between classical states. We have not rigorously defined the notion of a "channel," as there are different notions, but a classical channel is a map between classical states that satisfies some additional set of conditions. The precise set of conditions depends on the particular setting, but Scott continuity and linearity are desirable traits when analyzing a channel through the lens of the Bayesian order. We have already seen that the depolarization channel is not a restriction order isomorphism with respect to the Bayesian order, and Corollary 4.4.1 puts further limitations on the types of channels that could be restriction order isomorphisms.

References

[1] J. Bell, "On the Einstein Podolsky Rosen paradox," *Physics*, vol. 1, pp. 195–200, 1964.

[2] J. Clauser, A. Horne, A. Shimony, and R. Holt, "Proposed experiment to test local hidden-variable theories," *Phys. Rev. Lett.*, vol. 23, no. 15, pp. 880–884, 1969.

[3] E. Knill. Private communication.

[4] S. Kochen and E. Specker, "The problem of hidden variables in quantum mechanics," *Indiana Univ. Math. J.*, vol. 17, pp. 59–87, 1968.

[5] R. D. Gill and M. S. Keane, "A geometric proof of the Kochen-Specker no-go theorem," *J. Phys. A: Math. Gen.*, vol. 29, no. 12, p. L289, 1996.

[6] B. Coecke and K. Martin, *A Partial Order on Classical and Quantum States.*
Research Report PRG-RR-02-07, Oxford University Computing Laboratory, 2002.

[7] M. Nielsen and I. Chuang, *Quantum Computation and Quantum Information.*
Cambridge: Cambridge University Press, 2000.

[8] R. Bartle, *The Elements of Integration and Lebesgue Measure.*
New York: John Wiley & Sons, 1995.

[9] A. D. Wentzell, *Lecture Notes for the course Probability (Math 7550).*
Tulane Univesity, 2010.
Available online at 129.81.170.14/∼wentzell/755dir.html.

[10] K. L. Chung, *A Course in Probability Theory.*
Academic Press, 2nd ed., 1974.

[11] A. Brandenburger and N. Yanofsky, "A classification of hidden-variable properties," *J. Phys. A: Math. Theor.*, vol. 41, no. 425302, 2008.

[12] A. Brandenburger and H. J. Keisler, "A canonical hidden-variable space," 2012.
Available online at http://www.stern.nyu.edu/∼abranden.

[13] J. Clauser and M. Horne, "Experimental consequences of objective local theories," *Phys. Rev. D*, vol. 10, no. 2, pp. 526–535, 1974.

[14] J. Barrett, D. Collins, L. Hardy, A. Kent, and S. Popescu, "Quantum nononlocality, Bell inequalities, and the memory loophole," *Phys. Rev. A*, vol. 66, no. 042111, 2002.

[15] R. D. Gill, "Accardi contra Bell (cum mundi): The impossible coupling," *Mathematical Statistics and Applications: Festschrift for Constance van Eeden IMS Lecture Notes - Monograph*, vol. 42, pp. 133–154, 2003.

[16] G. Weihs, T. Jennewein, C. Simon, H. Weinfurter, and A. Zeilinger, "Violation of Bell's inequality under strict Einstein locality conditions," *Phys. Rev. Lett.*, vol. 81, pp. 5039–43, Dec 1998.

[17] M. A. Rowe, D. Kielpinski, V. Meyer, C. A. Sackett, W. M. Itano, C. Monroe, and D. J. Wineland, "Experimental violation of a Bell's inequality with efficient detection," *Nature*, vol. 409, pp. 791–4, 2001.

[18] M. Ansmann *et al.*, "Violation of Bell's inequality in Josephson phase qubits," *Nature*, vol. 461, pp. 504–6, 2009.

[19] N. Gisin and B. Gisin, "A local hidden variable model of quantum correlation exploiting the detection loophole," *Phys. Lett. A*, vol. 260, pp. 323–327, 1999.

[20] P. H. Eberhard, "Background level and counter efficiencies required for a loophole-free Einstein-Podolsky-Rosen experiment," *Phys. Rev. A*, vol. 47, pp. R747–R750, Feb 1993.

[21] J.-A. Larsson and J. Semitecolos, "Strict detector-efficiency bounds for n-site Clauser-Horne inequalities," *Phys. Rev. A*, vol. 63, p. 022117, 2001.

[22] I. Pitowsky, *Quantum Probability – Quantum Logic, Lecture Notes in Physics*, vol. 321.
Springer-Verlag, 1989.

[23] M. Giustina *et al.*, "Bell violation using entangled photons without the fair-sampling assumption," *Nature*, vol. 497, pp. 227–30, 2013.

[24] B. G. Christensen *et al.*, "Detection-loophole-free test of quantum nonlocality, and applications," *Phys. Rev. Lett.*, vol. 111, p. 130406, Sep 2013.

[25] J. Kofler, S. Ramelow, M. Giustina, and A. Zeilinger, "On 'Bell violation using entangled photons without the fair-sampling assumption'," 2013.
arXiv:1307.6475 [quant-ph].

[26] J.-A. Larsson, M. Giustina, J. Kofler, B. Wittmann, R. Ursin, and S. Ramelow, "Bell violation with entangled photons, free of the coincidence time loophole," 2013.
arXiv:1309.0712 [quant-ph].

[27] Y. Zhang, S. Glancy, and E. Knill, "Efficient quantification of experimental evidence against local realism," *Phys. Rev. A*, vol. 88, p. 052119, Nov 2013.

[28] G. Grimmett and D. Stirzaker, *Probability and Random Processes*.
Oxford: Oxford University Press, 3rd ed., 2001.

[29] C. McDiarmid, "On the method of bounded differences," in *Surveys in Combinatorics, 1989*, vol. 141, pp. 148–88, Cambridge: Cambridge Univ. Press, 1989.

[30] W. van Dam, R. D. Gill, and P. D. Grunwald, "The statistical strength of nonlocality proofs," *IEEE T. Inform. Theory*, vol. 51, pp. 2812–35, 2005.

[31] J. Barrett, L. Hardy, and A. Kent, "No signaling and quantum key distribution," *Phys. Rev. Lett.*, vol. 95, p. 010503, 2005.

[32] S. Pironio *et al.*, "Random numbers certified by Bell's theorem," *Nature*, vol. 464, pp. 1021–4, 2010.

[33] M. A. Nielsen, "Conditions for a class of entanglement transformations," *Phys. Rev. Lett.*, vol. 83, pp. 436–439, Jul 1999.

[34] B. Coecke, *Entropic Geometry from Logic*.
arXiv:quant-ph/0212065v3, 2003.

Biography

The author was born in New York City in 1982 and graduated from Harvard University with an AB in mathematics in 2004. After college, the author spent a few years tutoring math in New York City and Portland, OR before starting the PhD program at the Tulane University mathematics department in August 2008, eventually completing the program in June 2014 after earning an MS in statistics along the way.