The statistical strength of nonlocality proofs

Wim van Dam* Richard D. Gill† Peter D. Grünwald‡

September 26, 2003

Abstract

The strength of a nonlocality proof is examined in terms of the amount of evidence that the corresponding experiment provides for the nonlocality of Nature. An experimental implementation of such a proof gives data whose statistics will differ from the statistics that are possible under a local description of Nature. The strength of the experiment is quantified by the expected deviation between the observed frequencies, which are given by the laws of quantum mechanics, and the closest possible local theory. Varying the frequencies of the measurement settings gives different experimental implementations of a nonlocality proof, giving each implementation its own strength. The statistical strength of a nonlocality proof is thus determined by the experimental implementation that maximizes its statistical deviation from all possible local theories.

It is shown that the deviation between quantum mechanics and a local theory is best expressed by the Kullback-Leibler distance between the probability distributions over the measurement outcomes that the respective theories predict. Specifically, it is proven that the Kullback-Leibler distance is optimal for three methods of hypothesis testing: frequentist, Bayesian, and information theoretic hypothesis testing.

The nonlocality proofs that are analyzed in this article are: Bell's original proof, an improved version of Bell's proof, the CHSH inequality, Hardy's proof, a proof by Mermin, and the 3-party GHZ inequality. The outcome is that the GHZ proof is an order of magnitude stronger than all other proofs, while of the two party proofs, the CHSH inequality is the strongest.

Contents

*Computer Science Department, University of California, Soda Hall, Berkeley, CA 94720-1776, USA. Also at Mathematical Sciences Research Institute, Berkeley, and HP Labs, Palo Alto, CA, USA. vandam@cs.berkeley.edu

†Mathematical Institute, University Utrecht, Budapestlaan 6, NL-3584 CD Utrecht, The Netherlands. Also at EURANDOM, Eindhoven, The Netherlands. gill@math.uu.nl

‡CWI, P.O. Box 94079, NL-1090 GB The Netherlands. Also at EURANDOM, Eindhoven, The Netherlands. pdg@cwi.nl

1 Introduction

A plethora of proofs exist of Bell's theorem ("quantum mechanics violates local realism") encapsulated in inequalities and equalities of which the most celebrated are those of Bell [4], Clauser, Horne, Shimony and Holt (CHSH) [7], Greenberger, Horne and Zeilinger (GHZ) [13], Hardy [18], and Mermin [23]. Competing claims exist that one proof is stronger than another, for instance, a proof in which quantum predictions having probabilities 0 or 1 only are involved, is often said to be more strong than a proof which involves quantum predictions of probabilities between 0 and 1. Now one has to distinguish between a mathematical proof that the *predicted* probabilities of quantum theory are incompatible with local realism, and an experimental proof that *physical reality* conforms to those predictions and hence too is incompatible with local realism. That some outcomes should have zero probability, also needs experimental confirmation. If the event occurs just once, one can rule out that theory. However even if the outcome is never observed in millions of replications of the experiment in question, we never know for sure that it is impossible.

To put it another way, the strength of a mathematical proof is measured in terms of the weakness of its assumptions. The strength of an experimental proof is measured in statistical terms: how sure do we become that a certain theory is false, after observing a certain violation from that theory, in a certain number of experiments.

Our Game In fact, when comparing different potential experiments, various other aspects also come into play, such as: how easy is it to prepare certain types of particles in certain states? Can we arrange to have the time and spatial separations which are necessary to make the results convincing? Can we implement the necessary random changes in settings per trial, quickly enough? We shall neglect all these practical aspects and just analyze the statistical aspect of how much statistical information is provided per independent trial, in a given design corresponding to a given proof of Bell's theorem, independently of the costs and time necessary per trial. We propose to further analyze this from a game theoretic point of view. The two players involved are the pro-quantum theory experimenter QM, and a pro-local realism theoretician LR. The experimenter QM is armed with a specific proof of Bell's theorem. A given proof—Bell, CHSH, Hardy, Mermin, GHZ—involves a collection of equalities and inequalities between various experimentally accessible probabilities. The proof specifies a given quantum state (of a collection of entangled qubits, for instance) and experimental settings (orientations of polarization filters or Stern-Gerlach devices), such that the equalities hold under QM but are impossible under LR, or such that the inequalities hold under LR but are violated under QM. The QM experimenter still has a choice of the relative frequency, with which the different combinations of settings will be applied, in a long sequence of independent trials. In other words, he must still decide how to allocate his resources over the different combinations of settings. At the same time, the local realist can come up with all kinds of different local realistic theories, predicting different probabilities for the outcomes given the settings. She might put forward different theories in response to different specific experiments. Thus the quantum experimenter will choose that probability distribution over his settings, for which the *best* local realistic model explains the data *worst*, when compared with the true (quantum mechanical) description.

In the past this feature has been quantified by simply saying: the largest deviation in the Bell inequality is attained with such and such filter settings, and hence the experiment which is done with these settings gives (potentially) the strongest proof of nonlocality. The argument is however not very convincing. One should take account of the statistical variability in finite statistics (physicists' jargon for finite sample sizes). The experiment which might confirm the largest absolute deviation from local realistic theories, might be subject to the largest standard errors, and therefore be less convincing than an experiment where a much smaller deviation can be proportionally much more accurately determined.

Alternatively, the argument has just been that with a large enough sample size, even the smallest deviation between two theories can be made firm enough. For instance, [23] has said in the context of a particular example

> "...to produce the conundrum it is necessary to run the experiment sufficiently many times to establish with overwhelming probability that the observed frequencies (which will be close to 25% and 75%) are not chance fluctuations away from expected frequencies of 33% and 66%. (A million runs is more than enough for this purpose)..."

We want to replace the words "sufficiently", "overwhelming", "more than enough" with something more scientific. And as experiments are carried out which are harder and harder to prepare, it becomes important to design them so that they give conclusive results with the smallest possible sample sizes.

Two Paradigms of Statistics Matters are complicated by the fact that there are several statistical paradigms around, each prescribing different methods for designing experiments. The two prevailing paradigms are the 'frequentist' or 'orthodox' approach and the 'Bayesian' approach. Initial work in our direction has been done by [26] who adopts a Bayesian type of approach. In contrast[1], we provide a method which is neutral towards whatever kind of statistical paradigm one chooses, but which can be harnessed to provide useful information for designers of experiments, whether they will be frequentist or Bayesian. We choose an information–theoretic quantification, namely the Kullback–Leibler divergence (also known as *information deficiency* or *relative entropy* [8]). It turns out that, within our context, this notion captures the idea of 'statistical strength' both in the Bayesian and in the frequentist analysis. For a given type of experiment, we consider the game in which the experimenter wants to maximize the divergence, while the local theorist looks for theories, which minimize it. The experimenter's game space is the collection of probability distributions over joint settings, the local realist's game space is the space of local realistic theories. This game defines an experiment, such that each trial (assuming quantum mechanics is true) provides on average, the maximal support (both under the Bayesian and under the frequentist definition of 'support') for quantum theory against the best which local realism can provide, at those settings and for the corresponding true (quantum mechanical) probabilities. As will be explained, the amount of support provided on average by each trial, can be converted into numbers which a frequentist or a Bayesian can understand.

Statistical Strength of Nonlocality Proofs Since we can compare the statistical strength of each type of experiment, we can conclude by determining whether the Bell, CHSH, Hardy, GHZ, or Mermin proof, can yield most strong experimental evidence against local realism. Moreover, we can search more widely, for the best 'CHSH style proof', for instance, whereby we now also vary the settings or the quantum state, to produce the best game for QM. Similarly one could search more widely among 'GHZ style proofs'. It turns out that the original CHSH proof is much stronger than the original Hardy proof. Which CHSH-style proof is strongest, remains to be determined.

The GHZ proof was the first of a new class of proofs of Bell's theorem, "without inequalities". It specifies a state and collection of settings, such that certain QM probabilities are all zero or one, while this is impossible under LR. Now we would argue that this proof contains a hidden inequality, which is actually much more important. Suppose one could approximate the QM probabilities of zero and one, arbitrarily well, by appropriate LR theories, even if no LR theory would exactly reproduce the zero's and ones. Then no amount of experimentation could ever strictly rule out LR. However it is a fact, that there exists a positive ϵ such that any local realist theory which comes within ϵ of all the equalities but one, is forced to deviate by more than ϵ in the last. Thus, accompanying the GHZ style proof without inequalities, is an implied inequality, and it is this latter inequality that can be tested experimentally.

It should be noted that our approach is nonsymmetric between quantum mechanics and local realism. There is only one quantum theory, and we believe in it, but we must arm ourselves against any and all local realists. We leave the corresponding analysis for a local realist who believes in her theory, to that person to develop.

Related Work by Peres Earlier work by Peres [26] describes the same viewpoint on the strength of nonlocality proofs. Our work differs in that we allow the experimentalist to optimize her experimental settings, whereas [26] assumes that the frequencies over the measurement settings is uniform.

2 Formal Setup

A basic nonlocality proof ("quantum mechanics violates local realism") has the following ingredients. There are two parties A and B, who can each dispose over one of two entangled qubits. They may each choose out of two different measurement settings; let us denote these by "m_0" and "m_1". In each trial of the experiment, one randomly samples from the four different joint settings. In each trial, A and B each observe one of two different binary outcomes, say "**F**" (false) and "**T**" (true). Quantum mechanics enables us to compute the joint probability distribution of the outcomes, as function of the measurement settings and of the joint state of the two qubits. Thus possible design choices are: the state of the qubits, the values of the settings; and the probability distribution over the settings. More complicated experiments may involve more parties, more settings, and more outcomes. Such a generalized setting is formalized in Appendix A. In the main text, we focus on the basic $2 \times 2 \times 2$ case.

[1] Our work also differs from and extends [26] in a number of other ways; see the introduction of this article.

Now under local realism it is possible to talk about "the outcome which A would have observed, if she had used setting 1", independently of which setting was used by B and indeed of whether or not A actually did use setting 1 or 2. Thus we have four binary variables, which we will call X_1, X_2, Y_1 and Y_2. Here, as in the remainder of this paper, variables named X correspond to A's observations, and variables named Y correspond to B's observations. Thus, "$Y_1 = \mathbf{F}$" denotes that "if B had chosen measurement setting 1, he would have observed outcome $\mathbf{F}$".

Below we introduce notation for all ingredients involved in nonlocality proofs.

2.1 Frequency of Measurement Settings

Random variable A denotes the measurement setting at party A. Random variable B denotes the measurement setting at party B. Both A and B take values in $\{1, 2\}$. QM and LR must agree on the distribution σ on (A, B), giving the probabilities (and, after many repetitions of the experiment, the frequencies) witch which each measurement setting is sampled. The distribution σ is identified with its probability vector $\sigma = (\sigma_{11}, \sigma_{12}, \sigma_{21}, \sigma_{22}) \in \Sigma$, and Σ is the unit simplex in $\mathbb{R}^4$ defined by

$$\Sigma \quad := \quad \left\{ (\sigma_{11}, \sigma_{12}, \sigma_{21}, \sigma_{22}) \mid \sum_{a,b \in \{1,2\}} \sigma_{ab} = 1, \text{for all } a, b : \sigma_{ab} \geq 0 \right\}. \tag{1}$$

We use Σ^{UC} to denote the set of vectors representing *uncorrelated* distributions in Σ. Formally, $\sigma \in \Sigma^{\text{UC}}$ if and only if $\sigma_{ab} = (\sigma_{a1} + \sigma_{a2})(\sigma_{1b} + \sigma_{2b})$ for all $a, b \in \{1, 2\}$.

2.2 Measurement Outcomes

Random variable X denotes the set of possible experimental outcomes at party A; random variable Y denotes this set at party B. X and Y take values in $\{\mathbf{F}, \mathbf{T}\}$, $\mathbf{F}$ standing for 'false' and $\mathbf{T}$ standing for 'true'. Thus, the statement '$X = \mathbf{F}, Y = \mathbf{T}$' and denotes the event that party A observed $\mathbf{F}$ and party B observed $\mathbf{T}$.

The distribution of (X, Y) depends on the chosen setting $(a, b) \in \{1, 2\}^2$. Hence, the state of the entangled qubits determines four conditional distributions $Q_{11}, Q_{12}, Q_{21}, Q_{22}$ for (X, Y), one for each joint measurement setting. Q_{ab} is the distribution of (X, Y) given that measurement setting (a, b) has been chosen. For example, $Q_{ab}(X = \mathbf{F}, Y = \mathbf{T})$ abbreviated to $Q_{ab}(\mathbf{F}, \mathbf{T})$, denotes the probability that party A observes $\mathbf{F}$ and party B observes $\mathbf{T}$, given that the device of A is in setting a and the device of b is in setting b. According to QM, the outcome (X, Y) of a single experiment is then distributed as Q_σ defined by $Q_\sigma(X = x, Y = y, A = a, B = b) := \sigma_{ab} Q_{ab}(X = x, Y = y)$.

2.3 Definition of Nonlocality Proof

A *non-locality proof* for 2 parties, 2 measurement settings per party, and 2 outcomes, is identified with an entangled quantum state of two qubits (realized, by, e.g., two photons) and two measurement devices (e.g., polarization filters with some fixed orientation angles). Everything about the quantum state and the measurement devices that is relevant for the probability distribution on outcomes of experiments can be summarized by the four distributions Q_{ab} on (X, Y), $(a, b) \in \{1, 2\}^2$. Henceforth, we will simply *identify* a $2 \times 2 \times 2$ non-locality proof with the vector of distributions $Q = (Q_{11}, Q_{12}, Q_{21}, Q_{22})$.

This definition can be extended in an entirely straightforward manner to settings with more than two outcomes, parties and measurement settings per party. A formal definition can be found in Appendix A

We call a non-locality proof $Q = (Q_{11}, Q_{12}, Q_{21}, Q_{22})$ *proper* if and only if it violates local realism, i.e. if there exists no local realist distribution π such that $P_{ab;\pi}(\cdot) = Q_{ab}(\cdot)$ for all $(a, b) \in \{1, 2\}^2$.

2.4 Local Realist Theories

The local realist (LR) may provide any 'local' theory she likes to explain the results of the experiments. According to LR, each experiment determines values for the four random variables (X_1, X_2, Y_1, Y_2). For $a \in \{1, 2\}$, $X_a \in \{\mathbf{F}, \mathbf{T}\}$ denotes the outcome that party A would have observed if the measurement setting at A had been a. Similarly, for $b \in \{1, 2\}$, $Y_b \in \{\mathbf{F}, \mathbf{T}\}$ denotes the outcome that party B would have observed if the measurement setting at B had been b.

A local theory π may be viewed as a probability distribution for (X_1, X_2, Y_1, Y_2). Formally, we define π as a 16-dimensional probability vector with indices $(x_1, x_2, y_1, y_2) \in \{\mathbf{F}, \mathbf{T}\}^4$. By definition, $P_\pi(X_1 = x_1, X_2 = x_2, Y_1 = y_1, Y_2 = y_2) := \pi_{x_1 x_2 y_1 y_2}$. For example, $\pi_{\mathbf{FFFF}}$ denotes LR's probability that, in all possible measurement settings, A and B would both have observed $\mathbf{F}$. The set of local theories can thus be identified with the unit simplex in $\mathbb{R}^{16}$, which we will denote by Π.

Recall that the quantum state of the entangled qubits determines four distributions over measurement outcomes $Q_{ab}(X = \cdot, Y = \cdot)$, one for each joint setting $(a, b) \in \{1, 2\}^2$. Similarly, each LR theory $\pi \in \Pi$ determines four distributions $P_{ab;\pi}(X = \cdot, Y = \cdot)$. These are the marginal distributions, according to the local realist theory π, of random variables (X, Y) given that setting (a, b) has been chosen. The value $P_{ab;\pi}(X = \cdot, Y = \cdot)$ is defined as:

$$P_{ab;\pi}(X = x, Y = y) \quad := \sum_{\substack{x_1,x_2,y_1,y_2 \in \{\mathbf{F},\mathbf{T}\} \\ x_a = x; y_b = y}} \pi_{x_1 x_2 y_1 y_2}. \tag{2}$$

According to LR, the outcome of a single experiment is then distributed as $P_{\sigma;\pi}$ defined by $P_{\sigma;\pi}(X = x, Y = y, A = a, B = b) := \sigma_{ab} P_{ab;\pi}(X = x, Y = y)$.

3 The Nonlocality Proofs

In this section we briefly describe the five celebrated nonlocality proofs for which we will compute the statistical strength. In Appendix C, we provide further details about the entangled quantum states that give rise to the violations of the various inequalities.

Let us interpret the measurement outcomes $\mathbf{F}$ and $\mathbf{T}$ in terms of Boolean logic, i.e. $\mathbf{F}$ is "false" and $\mathbf{T}$ is "true". We can then use Boolean expressions such as $X_2 \& Y_2$, which evaluates to true whenever both X_2 and Y_2 evaluate to 'true', i.e. when both $X_2 = \mathbf{T}$ and $Y_2 = \mathbf{T}$. We derive the proofs by applying the rule that if the event $X = \mathbf{T}$ implies the event $Y = \mathbf{T}$ (in short "$X \implies Y$"), then $\Pr(X) \leq \Pr(Y)$. In similar vein, we will use rules like $\frac{1}{2}[\Pr(X) + \Pr(Y)] \leq \Pr(X \vee Y) \leq \Pr(X) + \Pr(Y)$ and $1 - \Pr(\neg X) - \Pr(\neg Y) \leq 1 - \Pr(\neg X \vee \neg Y) = \Pr(X \& Y) \leq \frac{1}{2}[\Pr(X) + \Pr(Y)]$.

As an aside we want to mention that the proofs of Bell, CHSH and Hardy all contain the following argument, which can be traced back to the nineteenth century logician Boole (1854). Consider four events such that $\neg B \cap \neg C \cap \neg D \implies \neg A$. Then it follows that $A \implies B \cup C \cup D$. And from this, it follows that $\Pr(A) \leq \Pr(B) + \Pr(C) + \Pr(D)$. In the CHSH argument and the Bell argument, the events concern the equality or inequality of one of the X_i with one of the Y_j. In the Hardy argument, the events concern the joint equality or inequality of one of the X_i, one of the Y_j, and a specific value $\mathbf{F}$ or $\mathbf{T}$. We provide some extracts from Boole (1854) in Appendix B.

3.1 CHSH and Bell's Argument

For the CHSH argument one notes that the implication

$$[(X_1 = Y_1) \& (X_1 = Y_2) \& (X_2 = Y_1)] \quad \implies \quad (X_2 = Y_2) \tag{3}$$

is logically true, and hence $(X_2 \neq Y_2) \implies [(X_1 \neq Y_1) \vee (X_1 \neq Y_2) \vee (X_2 \neq Y_1)]$ holds. As a result, local realism implies the following CHSH inequality

$$\Pr(X_2 \neq Y_2) \quad \leq \quad \Pr(X_1 \neq Y_1) + \Pr(X_1 \neq Y_2) + \Pr(X_2 \neq Y_1), \tag{4}$$

which can be violated by many choices of settings and states under quantum theory. In the example that CHSH gave, the first probability equals $0.85\ldots$ and the latter three equal to $0.15\ldots$. The probability distribution that corresponds with CHSH's proof is as follows

Pr	$X_1 = \mathbf{T}$	$X_1 = \mathbf{F}$	$X_2 = \mathbf{T}$	$X_2 = \mathbf{F}$
$Y_1 = \mathbf{T}$	0.4267766953	0.0732233047	0.4267766953	0.0732233047
$Y_1 = \mathbf{F}$	0.0732233047	0.4267766953	0.0732233047	0.4267766953
$Y_2 = \mathbf{T}$	0.4267766953	0.0732233047	0.0732233047	0.4267766953
$Y_2 = \mathbf{F}$	0.0732233047	0.4267766953	0.4267766953	0.0732233047

. (5)

By the same line of reasoning as above, one obtains Bell's inequality

$$\Pr(X_1 = Y_1) \quad \leq \quad \Pr(X_2 \neq Y_2) + \Pr(X_2 \neq Y_1) + \Pr(X_1 + Y_2). \tag{6}$$

See Sections C.1 and C.2 in the appendix for how this inequality can be violated.

3.2 Hardy's Argument

Hardy noted the following: if $(X_2 \& Y_2)$ is true, and $(X_2 \implies Y_1)$ is true, and $(Y_2 \implies X_1)$ is true, then $(X_1 \& Y_1)$ is true. Thus $(X_2 \& Y_2)$ implies: $\neg(X_2 \implies Y_1)$ or $\neg(Y_2 \implies X_1)$ or $(X_1 \& Y_1)$. Therefore

$$\Pr(X_2 \& Y_2) \quad \leq \quad \Pr(X_2 \& \neg Y_1) + \Pr(\neg X_1 \& Y_2) + \Pr(X_1 \& Y_1). \tag{7}$$

On the other hand, according to quantum mechanics it is possible that the first probability is positive, in particular, equals 0.09, while the three other probabilities here are all zero. See Section C.4 in the appendix for the precise probabilities.

3.3 Mermin's Argument

Mermin's argument uses three settings on both sides of the two parties, thus giving the set of six events $\{X_1, Y_1, X_2, Y_2, X_3, Y_3\}$. First, observe that the three equalities in $(X_1 = Y_1)\&(X_2 = Y_2)\&(X_3 = Y_3)$ implies at least one of the three statements in $((X_1 = Y_2)\&(X_2 = Y_1)) \vee ((X_1 = Y_3)\&(X_3 = Y_1)) \vee ((X_2 = Y_3)\&(X_3 = Y_2))$. By the standard arguments that we used before, we see that

$$1 - \Pr(X_1 \neq Y_1) - \Pr(X_2 \neq Y_2) - \Pr(X_3 \neq Y_3) \quad \leq \quad \Pr((X_1 = Y_1)\&(X_2 = Y_2)\&(X_3 = Y_3)),$$

and that

$$\begin{aligned} \Pr\begin{pmatrix} ((X_1 = Y_2)\&(X_2 = Y_1)) \\ \vee \\ ((X_1 = Y_3)\&(X_3 = Y_1)) \\ \vee \\ ((X_2 = Y_3)\&(X_3 = Y_2)) \end{pmatrix} \quad & \leq \quad \begin{pmatrix} \Pr((X_1 = Y_2)\&(X_2 = Y_1)) \\ + \\ \Pr((X_1 = Y_3)\&(X_3 = Y_1)) \\ + \\ \Pr((X_2 = Y_3)\&(X_3 = Y_2)) \end{pmatrix} \\ & \leq \quad \frac{1}{2}\begin{pmatrix} \Pr(X_1 = Y_2) + \Pr(X_2 = Y_1) \\ + \\ \Pr(X_1 = Y_3) + \Pr(X_3 = Y_1) \\ + \\ \Pr(X_2 = Y_3) + \Pr(X_3 = Y_2) \end{pmatrix}. \end{aligned}$$

As a result we have the 'Mermin inequality'

$$1 \quad \leq \quad \sum_{i=1}^{3} \Pr(X_i \neq Y_i) + \frac{1}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^{3} \Pr(X_i = Y_j),$$

which gets violated by a state and measurement setting that has probabilities $\Pr(X_i \neq Y_i) = 0$ and $\Pr(X_i = Y_j) = \frac{1}{4}$ for $i \neq j$.

3.4 GHZ's Argument

In our initial CHSH story we restricted attention to situations where four probabilities concerned were equal. However, there is no reason to restrict ourselves to quantum states where this equality will be true, if the aim is to obtain maximal experimental support of the failure of local realism. Similarly there is no reason to restrict oneself, in a Hardy type story, to quantum states making specific outcomes have probability zero. Obviously, by widening our resources we will be able to find stronger proofs.

Starting with [13], GHZ, proofs against local realism have been based on systems of three or more qubits, on systems of higher-dimensional quantum systems, and on larger sets of measurements (settings) per particle. We will investigate what payoff results from expanding our resources in these ways. Each time we are allowed to search over a wider space we may be able to do better, though each time the actual experiment may become harder to set up in the laboratory.

Let $\oplus$ denote the exclusive or operation such that $X \oplus Y$ is true if and only if $X \neq Y$. Then the following implication must hold

$$((X_1 \oplus Y_2 = Z_2)\&(X_2 \oplus Y_1 = Z_2)\&(X_2 \oplus Y_2 = Z_1)) \implies (X_1 \oplus Y_1 = Z_1). \tag{8}$$

Now, by considering the contrapositive, we get

$$\Pr(X_1 \oplus Y_1 \neq Z_1) \quad \leq \quad \Pr((X_1 \oplus Y_2 \neq Z_2) \vee (X_2 \oplus Y_1 \neq Z_2) \vee (X_2 \oplus Y_2 \neq Z_1)). \tag{9}$$

And because $\Pr(X \oplus Y \neq Z) = \Pr(X \oplus Y \oplus Z)$ and the convexity of the $\vee$-operation, this gives us GHZ's inequality:

$$\Pr(X_1 \oplus Y_1 \oplus Z_1) \quad \leq \quad \Pr(X_1 \oplus Y_2 \oplus Z_2) + \Pr(X_2 \oplus Y_1 \oplus Z_2) + \Pr(X_2 \oplus Y_2 \oplus Z_1). \tag{10}$$

This inequality can be violated by a three way entangled state and measurement settings that give $\Pr(X_1 \oplus Y_1 \oplus Z_1) = 1$ and $\Pr(X_1 \oplus Y_2 \oplus Z_2) = \Pr(X_2 \oplus Y_1 \oplus Z_2) = \Pr(X_2 \oplus Y_2 \oplus Z_1) = 0$.

4 Kullback-Leibler Divergence and Statistical Strength

4.1 Kullback-Leibler Divergence

In this section we formally define our notion of 'statistical strength of a non-locality proof'. The notion will be based on the KL divergence, an information theoretic quantity which we now introduce. Let $\mathcal{Z}$ be an arbitrary finite set. For a distribution Q over $\mathcal{Z}$, $Q(z)$ denotes the probability of event $\{z\}$. For two (arbitrary) distributions Q and P defined over $\mathcal{Z}$, the Kullback-Leibler (KL) divergence between Q and P is defined as

$$D(Q\|P) \quad := \quad \sum_{z\in\mathcal{Z}} Q(z) \log \frac{Q(z)}{P(z)} \tag{11}$$

where the logarithm is taken here, as in the rest of the paper, to base 2. We use the conventions that, for $y > 0$, $y \log 0 = \infty$, and $0 \log 0 = \lim_{y\downarrow 0} y \log y = 0$.

The KL divergence is also known as relative entropy, cross-entropy, information deficiency or I-divergence. Introduced in [20], KL divergence has become a central notion in information theory, statistics and large deviation theory. A good reference is [8]. It is straightforward to show (using concavity of the logarithm and Jensen's inequality) that $D(Q\|P) \geq 0$ with equality if and only if $P = Q$; in this sense, KL divergence behaves like a distance. However, in general $D(P\|Q) \neq D(Q\|P)$, so formally $D(\cdot\|\cdot)$ is not a distance.

KL divergence expresses the average disbelief in P, when observing random outcomes Z from Q. Thus occasionally (with respect to Q) one observes an outcome Z which is much more likely under P than Q, but on average (with respect to Q), the outcomes are more likely under Q than P, as illustrated by the fact that $D(Q\|P) \geq 0$. In Appendix D.1 we provide several properties and examples of the KL divergence.

KL divergence has several different interpretations and applications. Below we focus on the interpretation we are concerned with in this paper: KL divergence as a measure of 'statistical closeness' in the context of statistical hypothesis testing.

KL Divergence and Statistical Strength in Simple Hypothesis Testing Let $Z_1, Z_2, \ldots$ be a sequence of random variables independently generated either by some distribution P or by some distribution Q with $Q \neq P$. Suppose we are given a sample (sequence of outcomes) $z_1, \ldots, z_n$. We want to perform a statistical test in order to find out whether the sample is from P or Q. Suppose that the sample is, in fact, generated by Q ('Q is true'). Then, given enough data, the data will with very high (Q-) probability be overwhelmingly more likely according to Q than according to P. That is, the data strongly suggest that they were sampled from Q rather than P. The 'statistical distance' between P and Q indicates *how strongly* or, equivalently, *how convincingly* data that are generated by Q will suggest that they are from Q rather than P. It turns out that this notion of 'statistical distance' between two distributions is precisely captured by the Kullback Leibler divergence $D(Q\|P)$, which can be interpreted as *the average amount of support in favor of Q and against P* per trial. The larger the KL divergence, the larger the amount of support per trial. It turns out that

1. For a fixed sample size n, the larger $D(Q\|P)$, the more support there will be in the sample $z_1, \ldots, z_n$ for Q versus P (with high probability under Q) .
2. For a pre-determined fixed level of support in favor of Q against P (equivalently, level of 'confidence' in Q, level of 'convincingness' of Q), we have that the larger $D(Q\|P)$, the smaller the sample size before this level of support is achieved (with high probability under Q).
3. If, based on observed data $z_1, \ldots, z_n$, an experimenter decides that Q rather than P must have generated the data, then, the larger $D(Q\|P)$, the larger the *confidence* the experimenter should have in this decision (with high probability under Q).

Below we state some formal results relating KL divergence to standard formal definitions of 'convincingness', 'support' and 'confidence' taken from mathematical statistics. A more intuitive and informal explanation of how KL divergence is related to these notions is found in Appendix D.3. In Appendix D.2 we explain why, contrary to what has sometimes been implicitly assumed, *absolute deviations* between probabilities can be quite *bad* indicators of statistical closeness.

KL Divergence and Statistical Strength in Composite Hypothesis Testing Observing a sample generated by Q or P and trying to infer whether it was generated by Q or P is called *hypothesis testing* in the statistical literature. A hypothesis is *simple* if it consists of a single probability distribution. A hypothesis is called *composite* if it consists of a *set* of distributions. The composite hypothesis '$\mathcal{P}$' should be interpreted as 'there exists a $P \in \mathcal{P}$ that generated the data'. Above, we related the KL divergence to statistical strength when testing two simple hypotheses against each other. Yet in most practical applications (and in this paper) the aim is to test two hypotheses, at least one of which is composite. For concreteness, suppose we want to test the distribution Q against the set of distributions $\mathcal{P}$. In this case, under some regularity conditions on $\mathcal{P}$ and $\mathcal{Z}$, the element $P \in \mathcal{P}$ that is *closest* in statistical divergence to Q determines the statistical strength of the test of Q against $\mathcal{P}$. Formally, for a set of distributions $\mathcal{P}$ on $\mathcal{Z}$ we define (as is customary, [8])

$$D(Q\|\mathcal{P}) \quad := \quad \inf_{P \in \mathcal{P}} D(Q\|P). \tag{12}$$

Analogously to $D(Q\|P)$, $D(Q\|\mathcal{P})$ may be interpreted as the *average amount of support in favor of Q and against $\mathcal{P}$ per trial*, if data are generated according to Q.

In our case, QM claims that data are generated by some distribution Q_σ. LR claims that data are generated by some $P \in \mathcal{P}_\sigma$, where $\mathcal{P}_\sigma := \{P_{\sigma;\pi} \ : \ \pi \in \Pi\}$. QM and LR agree to test the hypothesis Q_σ against $\mathcal{P}_\sigma$. QM, who knows that data are really generated according to Q_σ, wants to select σ in such a way that the average amount of support in favor of Q and against $\mathcal{P}$ is maximized. The previous discussion suggests that he should pick the $\sigma \in \Sigma^{\text{UC}}$ that *maximizes statistical strength* $D(Q_\sigma\|\mathcal{P}_\sigma)$. Below we show that this is (in some sense) also the optimal choice according to statistical theory. Indeed, in Section 4.3 we will define the statistical strength of Q as $\sup_{\sigma \in \Sigma^{\text{UC}}} D(Q_\sigma\|\mathcal{P}_\sigma)$.

4.2 Formal Connection between KL Divergence and Statistical Strength

We consider three methods for statistical hypothesis testing: *frequentist* hypothesis testing [28], *Bayesian* hypothesis [21] testing and *information-theoretic* hypothesis testing [22, 29]. Nearly all state-of-the-art, theoretically motivated statistical methodology falls in either the Bayesian or the frequentist categories. Frequentist hypothesis testing is the most common, the most taught in statistics classes and is the standard method in, for example, the medical sciences. Bayesian hypothesis testing is becoming more and more popular in, for example, econometrics and biological applications. While theoretically important, the information-theoretic methods are less used in practice and are added mainly because they lead to a very concrete interpretation of statistical strength in terms of *bits* of information.

We illustrate below that in all three approaches the KL divergence indeed captures the notion of 'statistical strength'. We consider the general situation with a sample $Z_1, Z_2, \ldots$, with the Z_i independently and identically distributed according to some Q_σ, Q_σ being some distribution over some finite set $\mathcal{Z}$. For each n, the first n outcomes are distributed according to the n-fold product distribution of Q_σ, which we shall also refer to as Q_σ. Hence $Q_\sigma(z_1, \ldots, z_n) = \prod_{i=1}^n Q_\sigma(z_i)$. The independence assumption also induces a distribution over the set $\mathcal{Z}^\infty$ of all infinite sequences[2] which we shall *also* refer as Q_σ.

We test Q_σ against a set of distributions $\mathcal{P}_\sigma$. Thus, Q_σ and $\mathcal{P}_\sigma$ may, but do not necessarily refer to quantum and local realist theories – the statements below hold more generally.

4.2.1 Frequentist Justification

In frequentist hypothesis testing, $\mathcal{P}_\sigma$ is called the *null-hypothesis* and Q_σ the alternative hypothesis. Frequentist hypothesis testing can be implemented in a number of different ways, depending on what *statistical test* one adopts. A statistical test is a procedure that, when input an arbitrary sample of arbitrary length, outputs a *decision*. The decision is either 'Q_σ generated the data' or '$\mathcal{P}_\sigma$ generated the

[2] Readers familiar with measure theory should note that throughout this paper, we tacitly assume that $\mathcal{Z}^\infty$ is endowed with a suitable σ-algebra such that all sets mentioned in this paper become measurable.

data'. The confidence in a given decision is determined by a quantity known as *the p-value*. This is a function of the actually observed outcomes $z_1, \ldots, z_n$. It is defined as

$$\texttt{p-value} \quad := \quad \sup_{P \in \mathcal{P}_\sigma} P(\text{“The test outputs: } Q_\sigma \text{ generated the data”}). \tag{13}$$

Large p-values mean small confidence: if, for the actual data, the test says “Q_σ” but the p-value is large, then this is not very convincing to someone who considers the possibility that some $P \in \mathcal{P}_\sigma$ has generated the data: the large p-value indicates that the test may very well have given the wrong answer.

We call a test asymptotically optimal for identifying Q_σ if, under the assumption that Q_σ generated the data, the p-value goes to 0 at the fastest possible rate. Now let us assume that Q_σ generates the data, and an optimal test is used. A well-known result due to Bahadur [1, Theorem 1] says that, under some regularity conditions on Q_σ and $\mathcal{P}_\sigma$, with Q_σ-probability 1, for all large n, the test outputs “Q_σ generated the data” and that

$$\texttt{p-value} \quad = \quad \mathrm{e}^{-nD(Q_\sigma \| \mathcal{P}_\sigma) + o(n)}. \tag{14}$$

where $\lim_{n \to \infty} o(n)/n = 0$. We say ‘the p-value is determined, *to first order in the exponent*, by $D(Q_\sigma \| \mathcal{P}_\sigma)$’. It turns out that the regularity conditions, needed for Equation 14 to hold, apply when Q_σ is instantiated to a quantum theory Q with measurement setting distributions σ, and $\mathcal{P}_\sigma$ is instantiated to the corresponding set of LR theories as defined in Section 2.

Now imagine that QM, who knows that Q_σ generates the data, wonders whether to use the experimental setup corresponding to σ_1 or σ_2. Suppose that $D(Q_{\sigma_1} \| \mathcal{P}_{\sigma_1}) > D(Q_{\sigma_1} \| \mathcal{P}_{\sigma_2})$. It follows from Equation 14 that if the experiment corresponding to σ_1 is performed, the p-value will go to 0 exponentially faster (in the number of trials) than if the experiment corresponding to σ_2 is performed. It therefore makes sense to say that ‘the statistical strength of the experiment corresponding to σ_1 is larger than the strength of σ_2’. This provides a frequentist justification of adopting $D(Q_\sigma \| \mathcal{P}_\sigma)$ as an indicator of statistical strength.

Remarks

1. Most hypothesis testing as it occurs in practice (in, e.g., the medical sciences) follows the *Neyman-Pearson* approach, in which the p-value is used only indirectly. Before observing the outcomes, one agrees on a *significance level* ϵ. Typically choices are $\epsilon = 0.01$ or $\epsilon = 0.05$. One then observes the data $Z_1, \ldots, Z_n$, and one rejects $\mathcal{P}_\sigma$ (i.e., one adopts Q_σ) if the p-value for $Z_1, \ldots, Z_n$ turns out to be $\leq \epsilon$. Thus, in this form of hypothesis testing, the p-value denotes the minimum significance level at which $\mathcal{P}_\sigma$ is rejected.
2. Bahadur [1, Theorem 2] also provides a variation of Equation 14, which (roughly speaking) says the following: suppose Q_σ generates the data. For $\epsilon > 0$, let N_ϵ be the *minimum number of observations* such that, for all $n \geq N_\epsilon$, the test rejects $\mathcal{P}_\sigma$ (if $\mathcal{P}_\sigma$ is not rejected for infinitely many n, then N_ϵ is defined to be infinite). Suppose that an optimal (in the sense we used previously) test is used. Then, for small ϵ, N_ϵ is inversely proportional to $D(Q_\sigma \| \mathcal{P}_\sigma)$: with Q_σ-probability 1, the smaller $D(Q_\sigma \| \mathcal{P}_\sigma)$, the larger N_ϵ. If a ‘non-optimal’ test is used, then N_ϵ can only be larger, never smaller.

The rate at which the p-value of a test converges to 0 is known in statistics as *Bahadur efficiency*. For an overview of the area, see [15]. For an easy introduction to the main ideas, focusing on ‘Stein's lemma’ (a theorem related to Bahadur's), see [3, Chapter 12, Section 8]. For an introduction to Stein's lemma with a physicist audience in mind, see [2].

4.2.2 Bayesian Justification

In the Bayesian approach to hypothesis testing [5, 21], when testing Q_σ against $\mathcal{P}_\sigma$, we must first determine an *a priori probability distribution* over Q_σ and $\mathcal{P}_\sigma$. This distribution over distributions is usually just called ‘the prior’. It can be interpreted as indicating the prior (i.e., before seeing the data) ‘degree of belief’ in Q_σ vs. $\mathcal{P}_\sigma$. It is often used to incorporate prior knowledge into the statistical decision process. In order to set up the test as fairly as possible, QM and LR may agree to use the prior $\Pr(Q_\sigma) = \Pr(\mathcal{P}_\sigma) = 1/2$ (this should be read as ‘the prior probability that Q_σ obtains is equal to the prior probability that some $P \in \mathcal{P}_\sigma$ obtains’). Yet as long as both $\Pr(Q_\sigma), \Pr(\mathcal{P}_\sigma) > 0$ the specific values for the priors will be irrelevant for the result below.

For given prior probabilities and a given sample $z_1, \ldots, z_n$, Bayesian statistics provides a method to compute the *posterior probabilities* of the two hypotheses, *conditioned* on the observed data: $\Pr(Q_\sigma)$ is transformed into $\Pr(Q_\sigma \mid z_1, \ldots, z_n)$. Similarly, $\Pr(\mathcal{P}_\sigma)$ is transformed to $\Pr(\mathcal{P}_\sigma \mid z_1, \ldots, z_n)$. One then adopts the hypothesis $H \in \{Q_\sigma, \mathcal{P}_\sigma\}$ with the larger posterior probability $\Pr(H \mid z_1, \ldots, z_n)$. The confidence in this decision is given by the *posterior odds* of Q_σ against $\mathcal{P}_\sigma$, defined, for given sample $z_1, \ldots, z_n$, as

$$\texttt{post-odds}(Q_\sigma, \mathcal{P}_\sigma) \quad := \quad \frac{\Pr(Q_\sigma \mid z_1, \ldots, z_n)}{\Pr(\mathcal{P}_\sigma \mid z_1, \ldots, z_n)}. \tag{15}$$

The larger `post-odds`, the larger the confidence. Now suppose that data are distributed according to Q_σ. It can be shown that, under some regularity conditions on Q_σ and $\mathcal{P}_\sigma$, with Q_σ-probability 1,

$$\texttt{post-odds} \quad = \quad \mathrm{e}^{nD(Q_\sigma \| \mathcal{P}_\sigma) + O(\log n)}, \tag{16}$$

In our previously introduced terminology, 'the Bayesian confidence (posterior odds) is determined by $(Q_\sigma \| \mathcal{P}_\sigma)$, up to first order in the exponent'. We may now reason exactly as in the frequentist case to conclude that it makes sense to adopt $D(Q_\sigma \| \mathcal{P}_\sigma)$ as an indicator of statistical strength, and that it makes sense for QM to choose the setting probabilities σ so as to maximize $D(Q_\sigma \| \mathcal{P}_\sigma)$.

Equation 16 is a 'folklore result' which 'usually' holds. In Appendix D.4, we show that it does indeed holds with Q_σ and $\mathcal{P}_\sigma$ defined as nonlocality proofs and local realist theories, respectively.

4.2.3 Information-Theoretic Justification

There exist several approaches to information-theoretic or compression-based hypothesis testing; see, for example, [3, 22]. The most influential of these is the so-called *Minimum Description Length Principle* [29]. The basic idea is always that the more one can compress a given sequence of data, the more regularity one has extracted from the data, and thus, the better one has captured the 'underlying regularities in the data'. Thus, the hypothesis that allows for the maximum compression of the data should be adopted.

Let us first consider testing a simple hypothesis Q against another simple hypothesis P. Two basic facts of coding theory say that

1. There exists a uniquely decodeable code with lengths L_Q that satisfy, for all $z_1, \ldots, z_n \in \mathcal{Z}^n$,

$$L_Q(z_1, \ldots, z_n) \quad = \quad \lceil -\log Q(z_1, \ldots, z_n) \rceil. \tag{17}$$

   The code with lengths L_Q is called the *Shannon-Fano code*, and its existence follows from the so-called *Kraft Inequality*, [8].

2. If data $Z_1, \ldots, Z_n$ are independently identically distributed $\sim Q$, then among all uniquely decodeable codes, the code with length function L_Q has the shortest expected code-length. That is, let L be the length function of any uniquely decodeable code over n outcomes, then

$$E_Q[L(Z_1, \ldots, Z_n)] \quad \geq \quad E_Q[-\log Q(Z_1, \ldots, Z_n)]. \tag{18}$$

Thus, under the assumption that Q generated the data, the optimal (maximally compressing) code to use will be the Shannon-Fano code with lengths $-\log Q(Z^n)$ (here, as in the remainder of this section, we ignored the integer requirement for code lengths). Similarly, under the assumption that some P with $P \neq Q$ generated the data the optimal code will be the code with lengths $-\log P(Z^n)$. Thus, from the information-theoretic point of view, if one wants to find out whether P or Q better explains the data, one should check whether the optimal code under P or the optimal code under Q allows for more compression of the data. That is, one should look at the difference

$$\texttt{bit-diff} \quad := \quad -\log P(z_1, \ldots, z_n) - [-\log Q(z_1, \ldots, z_n)]. \tag{19}$$

If `bit-diff` > 0, then one decides that Q better explains the data. The confidence in this decision is given by the magnitude of `bit-diff`: the larger `bit-diff`, the more extra bits one needs to encode the data under P rather than Q, thus the larger the confidence in Q.

Now suppose that Q actually generates the data. The *expected code length difference*, measured in bits, between coding the data using the optimal code for Q and coding using the optimal code for P, is given by $E_Q[-\log P(Z^n) - [-\log Q(Z^n)]] = nD(Q\|P)$. Thus, the KL divergence can be interpreted as *the expected additional number of bits needed to encode outcomes generated by Q, if outcomes are encoded using a code that is optimal for P rather than for Q* . Thus, the natural 'unit' of $D(\cdot\|\cdot)$ is the 'bit', and $D(Q\|P)$ may be viewed as 'average amount of information about Z that is lost if Z is wrongfully regarded as being distributed by Q rather than P'. By the law of large numbers, Equation 19 implies that, with Q-probability 1, as $n \to \infty$,

$$\frac{1}{n}\texttt{bit-diff} \quad \to \quad D(Q\|P). \tag{20}$$

Thus, if Q generates the data, then the information-theoretic confidence `bit-diff` in decision "Q explains the data better than P" is, up to first order, determined by the KL divergence between Q and P: the larger $D(Q\|P)$, the larger the confidence. This gives an information-theoretic justification of the use of the KL divergence as an indicator of statistical strength for simple hypothesis testing. We now turn to composite hypothesis testing.

Composite Hypothesis Testing If one compares Q_σ against a set of hypotheses $\mathcal{P}_\sigma$, then one has to associate $\mathcal{P}_\sigma$ with a code that is 'optimal under the assumption that some $P \in \mathcal{P}_\sigma$ generated the data'. It turns out that there exist codes with lengths $L_{\mathcal{P}}$ satisfying, for all $z_1, \ldots, z_n \in \mathcal{Z}^n$,

$$L_{\mathcal{P}_\sigma}(z_1, \ldots, z_n) \quad \leq \quad \inf_{P \in \mathcal{P}_\sigma} - \log P(z_1, \ldots, z_n) + O(\log n). \tag{21}$$

An example of such a code is given in Appendix D.5. The code $L_{\mathcal{P}_\sigma}$ is optimal, up to logarithmic terms, for whatever distribution $P \in \mathcal{P}_\sigma$ that might actually generate data. The information theoretic approach to hypothesis testing now tells us that, to test Q_σ against $\mathcal{P}_\sigma$, we should compute the difference in code lengths

$$\texttt{bit-diff} \quad = \quad L_{\mathcal{P}_\sigma}(z_1, \ldots, z_n) - [-\log Q_\sigma(z_1, \ldots, z_n)]. \tag{22}$$

The larger this difference, the larger the confidence that Q_σ rather than $\mathcal{P}_\sigma$ generated the data. In Appendix D.5 we show that, in analogy to Equation 20, as $n \to \infty$,

$$\frac{1}{n}\texttt{bit-diff} \quad \to \quad D(Q_\sigma\|\mathcal{P}_\sigma) \tag{23}$$

Thus, up to sublinear terms, the information-theoretic confidence in Q_σ is given by $nD(Q_\sigma\|\mathcal{P}_\sigma)$. This provides an information-theoretic justification of adopting $D(Q_\sigma\|\mathcal{P}_\sigma)$ as an indicator of statistical strength.

4.3 Formal Definition of Statistical Strength

In this section, Q_σ denotes a nonlocality proof equipped with setting probabilities σ, and $\mathcal{P}_\sigma$ is (as in Section 2) defined as the set of corresponding local theories: $\mathcal{P}_\sigma := \{P_{\sigma;\pi} \ : \ \pi \in \Pi\}$. The discussion of the previous section implies that

1. For given probabilities over the settings σ, the statistical strength of Q_σ against $\mathcal{P}_\sigma$ is given by $D(Q_\sigma\|\mathcal{P}_\sigma)$.
2. This strength is maximized if the distribution σ is picked that maximizes $D(Q_\sigma\|\mathcal{P}_\sigma)$.

This leads to the following definition:

Definition 1 *The* statistical strength of non-locality proof Q *is denoted by* s_Q *and defined as*

$$\begin{aligned} \mathrm{s}_Q \quad &:= \quad \sup_{\Sigma^{\text{UC}}} D(Q_\sigma\|\mathcal{P}_\sigma) && (24) \\ &= \quad \sup_{\Sigma^{\text{UC}}} \inf_{\Pi} D(Q_\sigma\|P_{\sigma,\pi}), && (25) \end{aligned}$$

where here, as in the remainder of the paper, the subscript $\sigma \in \Sigma^{\text{UC}}$ is abbreviated to Σ^{UC}, and $\pi \in \Pi$ is abbreviated to Π. Note that we only maximize over *uncorrelated* distributions σ: if we would allow the probability of measurement setting at party A to be dependent on the measurement setting at party B, this would defeat the purpose of the experiment. Is Definition 1 the right characterization of statistical strength? There are several issues here. We defer discussion of all these issues to Sections 6; for the time being we simply assume that Definition 1 is appropriate. We proceed to list some essential properties of s_Q. We say that "nonlocality proof Q is *absolutely continuous with respect to local realist theory* π" if and only if for all $a, b \in \{1,2\}, x, y \in \{\mathbf{F}, \mathbf{T}\}$, it holds that if $Q_{ab}(x,y) > 0$ then $P_{ab;\pi}(x,y) > 0$.

Theorem 1 *Let Q be a given (not necessarily* $2 \times 2 \times 2$*) nonlocality proof and* Π *the corresponding set of local realist theories.*

1. *Let* $U(\sigma, \pi) := D(Q_\sigma\|P_{\sigma;\pi})$*, then:*

(a) For a $2 \times 2 \times 2$ proof, we have that

$$U(\sigma, \pi) = \sum_{a,b \in \{1,2\}} \sigma_{ab} D(Q_{ab}(\cdot) \| P_{ab;\pi}(\cdot)) \tag{26}$$

Hence, the KL divergence $D(Q_\sigma \| P_{\sigma;\pi})$ may alternatively be viewed as the average KL divergence between the distributions of (X, Y), where the average is over the settings (A, B). For a generalized nonlocality proof, the analogous generalization of Equation 26 holds.

(b) For fixed σ, $U(\sigma, \pi)$ is convex and lower semicontinuous on Π, and continuous and differentiable on the interior of Π.

(c) If Q is absolutely continuous with respect to some fixed π, then $U(\sigma, \pi)$ is linear in σ.

2. Let

$$U(\sigma) := \inf_{\pi \in \Pi} U(\sigma, \pi), \tag{27}$$

then

(a) For all $\sigma \in \Sigma$, the infimum in Equation 27 is achieved for some π^*.

(b) The function $U(\sigma)$ is nonnegative, bounded, concave and continuous on σ.

(c) If Q is not a proper nonlocality proof, then for all $\sigma \in \Sigma, U(\sigma) = 0$. If Q is a proper nonlocality proof, then $U(\sigma) > 0$ for all *σ in the interior of Σ.*

(d) For a 2 party, 2 measurement settings per party nonlocality proof, we further have that, even if Q is proper, then still $U(\sigma) = 0$ for all σ on the boundary of Σ.

3. Suppose that σ is in the interior of Σ, then:

(a) Let Q be a $2 \times 2 \times 2$ nonlocality proof. Suppose that Q is non-trivial in the sense that, for some a, b, Q_{ab} is not a point mass (i.e. $0 < Q_{ab}(x, y) < 1$ for some x, y). Then $\pi^* \in \Pi$ achieves the infimum in Equation 27 if and only if the following 16 (in)equalities hold:

$$\sum_{a,b \in \{1,2\}} \sigma_{ab} \frac{Q_{ab}(x_a, y_b)}{P_{ab;\pi^*}(x_a, y_b)} = 1 \tag{28}$$

for all $(x_1, x_2, y_1, y_2) \in \{\mathbf{F}, \mathbf{T}\}^4$ such that $\pi^*_{x_1,x_2,y_1,y_2} > 0$, and

$$\sum_{a,b \in \{1,2\}} \sigma_{ab} \frac{Q_{ab}(x_a, y_b)}{P_{ab;\pi^*}(x_a, y_b)} \leq 1 \tag{29}$$

for all $(x_1, x_2, y_1, y_2) \in \{\mathbf{F}, \mathbf{T}\}^4$ such that $\pi^*_{x_1,x_2,y_1,y_2} = 0$.
For generalized nonlocality proofs, $\pi^* \in \Pi$ achieves Equation 27 if and only if the corresponding analogues of Equations 28 and 29 both hold.

(b) Suppose that π^* and π° both achieve the infimum in Equation 27. Then for all $x, y \in \{\mathbf{F}, \mathbf{T}\}$, $a, b \in \{1, 2\}$ with $Q_{ab}(x, y) > 0$, we have $P_{ab;\pi^*}(x, y) = P_{ab;\pi^\circ}(x, y) > 0$. In words, π^* and π° coincide in every measurement setting for every measurement outcome that has positive probability according to Q_σ, and Q is absolutely continuous with respect to π^* and π°.

The proof is in Appendix E.

In general, $\inf_\Pi U(\sigma, \pi)$ may be achieved for several, different π. By part 2 of the theorem, these must induce the same four marginal distributions $P_{ab;\pi}$. It also follows directly from part 2 of the theorem that, for $2 \times 2 \times 2$ proofs, $\mathrm{s}_Q = \sup_{\sigma \in \Sigma^{\mathrm{UC}}} U(\sigma)$ is achieved for some $\sigma^* \in \Sigma^{\mathrm{UC}}$, where $\sigma^*_{ab} > 0$ for all $a, b \in \{1, 2\}$.

Computing Statistical Strength s_Q The question remains how to compute s_Q for given nonlocality proofs. In some special cases, we can make an educated guess of the σ^* achieving $\sup_{\sigma \in \Sigma^{\mathrm{UC}}} D(Q_\sigma \| \mathcal{P}_\sigma)$, and then verify it using Theorem 1, part 3(a) and the game-theoretic tools which we will develop in Section 6. Whenever this is not possible, we have to resort to numerical optimization. By convexity of $U(\sigma, \pi)$ in π, and concavity of $U(\sigma)$ as defined in Theorem 1, this is computationally feasible[3]

[3] Interestingly, it turns out that computing $\inf_{\pi \in \Pi} U(\sigma, \Pi)$ is formally equivalent to computing the maximum likelihood in a well-known statistical missing data problem.

4.4 Asymptopia

A possible objection to our definition is that it is asymptotic in nature. KL divergence only gives a reliable indication of statistical strength (defined in terms of, e.g. p-values or posterior odds) if experiments are repeated more than n_0 times, for some n_0. This problem cannot be solved by saying that 'our results are only valid if the experiment is repeated at least n_0 times'. The reason is that the value n_0 at which the asymptotic regime is reached depends on σ in Q_σ. Nevertheless, we argue that the KL divergence is the only reasonable indicator of statistical strength, for the following two reasons:

1. Even though both the frequentist and the Bayesian results are asymptotic, they can be used to obtain bounds on frequentist p-values and Bayesian posterior odds in the non-asymptotic case. We have not developed this possibility any further here.
2. The KL divergence is the only way to obtain a clean definition of strength, independent of the statistical method (Bayesian, frequentist, information-theoretic) that one adopts: every reasonable method leads asymptotically to KL divergence. But non-asymptotically, different methods and different definitions of confidence (p-values, posterior odds) may lead to quite different results.

5 The Results

In this section we give the statistical strength of the various nonlocality proofs. We considered three scenarios for the sampling frequencies of the measurement settings:

Uniform settings: Each measurement setting is sampled with equal probability

Uncorrelated settings: The parties sample their individual measurement settings according to distribution that is uncorrelated with the sample distributions of the other parties ($\sigma \in \Sigma^{\mathrm{UC}}$)

Correlated settings: The parties sample the joint measurement settings in a way that allows correlated sample distributions ($\sigma \in \Sigma$).

Following the explanation and tables of Appendix C we get the following values.

Strength	Uniform	Uncorrelated	Correlated
Original Bell	0.0141597409	0.0158003672	0.0169800305
Optimized Bell	0.0177632822	0.0191506613	0.0211293952
CHSH	0.0462738469	0.0462738469	0.0462738469
Hardy	0.0278585182	0.0279816333	0.0280347655
Mermin	0.0157895843	0.0191506613	0.0211293952
GHZ	0.2075187496	0.2075187496	0.4150374993

(30)

We thus see that in two-party setting, the nonlocality proof of CHSH is much stronger than that of Bell, Hardy or Mermin, and that this optimal strength is obtained for uniform measurement settings. Furthermore it is clear that the three-party proof of GHZ is an order of magnitude stronger than the two-party case.

We also note that the nonlocality proof of Mermin—in the case of non-uniform settings—is equally strong as the optimized version of Bell's proof. The measurement frequencies tables in Appendix C.5 shows why this is the case: the optimal measurement settings for Mermin exclude one setting on A's side, and one setting on B's side, thus reducing Mermin's proof to that of Bell. One can view this is as an example of how a proof that is easier to understand (Mermin) is not necessarily stronger than one that has more subtle arguments (Bell).

6 Game-Theoretic Considerations

Let us consider the following variation of our basic scenario. Suppose that, before the experiments are actually conducted, LR has to decide on a *single* local theory π_0 (rather than the set Π) as an explanation of the outcomes that will be observed. QM then gets to see this π, and can choose σ depending on the π_0 that has been chosen. Since QM wants to maximize the strength of the experiment, he will pick the σ achieving $\sup_{\sigma \in \Sigma^{\mathrm{UC}}} D(Q_\sigma \| P_{\sigma;\pi_0})$. In such a scenario, the 'best' LR theory, minimizing statistical strength, is the LR theory π_0 that minimizes, over $\pi \in \Pi$, $\sup_{\sigma \in \Sigma^{\mathrm{UC}}} D(Q_\sigma \| P_{\sigma;\pi})$. Thus, in this slightly different setup, the statistical strength is determined by

$$\mathrm{s}'_Q \quad = \quad \inf_{\pi \in \Pi} \sup_{\sigma \in \Sigma^{\mathrm{UC}}} D(Q_\sigma \| P_{\sigma;\pi}) \qquad (31)$$

rather than $\mathrm{S}_Q = \sup_{\sigma \in \Sigma^{\mathrm{UC}}} \inf_{\pi \in \Pi} D(Q_\sigma \| P_{\sigma;\pi})$. Below we show that $\mathrm{S}'_Q \geq \mathrm{S}_Q$. We consider the definition S_Q to be preferable over S'_Q. The reason is that, as quantum experimenters, we should try to convince LR that QM is true *in a setting about which QM cannot complain.* Thus, if LR wants to entertain several local theories at the same time (use Π rather than π_0), or wants to have a look at the probabilities σ_{ab} before the experiment is conducted, we (QM) should allow him to do so—we will still be able to convince LR, even though we may need to repeat the experiment a few more times.

Nevertheless, it is quite useful to investigate under what conditions $\mathrm{S}_Q = \mathrm{S}'_Q$. As we will see, the answer will sometimes allow us to compute S_Q directly, without resorting to numerical optimization. Von Neumann's famous minimax theorem of game theory [24] suggests that

$$\sup_{\sigma \in \Sigma^*} \inf_{\pi \in \Pi} D(Q_\sigma \| P_{\sigma;\pi}) \quad = \quad \inf_{\pi \in \Pi} \sup_{\sigma \in \Sigma^*} D(Q_\sigma \| P_{\sigma;\pi}), \tag{32}$$

if Σ^* is a convex subset of Σ. Indeed, Theorem 2 below shows that Equation 32 holds if we take $\Sigma^* = \Sigma$. Unfortunately, Σ^{UC} is *not* convex, and Equation 32 does not hold in general for $\Sigma^* = \Sigma^{\mathrm{UC}}$, whence in general $\mathrm{S}_Q \neq \mathrm{S}'_Q$. Nevertheless, Theorem 3 provides some conditions under which Equation 32 does hold with $\Sigma^* = \Sigma^{\mathrm{UC}}$. In Section 6.3 we put this fact to use in computing S_Q for the CHSH nonlocality proof. But before presenting Theorems 2 and 3, we first need to introduce some game-theoretic terminology.

6.1 Game-Theoretic Definitions

A *statistical game* ([11]) is a triplet (A, B, L) where A and B are arbitrary sets and $L : A \times B \rightarrow \mathbb{R} \cup \{-\infty, \infty\}$ is a *loss function.* We say that the game has *value* V if

$$\begin{aligned} V &= \sup_{\alpha \in A} \inf_{\beta \in B} L(\alpha, \beta) & (33) \\ &= \inf_{\beta \in B} \sup_{\alpha \in A} L(\alpha, \beta). & (34) \end{aligned}$$

If for some $(\alpha^*, \beta^*) \in A \times B$ we have

$$\begin{array}{ll} \text{For all } \alpha \in A\text{:} & L(\alpha, \beta^*) \leq L(\alpha^*, \beta^*) \\ \text{For all } \beta \in B\text{:} & L(\alpha^*, \beta) \geq L(\alpha^*, \beta^*) \end{array}$$

then we call (α^*, β^*) a *saddle point* of the game. It is easily seen (Proposition 1, Appendix E) that, if α° achieves $\sup_{\alpha \in A} \inf_{\beta \in B} L(\alpha, \beta)$ and β° achieves $\inf_{\beta \in B} L(\alpha, \beta)$ and the game has value V, then $(\alpha^\circ, \beta^\circ)$ is a saddle point and $L(\alpha^\circ, \beta^\circ) = V$.

The Correlated Game With each non-locality proof we associate a corresponding *correlated game* which is just the statistical game defined by the triple (Σ, Π, U), where $U : \Sigma \times \Pi \rightarrow \mathbb{R} \cup \{\infty\}$ is defined by

$$U(\sigma, \pi) \quad := \quad D(Q_\sigma \| P_{\sigma;\pi}). \tag{35}$$

By the definition above, this game has a value V defined by

$$\begin{aligned} V &:= \inf_{\Pi} \sup_{\Sigma} U(\sigma, \pi) & (36) \\ &= \sup_{\Sigma} \inf_{\Pi} U(\sigma, \pi). & (37) \end{aligned}$$

We call the game *correlated* because we allow distributions σ over measurement settings to be such that the probability that party A is in setting a is correlated with (is dependent of) the probability that party B is in setting b.

The Uncorrelated Game Recall that we use Σ^{UC} to denote the set of vectors representing uncorrelated distributions in Σ. With each non-locality proof we can associate the game $(\Sigma^{\mathrm{UC}}, \Pi, U)$ which we call the corresponding *uncorrelated game.*

6.2 Game-Theoretic, Saddle Point Theorems

Theorem 2 (Saddle point for Potentially Correlated Settings) *For every (generalized) non-locality proof, the correlated game (Π, Σ, U) corresponding to it has a finite value, i.e. there exist a $0 \leq V < \infty$ with $\inf_\Pi \sup_\Sigma U(\sigma, \pi) = V = \sup_\Sigma \inf_\Pi U(\sigma, \pi)$. The infimum on the left is achieved for some $\pi^* \in \Pi$; the supremum on the right is achieved for some σ^* in Σ, so that (π^*, σ^*) is a saddle point.*

The proof is in Appendix E.

Remark In the information-theoretic literature, several well-known minimax and saddle point theorems involving the Kullback-Leibler divergence exist; we mention [19, 31]. However, all these deal with settings that are substantially different from ours.

In the case where there are two parties and two measurement settings per party, we can say a lot more.

Theorem 3 *Fix any proper non-locality proof based on 2 parties with 2 measurement settings per party and let* (Σ, Π, U) *and* $(\Sigma^{\text{UC}}, \Pi, U)$ *be the corresponding correlated and uncorrelated games, then:*

1. *The correlated game has a saddle point with value* $V > 0$*. Moreover,*

$$\sup_{\Sigma^{\text{UC}}} \inf_{\Pi} U(\sigma, \pi) \leq \sup_{\Sigma} \inf_{\Pi} U(\sigma, \pi) = V, \tag{38}$$

$$\inf_{\Pi} \sup_{\Sigma^{\text{UC}}} U(\sigma, \pi) = \inf_{\Pi} \sup_{\Sigma} U(\sigma, \pi) = V. \tag{39}$$

2. *Let*

$$\Pi^* := \{\pi : \pi \text{ achieves } \inf_{\Pi} \sup_{\Sigma} U(\sigma, \pi)\}, \tag{40}$$

$$\Pi^{\text{UC}*} := \{\pi : \pi \text{ achieves } \inf_{\Pi} \sup_{\Sigma^{\text{UC}}} U(\sigma, \pi)\}, \tag{41}$$

then

(a) Π^* *is non-empty.*

(b) $\Pi^* = \Pi^{\text{UC}*}$.

(c) All $\pi^* \in \Pi^*$ *are 'equalizer strategies', i.e. for all* $\sigma \in \Sigma, U(\sigma, \pi^*) = V$.

3. *The uncorrelated game has a saddle point if and only if there exists* (π^*, σ^*)*, with* $\sigma^* \in \Sigma^{\text{UC}}$*, such that*

(a) π^* *achieves* $\inf_\pi U(\sigma^*, \pi)$.

(b) π^* *is an equalizer strategy.*

If such (σ^*, π^*) *exists, it is a saddle point.*

The proof is in Appendix E.

6.3 Example Application of Game-Theoretic Arguments

Consider the CHSH nonlocality argument. The quantum distributions Q, given in the table in Section 3 have traditionally been compared with the local theory $\tilde{\pi}$ defined by

$$\tilde{\pi}_{\mathbf{FFFF}} = \tilde{\pi}_{\mathbf{TTTT}} = \tilde{\pi}_{\mathbf{FFFT}} = \tilde{\pi}_{\mathbf{TTTF}} = \tilde{\pi}_{\mathbf{FFTF}} = \tilde{\pi}_{\mathbf{TTFT}} = \tilde{\pi}_{\mathbf{TFFT}} = \tilde{\pi}_{\mathbf{FTTF}} = \tfrac{1}{8} \tag{42}$$

and $\tilde{\pi}_{x_1 x_2 y_1 y_2} = 0$ otherwise. This gives rise to the following probability table:

$P_{ab;\tilde{\pi}}$	$X_1 = \mathbf{T}$	$X_1 = \mathbf{F}$	$X_2 = \mathbf{T}$	$X_2 = \mathbf{F}$
$Y_1 = \mathbf{T}$	0.375	0.125	0.375	0.125
$Y_1 = \mathbf{F}$	0.125	0.375	0.125	0.375
$Y_2 = \mathbf{T}$	0.375	0.125	0.125	0.375
$Y_2 = \mathbf{F}$	0.125	0.375	0.375	0.125

. (43)

There exists no local theory which has uniformly smaller absolute deviations from the quantum probabilities in all four tables. Even though, in general, absolute deviations are not a good indicator of statistical strength, based on the fact that all four tables 'look the same', we may still *guess* that, for uniform measurement settings $\tilde{\sigma}_{ab} = 1/4$, $a, b \in \{1, 2\}$, the optimal local realist theory is given by the $\tilde{\pi}$ defined above. We can now use Theorem 1, part 3(a) to check our guess. Checking the 16 equations 28 and 29 shows that our guess was correct: $\tilde{\pi}$ achieves $\inf U(\sigma, \pi)$ for the uniform measurement settings $\tilde{\sigma}$. It is clear that $\tilde{\pi}$ is an equalizer strategy and that $\tilde{\sigma}$ is uncorrelated. But now Theorem 3, part (3) tells us that $(\tilde{\sigma}, \tilde{\pi})$ is a saddle point in the uncorrelated game. This shows that $\tilde{\sigma}$ achieves $\sup_{\sigma \in \Sigma^{\text{UC}}} \inf_{\pi \in \Pi} D(Q_\sigma \| P_\sigma)$. Therefore, the statistical strength of the CHSH nonlocality proof must be given by

$$\mathbb{S}_Q = \sup_{\sigma \in \Sigma^{\text{UC}}} \inf_{\pi \in \Pi} D(Q_\sigma \| P_\sigma) = D(Q_{\tilde{\sigma}} \| P_{\tilde{\sigma};\tilde{\pi}}) \tag{44}$$

which is straightforward to evaluate.

7 Acknowledgments

The authors are very grateful to Piet Groenenboom for providing us with the software [14] needed to compute $\inf_{P \in \mathcal{P}} D(Q \| P)$, and helping us to adjust the program for our purposes.

The authors would like to thank Eurandom for financial support. Part of this research was performed while Peter Grünwald was visiting the University of California at Santa Cruz (UCSC). The visit was funded by NWO (the Netherlands Organization for Scientific Research) and the Applied Math and Computer Science Departments of UCSC. Wim van Dam is funded by an HP-MSRI postdoctoral fellowship. Richard Gill's research was partially funded by project RESQ (IST-2001-37559) of the IST-FET programme of the European Union. Richard Gill is grateful for the hospitality of the Quantum Probability group at the department of mathematics of the University of Greifswald, Germany. His research there was supported by European Commission grant HPRN-CT-2002-00279, RTN QP-Applications.

References

[1] R.R. Bahadur, "An optimal property of the likelihood ratio statistic", In *Proc. Fifth Berkeley Symp. Math. Stat. Prob.,* Volume 1, pp. 13–26, 1967.

[2] V. Balasubramanian, "A Geometric Formulation of Occam's Razor for Inference of Parametric Distributions", available at `http://arxiv.org/`, *adap-org/9601001*, 1996.

[3] A. Barron and T. Cover, "Minimum complexity density estimation", *IEEE Transactions on Information Theory,* Volume 37(4), pp. 1034–1054, 1991.

[4] J.S. Bell, "On the Einstein-Podolsky-Rosen paradox", *Physics,* Volume 1, pp. 195–200, 1964.

[5] J.M. Bernardo and A.F.M. Smith, *Bayesian theory,* John Wiley, 1994.

[6] G. Boole, *An Investigation of the Laws of Thought (on which are founded the mathematical theories of logic and probabilities),* MacMillan and Co., Cambridge, 1854. Reprinted by Dover, 1951.

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

[8] T.M. Cover and J.A. Thomas, *Elements of Information Theory,* Wiley Interscience, New York, 1991.

[9] W. Feller, *An Introduction to Probability Theory and Its Applications,* Wiley, Volume 1, third edition, 1969.

[10] W. Feller, *An Introduction to Probability Theory and Its Applications,* Wiley, Volume 2, third edition, 1969.

[11] T.S. Ferguson, *Mathematical Statistics – a decision-theoretic approach,* Academic Press, 1967.

[12] T.L. Fine, *Theories of Probability,* Academic Press, 1973.

[13] D.M. Greenberger, M. Horne, and A. Zeilinger, "Going beyond Bell's theorem", In M. Kafatos, editor, *Bell's Theorem, Quantum Theory, and Conceptions of the Universe,* pp. 73–76, Kluwer, Dordrecht, 1989.

[14] P. Groeneboom, G. Jongbloed, and J.A. Wellner "Vertex direction algorithms for computing nonparametric estimates in mixture models", manuscript, 2002.

[15] P. Groeneboom and J. Oosterhoff, "Bahadur efficiency and probabilities of large deviations", *Statistica Neerlandica,* Volume 31, pp. 1–24, 1977.

[16] P.D. Grünwald, *The Minimum Description Length Principle and Reasoning under Uncertainty,* PhD thesis, University of Amsterdam, The Netherlands, October 1998. Available as ILLC Dissertation Series 1998-03.

[17] P. Grünwald and A.P. Dawid, "Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory", Technical Report 223, University College London. http://www.ucl.ac.uk/˜ucak06d/reports.html, 2002.

[18] L. Hardy, "Nonlocality for two particles without inequalities for almost all entangled states", *Physical Review Letters,* Volume 71, pp. 1665–1668, 1993.

[19] D. Haussler, "A general minimax result for relative entropy", *IEEE Transactions on Information Theory,* Volume 43(4), pp. 1276–1280.

[20] S. Kullback and R.A. Leibler, “On information and sufficiency”, *Annals of Mathematical Statistics,* Volume 22, pp. 76–86, 1951.

[21] P.M. Lee, *Bayesian Statistics – an introduction,* Arnold & Oxford University Press, 1997.

[22] M. Li and P.M.B. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications,* (revised and expanded Second edition), New York, Springer-Verlag, 1997.

[23] N.D. Mermin, “Quantum mysteries for anyone”, *J. Philos.,* Volume 781, pp. 397–408, 1981.

[24] J. Von Neumann, “Zur Theorie der Gesellschaftsspiele”, *Mathematische Annalen,* Volume 100, pp. 295–320, 1928.

[25] A. Peres, *Quantum Theory: Concepts and Methods,* Fundamental Theories of Physics, Volume 57, Kluwer Academic Publishers, 1995.

[26] A. Peres, “Bayesian analysis of Bell inequalities”, *Fortsch. Phys.,* Volume 48, pp. 531–535, 2000.

[27] E. Posner, “Random coding strategies for minimum entropy”, *IEEE Transactions on Information Theory,* Volume 21, pp. 388–91, 1975.

[28] J.A. Rice, *Mathematical Statistics and Data Analysis,* Duxbury Press, 1995.

[29] J. Rissanen, *Stochastic Complexity in Statistical Inquiry,* World Scientific Publishing Company, 1989.

[30] R.T. Rockafellar, *Convex Analysis,* Princeton University Press, Princeton, New Jersey, 1970.

[31] F. Topsøe, “Information-theoretical optimization techniques”, *Kybernetika,* Volume 15(1), 1989.

[32] V. Vapnik, *Statistical Learning Theory,* John Wiley, 1998.

A Beyond $2 \times 2 \times 2$: General Case of Non-Locality Proofs

Here we extend the $2 \times 2 \times 2$ setting to more than two parties, settings and outcomes. A general *non-locality proof* is defined as a tuple $(k, \mathcal{S}, \mathcal{X}, Q)$ where

1. k is the number of parties,
2. $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_k$ is the set of possible measurement settings.
   (a) $\mathcal{S}_j = \{1, 2, \ldots, N_j^s\}$ is the set of measurement settings for party j.
   (b) N_j^s is the number of settings at party j.
3. $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_k$ is the set of possible measurement outcomes.
   (a) $\mathcal{X}_j = \mathcal{X}_{(j,1)} \times \cdots \times \mathcal{X}_{(j,N_j^s)}$ is the set of measurement outcomes for party j.
   (b) $\mathcal{X}_{(j,s)} = \{1, 2, \ldots, N_{(j,s)}^x\}$ is the set of measurement outcomes for party j *when party j is in setting s.*
   (c) $N_{(j,s)}^x$ is the number of measurement outcomes for party j when party j is in setting s.
   (d) $(X_1, \ldots, X_k)$ are the random variables indicating the outcome at parties $1, 2, \ldots, k$.
4. $Q = (Q_{s_1 \ldots s_k} : (s_1, \ldots, s_k) \in \mathcal{S})$ is a list of all the distributions $Q_{s_1 \ldots s_k}(X_1 = \cdot, \ldots, X_k = \cdot)$, one for each joint measurement setting $(s_1, \ldots, s_k) \in \mathcal{S}$. These are the distributions on outcomes induced by the state that the quantum experimenter's entangled qubits are in.

To each non-locality proof $(k, \mathcal{S}, \mathcal{X}, Q)$ there corresponds a set of local realist distributions Π. Each such distribution is identified with its probability vector π. Formally, π is a distribution for the tuple of random variables

$$\begin{matrix} X_{(1,1)} & \cdots & X_{(1,N_1^s)} \\ \vdots & \ddots & \vdots \\ X_{(k,1)} & \cdots & X_{(1,N_k^s)} \end{matrix} \tag{45}$$

Here $X_{(j,s)}$ denotes LR's distribution of Z_j when party j's measurement device is in setting s.

Once again, we call a non-locality proof *proper* if and only if it violates local realism, i.e. if there exists no local realist distribution π such that $P_{s_1 \ldots s_k;\pi}(\cdot) = Q_{s_1 \ldots s_k}(\cdot)$ for all $(s_1, \ldots, s_k) \in \mathcal{S}$.

The definition of statistical strength remains unchanged.

B George Boole's Contribution

One of the contributions of [6] is "a fairly general approach, given the respective numbers of individuals comprised in any classes s, t, &c, being logically defined, to deduce a system of numerical limits of any other class, being logically defined". Interesting is also a footnote "The author regrets the loss of a manuscript written about four years ago in which this method he believes was developed at considerable length. His recollection of the contents is almost entirely confined to the *impression* that the principle of the method was the same as above described, and that its sufficiency was proved. The prior methods of this chapter are, it is almost needless to say, easier, though certainly less general".

Thus Boole does not claim that the methods he presents here, are absolutely general, though they are supposed to provide the least upper bound and greatest lower bound, when they can be applied. The method is good enough to reproduce the original "three variables" Bell inequality, for instance as an application of the last example of Boole's Chapter 18. Here, it is given that: $\Pr(x + (1-x)yz) = p$, $\Pr(y + (1-y)xz) = q$, and $\Pr(z + (1-z)xy) = r$, where x, y and z are (binary) logical variables. Boole's method results in the inequalities $1+p \geq q+r$, $1+q \geq p+r$, and $1+r \geq p+q$. Now, take x, y and z to be the events $X_1 = Y_1$, $Y_1 = X_2$, $X_2 = X_1$. Note that $(1-x)yz$ is impossible and so are the two other similar terms. We therefore may conclude $1 + \Pr(X_2 = X_1) \leq \Pr(X_1 = Y_1) + \Pr(Y_1 = X_2)$. The probability on the left hand side does not correspond to an empirically accessible event, but if we have two spin half particles in the singlet or Bell state, we know from quantum mechanics that $\Pr(X_1 \neq Y_1) = 1$. Hence we conclude

$$2 - \Pr(X_2 = Y_1) \quad \leq \quad \Pr(X_1 = Y_1) + \Pr(Y_1 = X_2). \tag{46}$$

C The Nonlocality Proofs and Their Optimal Settings

In this appendix we list the nonlocality proofs of Bell, an optimized version of Bell, CHSH, Hardy, Mermin and GHZ and their solutions. The proofs themselves are described by a multipartite quantum state and the measurement bases $|m_{\cdot}^{\cdot}\rangle$ of the parties. Because all bases are two dimensional in the proofs below, it is sufficient to only describe the vector $|m_{\cdot}^{\cdot}\rangle$, where it is understood that the other basis vector $(|\perp m_{\cdot}^{\cdot}\rangle)$ is the orthogonal one. Because of its frequent use, we define for the whole appendix the rotated vector $|R(\phi)\rangle := \cos(\phi)|0\rangle + \sin(\phi)|1\rangle$. A *measurement setting* refers to the bases that parties use during a trial of the experiment. All proofs, except Mermin's, have two different settings per party (in Mermin they have three).

Given the state and the measurement bases, the proof is summarized in a table of probabilities of the possible measurement outcomes. Here we list these probabilities conditionally on the specific measurement settings. For example, for Bell's original nonlocality proof, which uses the state $|\Psi\rangle := \frac{1}{\sqrt{2}}(|0_A 0_B\rangle + |1_A 1_B\rangle)$ and the measurement vectors $|X = \mathbf{T}\rangle_{a=1} := |R(0)\rangle$ and $|Y = \mathbf{T}\rangle_{b=1} := |R(\frac{\pi}{8})\rangle$, we list the probability $Q_{11}(X = \mathbf{T}, Y = \mathbf{T}) = |\langle\Psi|X = \mathbf{T}, Y = \mathbf{T}\rangle_{a=1,b=1}|^2 \approx 0.4268$ in the table.

As discussed in the article (Sections 2.1 and 4.3), the strength of a nonlocality proof will depend on the probabilities σ with which the parties use the different measurement settings. With respect to this we distinguish three different scenarios:

Uniform settings: Each measurement setting is sampled with equal probability

Uncorrelated settings: The parties sample their individual measurement settings according to distribution that is uncorrelated with the sample distributions of the other parties ($\sigma \in \Sigma^{\text{UC}}$)

Correlated settings: The parties sample the joint measurement settings in a way that allows correlated sample distributions ($\sigma \in \Sigma$).

For both the correlated and the uncorrelated settings, the parties can optimize their measurement frequencies to get the strongest possible statistics to prove the nonlocality of their measurement outcomes. We list these optimal frequencies below where, for example, $\Pr(a = 1)$ stands for the probability that party A uses the measurement basis $\{|(X = \mathbf{T}|a = 1)\rangle, |(X = \mathbf{F}|a = 1)\rangle\}$ and $\Pr(a = 1, b = 2)$ is the probability that A uses the basis $\{|(X = \mathbf{T}|a = 1)\rangle, |(X = \mathbf{F}|a = 1)\rangle\}$ while B uses the basis $\{|(Y = \mathbf{T}|b = 2)\rangle, |(Y = \mathbf{F}|b = 2)\rangle\}$, etc.

Associated with these optimal frequencies there is an optimal local theorist theory $\pi \in \Pi$ (see Section 4.3). The probabilities for such optimal classical theories are listed below as well and should be compared with the tables of the nonlocality proofs. Combining these data tables for each proof and each scenario we obtain the strengths that were listed in Section 5.

C.1 Original Bell

For Bell's proof of nonlocality we have to make a distinction between the original version, which Bell described [4], and the optimized version, which is described by Peres in [25].

First we discuss Bell's original proof. Take the bipartite state $\frac{1}{\sqrt{2}}|0_A 0_B\rangle + \frac{1}{\sqrt{2}}|1_A 1_B\rangle$, and the measurement settings

$$|X = \mathbf{T}\rangle_{a=1} := |R(0)\rangle \quad \text{and} \quad |X = \mathbf{T}\rangle_{a=2} := |R(\tfrac{\pi}{8})\rangle$$
$$|Y = \mathbf{T}\rangle_{b=1} := |R(\tfrac{\pi}{8})\rangle \quad \text{and} \quad |Y = \mathbf{T}\rangle_{b=2} := |R(\tfrac{\pi}{4})\rangle$$

With these settings, quantum mechanics predicts the following conditional probabilities.

		$a=1$		$a=2$	
$Q_{ab}(X=x,Y=y)$		$x=\mathbf{T}$	$x=\mathbf{F}$	$x=\mathbf{T}$	$x=\mathbf{F}$
$b=1$	$y=\mathbf{T}$	0.4267766953	0.0732233047	0.5	0
	$y=\mathbf{F}$	0.0732233047	0.4267766953	0	0.5
$b=2$	$y=\mathbf{T}$	0.25	0.25	0.4267766953	0.0732233047
	$y=\mathbf{F}$	0.25	0.25	0.0732233047	0.4267766953

with $\frac{1}{4} + \frac{1}{8}\sqrt{2} \approx 0.4267766953$ and $\frac{1}{4} - \frac{1}{8}\sqrt{2} \approx 0.0732233047$.

C.1.1 Uniform Settings, Original Bell

When the two parties use uniform frequencies for their measurement settings, the optimal classical theory is as follows.

		$a=1$		$a=2$	
$P_{ab}(X=x,Y=x)$		$x=\mathbf{T}$	$x=\mathbf{F}$	$x=\mathbf{T}$	$x=\mathbf{F}$
$b=1$	$y=\mathbf{T}$	0.3970311357	0.1029688643	0.5000000000	0.0000000000
	$y=\mathbf{F}$	0.1029688643	0.3970311357	0.0000000000	0.5000000000
$b=2$	$y=\mathbf{T}$	0.2940622714	0.2059377286	0.3970311357	0.1029688643
	$y=\mathbf{F}$	0.2059377286	0.2940622714	0.1029688643	0.3970311357

The corresponding KL distance is 0.0141597409.

C.1.2 Uncorrelated Settings, Original Bell

The optimized, uncorrelated measurement frequencies are

$\Pr(A=a,B=b)=\sigma_{ab}\in\Sigma^{\text{UC}}$	$a=1$	$a=2$	$\Pr(B=b)$
$b=1$	0.2316110419	0.1327830656	0.3643941076
$b=2$	0.4039948505	0.2316110419	0.6356058924
$\Pr(A=a)$	0.6356058924	0.3643941076	

The probabilities of the best classical theory for these frequencies are:

		$a=1$		$a=2$	
$P_{ab}(X=x,Y=x)$		$x=\mathbf{T}$	$x=\mathbf{F}$	$x=\mathbf{T}$	$x=\mathbf{F}$
$b=1$	$y=\mathbf{T}$	0.3901023259	0.1098976741	0.5000000000	0.0000000000
	$y=\mathbf{F}$	0.1098976741	0.3901023259	0.0000000000	0.5000000000
$b=2$	$y=\mathbf{T}$	0.2802046519	0.2197953481	0.3901023259	0.1098976741
	$y=\mathbf{F}$	0.2197953481	0.2802046519	0.1098976741	0.3901023259

As a result, the KL distance for Bell's original proof, with uncorrelated measurement settings is 0.0158003672.

C.1.3 Correlated Settings, Original Bell

The optimized, correlated measurement frequencies are

$\Pr(A=a,B=b)=\sigma_{ab}\in\Sigma$	$a=1$	$a=2$
$b=1$	0.2836084841	0.1020773549
$b=2$	0.3307056768	0.2836084841

The probabilities of the best classical theory for these frequencies are:

		$a = 1$		$a = 2$	
$P_{ab}(X = x, Y = x)$		$x = \mathbf{T}$	$x = \mathbf{F}$	$x = \mathbf{T}$	$x = \mathbf{F}$
$b = 1$	$y = \mathbf{T}$	0.3969913979	0.1030086021	0.4941498806	0.0058501194
	$y = \mathbf{F}$	0.1030086021	0.3969913979	0.0058501194	0.4941498806
$b = 2$	$y = \mathbf{T}$	0.2881326764	0.2118673236	0.3969913979	0.1030086021
	$y = \mathbf{F}$	0.2118673236	0.2881326764	0.1030086021	0.3969913979

The corresponding KL distance is 0.0169800305.

C.2 Optimized Bell

Take the bipartite state $\frac{1}{\sqrt{2}}|0_A 0_B\rangle + \frac{1}{\sqrt{2}}|1_A 1_B\rangle$, and the measurement settings

$$|X = \mathbf{T}\rangle_{a=1} := |R(0)\rangle \quad \text{and} \quad |X = \mathbf{T}\rangle_{a=2} := |R(\tfrac{\pi}{6})\rangle$$
$$|Y = \mathbf{T}\rangle_{b=1} := |R(0)\rangle \quad \text{and} \quad |Y = \mathbf{T}\rangle_{b=2} := |R(\tfrac{\pi}{3})\rangle.$$

With these settings, quantum mechanics predicts the following conditional probabilities.

		$a = 1$		$a = 2$	
$Q_{ab}(X = x, Y = y)$		$x = \mathbf{T}$	$x = \mathbf{F}$	$x = \mathbf{T}$	$x = \mathbf{F}$
$b = 1$	$y = \mathbf{T}$	0.5	0	0.375	0.125
	$y = \mathbf{F}$	0	0.5	0.125	0.375
$b = 2$	$y = \mathbf{T}$	0.125	0.375	0.375	0.125
	$y = \mathbf{F}$	0.375	0.125	0.125	0.375

C.2.1 Uniform Settings, Optimized Bell

For uniform measurement settings frequencies, the best classical approximation is

		$a = 1$		$a = 2$	
$P_{ab}(X = x, Y = x)$		$x = \mathbf{T}$	$x = \mathbf{F}$	$x = \mathbf{T}$	$x = \mathbf{F}$
$b = 1$	$y = \mathbf{T}$	0.5000000000	0.0000000000	0.3333333333	0.1666666667
	$y = \mathbf{F}$	0.0000000000	0.5000000000	0.1666666667	0.3333333333
$b = 2$	$y = \mathbf{T}$	0.1666666667	0.3333333333	0.3333333333	0.1666666667
	$y = \mathbf{F}$	0.3333333333	0.1666666667	0.1666666667	0.3333333333

KL-divergence: 0.0177632822.

C.2.2 Uncorrelated Settings, Optimized Bell

For uncorrelated measurements, the optimal frequencies are

$\Pr(A = a, B = b) = \sigma_{ab} \in \Sigma^{\text{UC}}$	$a = 1$	$a = 2$	$\Pr(B = b)$
$b = 1$	0.1497077788	0.2372131160	0.3869208948
$b = 2$	0.2372131160	0.3758659893	0.6130791052
$\Pr(A = a)$	0.3869208948	0.6130791052	

The probabilities of the best classical theory for these frequencies are:

		$a = 1$		$a = 2$	
$P_{ab}(X = x, Y = x)$		$x = \mathbf{T}$	$x = \mathbf{F}$	$x = \mathbf{T}$	$x = \mathbf{F}$
$b = 1$	$y = \mathbf{T}$	0.5000000000	0.0000000000	0.3267978563	0.1732021436
	$y = \mathbf{F}$	0.0000000000	0.5000000000	0.1732021436	0.3267978563
$b = 2$	$y = \mathbf{T}$	0.1732021436	0.3267978563	0.3464042873	0.1535957127
	$y = \mathbf{F}$	0.3267978563	0.1732021436	0.1535957127	0.3464042873

The corresponding KL distance is 0.0191506613.

C.2.3 Correlated Settings, Optimized Bell

Here are the optimized, correlated measurement frequencies

$\Pr(A=a,B=b)=\sigma_{ab}\in\Sigma$	$a=1$	$a=2$
$b=1$	0.1046493146	0.2984502285
$b=2$	0.2984502285	0.2984502285

The probabilities of the best classical theory for these frequencies are:

		$a=1$		$a=2$	
$P_{ab}(X=x,Y=x)$		$x=\mathbf{T}$	$x=\mathbf{F}$	$x=\mathbf{T}$	$x=\mathbf{F}$
$b=1$	$y=\mathbf{T}$	0.4927305107	0.0072694892	0.3357564964	0.1642435036
	$y=\mathbf{F}$	0.0072694892	0.4927305107	0.1642435036	0.3357564964
$b=2$	$y=\mathbf{T}$	0.1642435036	0.3357564964	0.3357564964	0.1642435036
	$y=\mathbf{F}$	0.3357564964	0.1642435036	0.1642435036	0.3357564964

The corresponding KL distance is 0.0211293952.

C.3 CHSH

The bipartite state $\frac{1}{\sqrt{2}}|0_A0_B\rangle+\frac{1}{\sqrt{2}}|1_A1_B\rangle$. A's and B's measurement settings are:

$$|X=\mathbf{T}\rangle_{a=1}:=|R(0)\rangle \quad\text{and}\quad |X=\mathbf{T}\rangle_{a=2}:=|R(\tfrac{\pi}{4})\rangle, \tag{47}$$

$$|Y=\mathbf{T}\rangle_{b=1}:=|R(\tfrac{\pi}{8})\rangle \quad\text{and}\quad |Y=\mathbf{T}\rangle_{b=2}:=|R(-\tfrac{\pi}{8})\rangle. \tag{48}$$

With these settings, quantum mechanics predicts the following conditional probabilities.

		$a=1$		$a=2$	
$Q_{ab}(X=x,Y=y)$		$x=\mathbf{T}$	$x=\mathbf{F}$	$x=\mathbf{T}$	$x=\mathbf{F}$
$b=1$	$y=\mathbf{T}$	0.4267766953	0.0732233047	0.4267766953	0.0732233047
	$y=\mathbf{F}$	0.0732233047	0.4267766953	0.0732233047	0.4267766953
$b=2$	$y=\mathbf{T}$	0.4267766953	0.0732233047	0.0732233047	0.4267766953
	$y=\mathbf{F}$	0.0732233047	0.4267766953	0.4267766953	0.0732233047

with $\frac{1}{4}+\frac{1}{8}\sqrt{2}\approx 0.4267766953$ and $\frac{1}{4}-\frac{1}{8}\sqrt{2}\approx 0.0732233047$.

C.3.1 Uniform, Uncorrelated and Correlated Settings, CHSH

The optimal measurement settings is the uniform settings, where both A and B perform uses one of the two measurements with probability 0.5 (that is $\sigma_{ab}=0.25$)

The optimal classical theory in this scenario has the following probabilities.

		$a=1$		$a=2$	
$P_{ab}(X=x,Y=x)$		$x=\mathbf{T}$	$x=\mathbf{F}$	$x=\mathbf{T}$	$x=\mathbf{F}$
$b=1$	$y=\mathbf{T}$	0.375	0.125	0.375	0.125
	$y=\mathbf{F}$	0.125	0.375	0.125	0.375
$b=2$	$y=\mathbf{T}$	0.375	0.125	0.125	0.375
	$y=\mathbf{F}$	0.125	0.375	0.375	0.125

KL-divergence: 0.0462738469.

C.4 Hardy

The bipartite state $\alpha|0_A0_B\rangle-\beta|1_A1_B\rangle$, with $\alpha:=\frac{1}{2}\sqrt{2+2\sqrt{-13+6\sqrt{5}}}\approx 0.907$ and $\beta:=\sqrt{1-\alpha^2}\approx 0.421$ (such that indeed $\alpha^2+\beta^2=1$). A's and B's measurement settings are now identical and given by:

$$|X=\mathbf{T}\rangle_{a=1}=|Y=\mathbf{T}\rangle_{b=1} \quad := \quad \sqrt{\frac{\beta}{\alpha+\beta}}|0\rangle+\sqrt{\frac{\alpha}{\alpha+\beta}}|1\rangle, \tag{49}$$

$$|X=\mathbf{T}\rangle_{a=2}=|Y=\mathbf{T}\rangle_{b=2} \quad := \quad -\sqrt{\frac{\beta^3}{\alpha^3+\beta^3}}|0\rangle+\sqrt{\frac{\alpha^3}{\alpha^3+\beta^3}}|1\rangle. \tag{50}$$

With these settings, quantum mechanics predicts the following conditional probabilities.

		$a = 1$		$a = 2$	
$Q_{ab}(X = x, Y = y)$		$x = \mathbf{T}$	$x = \mathbf{F}$	$x = \mathbf{T}$	$x = \mathbf{F}$
$b = 1$	$y = \mathbf{T}$	0	0.38196601125	0.23606797750	0.14589803375
	$y = \mathbf{F}$	0.38196601125	0.23606797750	0	0.61803398875
$b = 2$	$y = \mathbf{T}$	0.23606797750	0	0.09016994375	0.14589803375
	$y = \mathbf{F}$	0.14589803375	0.61803398875	0.14589803375	0.61803398875

C.4.1 Uniform Settings, Hardy

For uniform measurement settings, this is the best classical theory to describe the statistics.

		$a = 1$		$a = 2$	
$P_{ab}(X = x, Y = x)$		$x = \mathbf{T}$	$x = \mathbf{F}$	$x = \mathbf{T}$	$x = \mathbf{F}$
$b = 1$	$y = \mathbf{T}$	0.0338829434	0.3543640363	0.2190090188	0.1692379609
	$y = \mathbf{F}$	0.3543640363	0.2573889840	0.0075052045	0.6042478158
$b = 2$	$y = \mathbf{T}$	0.2190090188	0.0075052045	0.0488933524	0.1776208709
	$y = \mathbf{F}$	0.1692379609	0.6042478158	0.1776208709	0.5958649058

KL-divergence: 0.0278585182.

C.4.2 Uncorrelated Settings, Hardy

Here are the optimized, uncorrelated measurement frequencies

$\Pr(A = a, B = b) = \sigma_{ab} \in \Sigma^{\text{UC}}$	$a = 1$	$a = 2$	$\Pr(B = b)$
$b = 1$	0.2603092699	0.2498958554	0.5102051253
$b = 2$	0.2498958554	0.2398990193	0.4897948747
$\Pr(A = a)$	0.5102051253	0.4897948747	

The probabilities of the best classical theory for these frequencies are:

		$a = 1$		$a = 2$	
$P_{ab}(X = x, Y = x)$		$x = \mathbf{T}$	$x = \mathbf{F}$	$x = \mathbf{T}$	$x = \mathbf{F}$
$b = 1$	$y = \mathbf{T}$	0.0198831449	0.3612213769	0.2143180373	0.1667864844
	$y = \mathbf{F}$	0.3612213769	0.2576741013	0.0141212511	0.6047742271
$b = 2$	$y = \mathbf{T}$	0.2143180373	0.0141212511	0.0481256471	0.1803136414
	$y = \mathbf{F}$	0.1667864844	0.6047742271	0.1803136414	0.5912470702

The corresponding KL distance is 0.0279816333.

C.4.3 Correlated Settings, Hardy

Here are the optimized, correlated measurement frequencies

$\Pr(A = a, B = b) = \sigma_{ab} \in \Sigma$	$a = 1$	$a = 2$
$b = 1$	0.2562288294	0.2431695652
$b = 2$	0.2431695652	0.2574320402

The probabilities of the best classical theory for these frequencies are:

		$a = 1$		$a = 2$	
$P_{ab}(X = x, Y = x)$		$x = \mathbf{T}$	$x = \mathbf{F}$	$x = \mathbf{T}$	$x = \mathbf{F}$
$b = 1$	$y = \mathbf{T}$	0.0173443545	0.3620376608	0.2123471649	0.1670348504
	$y = \mathbf{F}$	0.3620376608	0.2585803238	0.0165954828	0.6040225019
$b = 2$	$y = \mathbf{T}$	0.2123471649	0.0165954828	0.0505353201	0.1784073276
	$y = \mathbf{F}$	0.1670348504	0.6040225019	0.1784073276	0.5926500247

The corresponding KL distance is 0.0280347655.

C.5 Mermin

In [23], we find the following nonlocality proof with two parties, three measurement settings, and two possible outcomes. Let the entangled state be $\frac{1}{\sqrt{2}}(|0_A 0_B\rangle + |1_A 1_B\rangle)$, and the measurement settings:

$$\begin{aligned} |X = \mathbf{T}\rangle_{a=1} = |Y = \mathbf{T}\rangle_{b=1} &:= |0\rangle, \\ |X = \mathbf{T}\rangle_{a=2} = |Y = \mathbf{T}\rangle_{b=2} &:= |R(\tfrac{2}{3}\pi)\rangle, \\ |X = \mathbf{T}\rangle_{a=3} = |Y = \mathbf{T}\rangle_{b=3} &:= |R(\tfrac{4}{3}\pi)\rangle. \end{aligned}$$

With these settings, quantum mechanics predicts the following conditional probabilities.

		$a=1$		$a=2$		$a=3$	
$Q_{ab}(X=x,Y=y)$		$x=\mathbf{T}$	$x=\mathbf{F}$	$x=\mathbf{T}$	$x=\mathbf{F}$	$x=\mathbf{T}$	$x=\mathbf{F}$
$b=1$	$y=\mathbf{T}$	0.5	0	0.125	0.375	0.125	0.375
	$y=\mathbf{F}$	0	0.5	0.375	0.125	0.375	0.125
$b=2$	$y=\mathbf{T}$	0.125	0.375	0.5	0	0.125	0.375
	$y=\mathbf{F}$	0.375	0.125	0	0.5	0.375	0.125
$b=3$	$y=\mathbf{T}$	0.125	0.375	0.125	0.375	0.5	0
	$y=\mathbf{F}$	0.375	0.125	0.375	0.125	0	0.5

C.5.1 Uniform Settings, Mermin

Here are the probabilities of the best classical theory for the uniform measurement settings:

		$a=1$		$a=2$		$a=3$	
$P_{ab}(X=x,Y=y)$		$x=\mathbf{T}$	$x=\mathbf{F}$	$x=\mathbf{T}$	$x=\mathbf{F}$	$x=\mathbf{T}$	$x=\mathbf{F}$
$b=1$	$y=\mathbf{T}$	0.50000	0.00000	0.16667	0.33333	0.16667	0.33333
	$y=\mathbf{F}$	0.00000	0.50000	0.33333	0.16667	0.33333	0.16667
$b=2$	$y=\mathbf{T}$	0.16667	0.33333	0.50000	0.00000	0.16667	0.33333
	$y=\mathbf{F}$	0.33333	0.16667	0.00000	0.50000	0.33333	0.16667
$b=3$	$y=\mathbf{T}$	0.16667	0.33333	0.16667	0.33333	0.50000	0.00000
	$y=\mathbf{F}$	0.33333	0.16667	0.33333	0.16667	0.00000	0.50000

The corresponding KL distance is 0.0157895843.

C.5.2 Uncorrelated Settings, Mermin

Here are the optimized, uncorrelated measurement frequencies $\Pr(A = a, B = b)$:

$\sigma_{ab} \in \Sigma^{\text{UC}}$	$a=1$	$a=2$	$a=3$	$\Pr(B=b)$
$b=1$	0.1497077711	0	0.2372131137	0.3869208848
$b=2$	0.2372131137	0	0.3758660015	0.6130791152
$b=3$	0	0	0	0
$\Pr(A=a)$	0.3869208848	0	0.6130791152	

The probabilities of the best classical theory for these frequencies are:

		$a=1$		$a=2$		$a=3$	
$P_{ab}(X=x,Y=y)$		$x=\mathbf{T}$	$x=\mathbf{F}$	$x=\mathbf{T}$	$x=\mathbf{F}$	$x=\mathbf{T}$	$x=\mathbf{F}$
$b=1$	$y=\mathbf{T}$	0.50000	0.00000	0.50000	0.00000	0.17320	0.32680
	$y=\mathbf{F}$	0.00000	0.50000	0.50000	0.00000	0.32680	0.17320
$b=2$	$y=\mathbf{T}$	0.17320	0.32680	0.50000	0.00000	0.15360	0.34640
	$y=\mathbf{F}$	0.32680	0.17320	0.50000	0.00000	0.34640	0.15360
$b=3$	$y=\mathbf{T}$	0.50000	0.50000	1.00000	0.00000	0.50000	0.50000
	$y=\mathbf{F}$	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

The corresponding KL distance is 0.0191506613.

C.5.3 Correlated Settings, Mermin

Here are the optimized, correlated measurement frequencies (note that there are also other optimal frequencies):

$\Pr(A=a,B=b)=\sigma_{ab}\in\Sigma$	$a=1$	$a=2$	$a=3$
$b=1$	0.1046493071	0	0.2984502310
$b=2$	0.2984502310	0	0.2984502310
$b=3$	0	0	0

The probabilities of the best classical theory for these frequencies are:

$P_{ab}(X=x,Y=y)$		$a=1$		$a=2$		$a=3$	
		$x=\mathbf{T}$	$x=\mathbf{F}$	$x=\mathbf{T}$	$x=\mathbf{F}$	$x=\mathbf{T}$	$x=\mathbf{F}$
$b=1$	$y=\mathbf{T}$	0.49273	0.00727	0.50000	0.00000	0.16424	0.33576
	$y=\mathbf{F}$	0.00727	0.49273	0.50000	0.00000	0.33576	0.16424
$b=2$	$y=\mathbf{T}$	0.16424	0.33576	0.50000	0.00000	0.16424	0.33576
	$y=\mathbf{F}$	0.33576	0.16424	0.50000	0.00000	0.33576	0.16424
$b=3$	$y=\mathbf{T}$	0.50000	0.50000	1.00000	0.00000	0.50000	0.50000
	$y=\mathbf{F}$	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

The corresponding KL distance is 0.0211293952.

C.6 GHZ

The tripartite state $\frac{1}{\sqrt{2}}|0_A0_B0_C\rangle+\frac{1}{\sqrt{2}}|1_A1_B1_C\rangle$. The settings for all three parties are identical:

$$|X=\mathbf{T}\rangle_{a=1}=|Y=\mathbf{T}\rangle_{b=1}=|Z=\mathbf{T}\rangle_{c=1} \quad := \quad \tfrac{1}{\sqrt{2}}|0\rangle+\tfrac{1}{\sqrt{2}}|1\rangle, \tag{51}$$

$$|X=\mathbf{T}\rangle_{a=2}=|Y=\mathbf{T}\rangle_{b=2}=|Z=\mathbf{T}\rangle_{c=2} \quad := \quad \tfrac{1}{\sqrt{2}}|0\rangle+\tfrac{\mathrm{i}}{\sqrt{2}}|1\rangle. \tag{52}$$

With these settings, quantum mechanics predicts the following conditional probabilities.

$Q_{abc}(X=x,Y=y,Z=z)$				$a=1$		$a=2$	
				$x=\mathbf{T}$	$x=\mathbf{F}$	$x=\mathbf{T}$	$x=\mathbf{F}$
$c=1$	$b=1$	$z=\mathbf{T}$	$y=\mathbf{T}$	0.25	0	0.125	0.125
			$y=\mathbf{F}$	0	0.25	0.125	0.125
		$z=\mathbf{F}$	$y=\mathbf{T}$	0	0.25	0.125	0.125
			$y=\mathbf{F}$	0.25	0	0.125	0.125
	$b=2$	$z=\mathbf{T}$	$y=\mathbf{T}$	0.125	0.125	0	0.25
			$y=\mathbf{F}$	0.125	0.125	0.25	0
		$z=\mathbf{F}$	$y=\mathbf{T}$	0.125	0.125	0.25	0
			$y=\mathbf{F}$	0.125	0.125	0	0.25
$c=2$	$b=1$	$z=\mathbf{T}$	$y=\mathbf{T}$	0.125	0.125	0	0.25
			$y=\mathbf{F}$	0.125	0.125	0.25	0
		$z=\mathbf{F}$	$y=\mathbf{T}$	0.125	0.125	0.25	0
			$y=\mathbf{F}$	0.125	0.125	0	0.25
	$b=2$	$z=\mathbf{T}$	$y=\mathbf{T}$	0	0.25	0.125	0.125
			$y=\mathbf{F}$	0.25	0	0.125	0.125
		$z=\mathbf{F}$	$y=\mathbf{T}$	0.25	0	0.125	0.125
			$y=\mathbf{F}$	0	0.25	0.125	0.125

C.6.1 Uniform and Uncorrelated Settings, GHZ

For all three settings, the best possible classical statistics that approximate the GHZ experiment is:

				$a = 1$		$a = 2$	
$P_{abc}(X = x, Y = y, Z = z)$				$x = \mathbf{T}$	$x = \mathbf{F}$	$x = \mathbf{T}$	$x = \mathbf{F}$
$c = 1$	$b = 1$	$z = \mathbf{T}$	$y = \mathbf{T}$	0.1875	0.0625	0.125	0.125
			$y = \mathbf{F}$	0.0625	0.1875	0.125	0.125
		$z = \mathbf{F}$	$y = \mathbf{T}$	0.0625	0.1875	0.125	0.125
			$y = \mathbf{F}$	0.1875	0.0625	0.125	0.125
	$b = 2$	$z = \mathbf{T}$	$y = \mathbf{T}$	0.125	0.125	0.0625	0.1875
			$y = \mathbf{F}$	0.125	0.125	0.1875	0.0625
		$z = \mathbf{F}$	$y = \mathbf{T}$	0.125	0.125	0.1875	0.0625
			$y = \mathbf{F}$	0.125	0.125	0.0625	0.1875
$c = 2$	$b = 1$	$z = \mathbf{T}$	$y = \mathbf{T}$	0.125	0.125	0.0625	0.1875
			$y = \mathbf{F}$	0.125	0.125	0.1875	0.0625
		$z = \mathbf{F}$	$y = \mathbf{T}$	0.125	0.125	0.1875	0.0625
			$y = \mathbf{F}$	0.125	0.125	0.0625	0.1875
	$b = 2$	$z = \mathbf{T}$	$y = \mathbf{T}$	0.0625	0.1875	0.125	0.125
			$y = \mathbf{F}$	0.1875	0.0625	0.125	0.125
		$z = \mathbf{F}$	$y = \mathbf{T}$	0.1875	0.0625	0.125	0.125
			$y = \mathbf{F}$	0.0625	0.1875	0.125	0.125

The optimal uncorrelated setting is the uniform settings that samples all eight measurement settings with equal probability. The corresponding KL divergence is: 0.2075187496.

C.6.2 Correlated Settings, GHZ

The optimal correlated setting samples with equal probability those four settings that yield the $(0.125, 0)$ outcome probabilities (those are the settings were an even number of the measurements are measuring along the m_1 axis). The KL divergence in this setting is twice that of the previous uniform setting: 0.4150374993.

D The Kullback-Leibler Divergence

This appendix provides in-depth information about the Kullback-Leiber divergence and its relation to statistical strength. Appendix D.1 discusses some general properties of KL divergence. Appendix D.2 compares it to variation distance. Appendix D.3 informally explains why KL divergence is related to statistical strength.

D.1 Properties of KL Divergence

General Properties Let $\mathcal{P}$ be the set of distributions on $\mathcal{Z}$. We equip $\mathcal{P}$ with the Euclidean topology by identifying each $P \in \mathcal{P}$ with its probability vector. Then $D(P\|Q)$ is jointly continuous in P and Q on the interior of $\mathcal{P}$. It is jointly *lower semicontinuous* (for a definition see, e.g., [30]), but not continuous, on $\mathcal{P}$. It is also jointly strictly convex on $\mathcal{P}$.

Because $Q(z) \log Q(z) = 0$ as $z \downarrow 0$ we can ignore the $Q(z) = 0$ parts in the summation, and hence

$$D(Q\|P) = \sum_{\substack{z \in \mathcal{Z} \\ Q(z) > 0}} Q(z)[-\log P(z) + \log Q(z)]. \tag{53}$$

The Additivity Property The KL divergence has the following additivity property. Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sample spaces, and let P and Q be distributions over the product space $\mathcal{X} \times \mathcal{Y}$. Let P_X (Q_X) denote the marginal distribution of P (Q) over $\mathcal{X}$, and for each $x \in \mathcal{X}$, let $P_{Y|x}$ ($Q_{Y|x}$) denote the conditional distribution over $\mathcal{Y}$ conditioned on $X = x$, i.e. for all $y \in \mathcal{Y}$, $P_{Y|x}(y) := P(y|x)$. Then

$$\begin{aligned} D(Q\|P) &= \sum_{x \in \mathcal{X}} Q(x) D(Q_{Y|x}\|P_{Y|x}) + D(Q_X\|P_X) & (54) \\ &= \mathbf{E}_{Q_X}[D(Q_{Y|X}\|P_{Y|X})] + D(Q_X\|P_X). & (55) \end{aligned}$$

An important consequence of this property is that the divergence between the joint distribution of n independent drawings from Q to that of n independent drawings from P, is n times the divergence for one drawing. It also implies Equation 26 in Section 4.

D.2 Kullback-Leibler vs. Total Variation Distance

In discussions about the strengths of nonlocality proofs, it has sometimes been claimed that QM should use the filter settings that give the largest deviation in the Bell inequality. This would mean that QM should try to set up the experiment such that the distribution of outcomes Q is as distant as possible to LR's distribution over outcomes P where distance is measured by the so-called *total variation distance*, [10] between Q and P, defined as $\sum_{z \in \mathcal{Z}} |P(z) - Q(z)|$. While it is true that this defines a distance between probability distributions, it is only one of large number of possible distances or divergences that can be defined for probability distributions. But if one is interested in measuring 'statistical distance', total variation is *not* the appropriate distance measure to use. Instead, one should use the KL divergence. To get some feel for how different KL and total variation can be, let $\mathcal{Z} = \{1, 2\}$ and consider the following possibilities for P and Q:

1. $P(1) = 0.99$ and $Q(1) = 1$. Then the absolute difference in probabilities between P and Q is very small (0.02); however, if data are sampled from P, then, with high probability, after a few hundred trials we will have observed at least one 0. From that point on, we are *100% certain* that P, and not Q, has generated the data. This is reflected by the fact that $D(P\|Q) = \infty$.
2. Let P and Q be as above but consider $D(Q\|P)$. We have $D(Q\|P) = -1 \cdot \log 0.99 = 0.015$. This illustrates that, if Q rather than P generates the data, we typically need an enormous amount of data before we can be reasonably sure that Q indeed generated the data.
3. $P(1) = 0.49, Q(1) = 0.5$. In this case, $D(P\|Q) = 0.49 \log 0.98 + 0.51 \log 1.02 \approx 0.000289$ and $D(Q\|P) = 0.5(-\log 0.98 - \log 1.02) \approx 0.000289$. Now the average support per trial in favor of Q under distribution Q is about equal to the average support per trial in favor of P under P.
4. Note that the KL divergences for the 'near uniform' distributions with $P(1), Q(1) \approx 0.5$ is much smaller than the divergences for the skewed distributions with $P(1), Q(1) \approx 1$, while the total variation distance is the same for all these distributions.

The example stresses the asymmetry of KL divergence as well as its difference from the absolute deviations between probabilities.

D.3 Intuition Behind It All

Here we give some intuition on the relation between KL divergence and statistical strength. It can be read without any statistical background. Let $Z_1, Z_2, \ldots$ be a sequence of random variables independently generated either by some distribution P or by some distribution Q with $Q \neq P$. Suppose we are given a sample (sequence of outcomes) $z_1, \ldots, z_n$. Perhaps simplest (though by no means only) way of finding out whether Q or P generated this data is to compare the *likelihood* (in our case, 'likelihood' = 'probability') of the data $z_1, \ldots, z_n$ according to the two distributions. That is, we look at the ratio

$$\frac{Q(z_1, \ldots, z_n)}{P(z_1, \ldots, z_n)} \quad = \quad \frac{\prod_{i=1}^n Q(z_i)}{\prod_{i=1}^n P(z_i)}. \tag{56}$$

Intuitively, if this ratio is larger than 1, the data is more typical for Q than for P, and we might decide that Q rather than P generated the data. Again intuitively, the magnitude of the ratio in Equation 56 might give us an idea of the confidence we should have in this decision.

Now assume that the data are actually generated according to Q, i.e. 'Q is true'. We will study the behavior of the logarithm of the likelihood ratio in Equation 56 under this assumption (the use of the logarithm is only to simplify the analysis; using Equation 56 directly would have led to the same conclusions). The *Law of Large Numbers* [9] tells us that, with Q-probability 1, averages of bounded random variables will converge to their Q-expectations. In particular, if the Z_i take values in a finite set $\mathcal{Z}$, and P and Q are such that for all $z \in \mathcal{Z}, P(z), Q(z) > 0$, then with Q-probability 1,

$$\frac{1}{n} \sum_{i=1}^{n} L_i \quad \to \quad \mathbf{E}_Q[L] \tag{57}$$

where $L_i := \log(Q(Z_i)/P(Z_i))$, and $\mathbf{E}_Q[L] = \mathbf{E}_Q[L_1] = \cdots = \mathbf{E}_Q[L_n]$ is given by $\mathbf{E}_Q[L] = \mathbf{E}_Q \log(Q/P) = \sum_z Q(z) \log(Q(z)/P(z)) = D(Q\|P)$. Therefore, with Q-probability 1,

$$\frac{1}{n} \log \frac{Q(Z_1, \ldots, Z_n)}{P(Z_1, \ldots, Z_n)} \quad \to \quad D(Q\|P). \tag{58}$$

Thus, with Q-probability 1, the *average log-likelihood ratio between P and Q will converge to the KL divergence between P and Q.* This means that the likelihood ratio, which may be viewed as the amount of evidence for Q vs. P, is asymptotically determined by the KL divergence, to first order in the exponent. For example, let us test Q first against P_1 with $D(Q\|P_1) = \epsilon_1$, and then against P_2 with $D(Q\|P_2) = \epsilon_2 > \epsilon_1$, then, with Q-probability 1,

$$\frac{1}{n}\log\frac{Q(Z_1,\ldots,Z_n)}{P_1(Z_1,\ldots,Z_n)} \rightarrow \epsilon_1 \quad \text{and} \quad \frac{1}{n}\log\frac{Q(Z_1,\ldots,Z_n)}{P_2(Z_1,\ldots,Z_n)} \rightarrow \epsilon_2. \tag{59}$$

This implies that with increasing n, the likelihood ratio Q/P_1 becomes exponentially smaller than the likelihood ratio Q/P_2:

$$\frac{Q(Z_1,\ldots,Z_n)}{P_1(Z_1,\ldots,Z_n)} \quad \leq \quad \frac{Q(Z_1,\ldots,Z_n)}{P_1(Z_2,\ldots,Z_n)}\cdot \mathrm{e}^{-n(\epsilon_2-\epsilon_1)+o(n)}. \tag{60}$$

Returning to the setting discussed in this paper, this preliminary analysis suggests that from QM's point of view (who knows that Q is true), the most convincing experimental results (highest likelihood ratio of Q_σ vs. $P_{\sigma;\pi}$) are obtained if the KL divergence between Q_σ and $P_{\sigma;\pi}$ is as large as possible. If Q_σ is compared against a set $\mathcal{P}_\sigma$, then the analysis suggests that the most convincing experimental results are obtained if the KL divergence between Q_σ and $\mathcal{P}_\sigma$ is as large as possible, that is, if $\inf_{P\in\mathcal{P}_\sigma} D(Q_\sigma\|P_\sigma)$ is as large as possible.

D.4 Bayesian Analysis

In this appendix we assume some basic knowledge of Bayesian statistics. We only give the derivation for the $2\times2\times2$ nonlocality proofs. Extension to generalized nonlocality proofs is straightforward.

Let us identify $H_1 := Q_\sigma$ and $H_0 := \mathcal{P}_\sigma$, where Q_σ and $\mathcal{P}_\sigma$ are defined as quantum and local realist theories respectively, as in Section 2. We start with a prior Pr on H_1 and H_0, and we assume $0 < \Pr(H_1) < 1$.

Now, *conditioned* on H_0 being the case, the actual distribution generating the data may still be any $P_{\sigma;\pi}\in H_0$. To indicate the prior degree of belief in these, we further need a *conditional prior distribution* $\Pr(\cdot|H_0)$ over all the distributions in H_0. Since σ is fixed, $H_0 = \mathcal{P}_\sigma$ is parameterized by the set π. We can therefore define the prior $\Pr(\cdot|H_0)$ in terms of a probability density function w over π, where we define for each (measurable) $A\subset\Pi$,

$$\Pr(\{P_{\sigma;\pi} \;:\; \pi\in A\}\mid H_0) \quad := \quad \int_{\pi\in A} w(\pi)\mathrm{d}\pi. \tag{61}$$

We restrict attention to prior densities $w(\cdot)$ that are continuous and uniformly bounded away from 0. By the latter we mean that there exists $w_{\min} > 0$ such that for all $\pi\in\Pi$, $w(\pi) > w_{\min}$. For concreteness one may take w to be uniform (constant over π), although this will not affect the analysis.

In order to apply Bayesian inference, we further have to define $\Pr(z_1,\ldots,z_n|H_i)$, 'the probability of the data given that H_i is true'. We do this in the standard Bayesian manner:

$$\begin{aligned}\Pr(z_1,\ldots,z_n|H_1) \quad &:= \quad Q_\sigma(z_1,\ldots,z_n),\\ \Pr(z_1,\ldots,z_n|H_0) \quad &:= \quad \int_{\pi\in\Pi} P_{\sigma;\pi}(z_1,\ldots,z_n)w(\pi)\mathrm{d}\pi.\end{aligned} \tag{62}$$

Here each outcome z_i consists of a realized measurement setting and an experimental outcome in that setting; hence we can write $z_i = (a_i,b_i,x_i,y_i)$ for $a_i,b_i\in\{1,2\}$ and $x_i,y_i\in\{\mathbf{F},\mathbf{T}\}$.

Together with the prior over $\{H_1,H_0\}$, Equation 62 defines a probability distribution over the product space $\{H_1,H_0\}\times\mathcal{Z}^n$ where $\mathcal{Z} := \{1,2\}\times\{1,2\}\times\{\mathbf{F},\mathbf{T}\}\times\{\mathbf{F},\mathbf{T}\}$. Given experimental data $z_1,\ldots,z_n$ and prior distribution Pr, we can now use Bayes' rule [9, 21] to compute the *posterior distribution* of H_i:

$$\Pr(H_i|z_1,\ldots,z_n) \quad = \quad \frac{\Pr(z_1,\ldots,z_n|H_i)\Pr(H_i)}{\sum_i \Pr(z_1,\ldots,z_n|H_i)\Pr(H_i)} \tag{63}$$

According to Bayesian hypothesis testing, we should select the H_i maximizing the posterior probability of Equation 63. The *confidence* in the decision, which we denote by `post-odds`, is given by the posterior odds against H_0:

$$\begin{aligned} \texttt{post-odds} &= \frac{\Pr(H_1|z_1,\ldots,z_n)}{\Pr(H_0|z_1,\ldots,z_n)} && (64) \\ &= \frac{\Pr(z_1,\ldots,z_n|H_1)\Pr(H_1)}{\Pr(z_1,\ldots,z_n|H_0)\Pr(H_0)} && (65) \\ &= \frac{Q_\sigma(z_1,\ldots,z_n)}{\int_{\pi\in\Pi} P_{\sigma;\pi}(z_1,\ldots,z_n)w(\pi)\mathrm{d}\pi} \cdot \frac{\Pr(H_1)}{\Pr(H_0)} && (66) \end{aligned}$$

Note that `post-odds` depends on H_0, H_1 and the data $z_1,\ldots,z_n$. The factor on the left of Equation 64 is called the *Bayes factor*, and the factor on the right is called the *prior odds*. Since the Bayes factor typically increases exponentially with n, the influence of the prior odds on the posterior odds is negligible for all but the smallest n. Below we show that, if H_1 is true ('QM is right'), then with probability 1,

$$\frac{1}{n}\log \texttt{post-odds} \quad\to\quad \inf_{\pi\in\Pi} D(Q_\sigma\|P_{\sigma;\pi}). \tag{67}$$

Therefore Equation 16 holds: the confidence `post-odds` will be determined, to first order in the exponent, by $\inf_{\pi\in\Pi} D(Q_\sigma\|P_{\sigma;\pi})$. This gives a Bayesian justification of adopting $D(Q_\sigma\|\mathcal{P}_\sigma)$ as an indicator of statistical strength – provided that we can show that Equation 67 holds. We proceed to show this.

Proof of Equation 67 We first note that

$$\begin{aligned} \log P_{\sigma;\pi}(z_1,\ldots,z_n) &= \log P_{\sigma;\pi}((a_1,b_1,x_1,y_1),\ldots,(a_n,b_n,x_n,y_n)) && (68) \\ &= n\cdot \sum_{a,b\in\{1,2\}} \mathbb{P}(a,b)\sigma_{ab}\log\mathbb{P}(a,b) \;+ && (69) \\ & \quad n\cdot \sum_{\substack{a,b\in\{1,2\}\\ x,y\in\{\mathbf{F},\mathbf{T}\}}} \mathbb{P}(a,b,x,y)\log\left(\sum_{\substack{x_1,x_2,y_1,y_2\\ x_a=x,y_b=y}} \pi_{x_1x_2y_1y_2}\right). && (70) \end{aligned}$$

Here $\mathbb{P}(a,b)$ is the frequency (number of occurrences in the sample divided by n) of experimental outcomes with measurement setting (a,b) and $\mathbb{P}(a,b,x,y)$ is the frequency of experimental outcomes with measurement setting (a,b) and outcome $X=x, Y=y$.

Let $\tilde{\pi}$ be any π achieving $\inf_{\pi\in\pi} D(Q_\sigma\|P_{\sigma;\pi})$. By Theorem 1, such a $\tilde{\pi}$ must exist, and Q_σ must be absolutely continuous with respect to $P_{\sigma;\tilde{\pi}}(a,b,x,y)$. It follows that

$$P_{\sigma;\tilde{\pi}}(a,b,x,y) \quad=\quad \sigma_{ab}\sum_{\substack{x_1,x_2,y_1,y_2\\ x_a=x,y_b=y}} \pi_{x_1x_2y_1y_2} \tag{71}$$

may be equal to 0 *only* if $\sigma_{ab}Q_\sigma(x,y,a,b)=0$. From this it follows (with some calculus) that there must be a constant c and an open (L_1-distance) ball $B_\epsilon := \{\pi\in\Pi \mid \;\; |\pi-\tilde{\pi}|<\epsilon\}$ around $\tilde{\pi}$ such that for all $\pi\in B_\epsilon$, for all n, all sequences $z_1,\ldots,z_n$. with $Q_\sigma(z_1,\ldots,z_n)>0$,

$$\frac{|\frac{1}{n}\log P_{\sigma;\pi}(z_1,\ldots,z_n) - \frac{1}{n}\log P_{\sigma;\tilde{\pi}}(z_1,\ldots,z_n)|}{|\pi-\tilde{\pi}|} \quad\leq\quad c \tag{72}$$

whence $|\log P_{\sigma;\pi}(z_1,\ldots,z_n) - \log P_{\sigma;\tilde{\pi}}(z_1,\ldots,z_n)| \leq nc|\pi-\tilde{\pi}| \leq nc\epsilon$. By choosing $\epsilon = n^{-2}$ we find that for sufficiently large n,

$$\begin{aligned} \Pr(z_1,\ldots,z_n|H_0) &= \int_{\pi\in\Pi} P_{\sigma;\pi}(z_1,\ldots,z_n)w(\pi)\mathrm{d}\pi && (73) \\ &\geq \int_{\pi\in B_\epsilon} P_{\sigma;\pi}(z_1,\ldots,z_n)w(\pi)\mathrm{d}\pi && (74) \\ &\geq w_{\min}c(\frac{1}{n^2})^k \mathrm{e}^{-c/n} P_{\sigma;\tilde{\pi}}(z_1,\ldots,z_n). && (75) \end{aligned}$$

Hence, $-\log\Pr(z_1,\ldots,z_n|H_0) \leq \log P_{\sigma;\tilde{\pi}}(z_1,\ldots,z_n) + -O(\log n)$. By applying the strong law of large numbers to $n^{-1}\sum\log P_{\sigma;\tilde{\pi}}(Z_i)$, we find that, with Q_σ-probability 1,

$$\begin{aligned} -\frac{1}{n}\log\Pr(Z_1,\ldots,Z_n|H_0) &\leq \mathbf{E}_{Q_\sigma}[-\log P_{\sigma;\tilde{\pi}}(Z)] + O(\frac{\log n}{n}) && (76) \\ &= \inf_{\pi\in\Pi}\mathbf{E}_{Q_\sigma}[-\log P_{\pi,\sigma}(Z)] + O(\frac{\log n}{n}). && (77) \end{aligned}$$

This bounds $-\frac{1}{n}\log \Pr(Z^n|H_0)$ from above. We proceed to bound it from below. Note that for all n, $z_1, \ldots, z_n$,

$$-\frac{1}{n}\log \int_{\pi\in\Pi} P_{\sigma;\pi}(z_1,\ldots,z_n)w(\pi)\mathrm{d}\pi \quad \geq \quad \inf_{\pi\in\Pi} -\frac{1}{n}\log P_{\sigma;\pi}(z_1,\ldots,z_n). \tag{78}$$

To complete the proof, we need to relate

$$\inf_{\pi\in\Pi} -\frac{1}{n}\log P_{\sigma;\pi}(z_1,\ldots,z_n) \quad = \quad \inf_{\pi\in\Pi} -\frac{1}{n}\sum_{i=1}^{n}\log P_{\sigma;\pi}(z_i) \tag{79}$$

(which depends on the data) to its 'expectation version' $\inf_{\pi\in\Pi}\mathbf{E}_{Q_\sigma}[-\log P_{\sigma;\pi}(Z)]$. This can be done using a version of the *uniform law of large numbers* [32]. Based on such a uniform law of large numbers, (for example, [16, Chapter 5, Lemma 5.14]) one can show that for all distributions Q over $\mathcal{Z}$, with Q-probability 1, as $n\rightarrow\infty$,

$$\inf_{\pi\in\Pi} -\frac{1}{n}\log P_{\sigma;\pi}(Z_1,\ldots,Z_n) \quad \rightarrow \quad \inf_{\pi\in\Pi}\mathbf{E}_{Q}[-\log P_{\sigma;\pi}(Z)]. \tag{80}$$

Together, Equations 76, 78 and 80 show that, with Q_σ-probability 1, as $n\rightarrow\infty$,

$$-\frac{1}{n}\log \Pr(Z_1,\ldots,Z_n|H_0) \quad \rightarrow \quad \inf_{\pi\in\Pi}\mathbf{E}_{Q_\sigma}[-\log P_{\sigma;\pi}(Z)] \tag{81}$$

Together with the law of large numbers applied to $n^{-1}\sum\log\Pr(Z_i|H_1)$, we find that

$$\frac{1}{n}\log\frac{\Pr(Z_1,\ldots,Z_n|H_1)}{\Pr(Z_1,\ldots,Z_n|H_0)} \quad \rightarrow \quad \inf_{\pi\in\pi}\mathbf{E}_{Q_\sigma}[\log\frac{Q_\sigma(Z)}{P_{\sigma;\pi}(Z)}]. \tag{82}$$

Noting that the right hand side is equal to $\inf_{\pi\in\Pi}\ D(Q_\sigma\|P_{\sigma;\pi})$ and plugging this in into Equation 64, we see that with H_1-probability 1, Equation 67 holds. This is what we had to prove.

D.5 Information Theoretic Analysis

In this appendix we assume that the reader is familiar with the basics of information theory.

The code with lengths $L_{\mathcal{P}_\sigma}$ is simply the Shannon-Fano code for the Bayesian marginal likelihood $\Pr(z_1,\ldots,z_n \mid \mathcal{P}_\sigma) = \Pr(z_1,\ldots,z_n \mid H_0)$ as defined in Equation 62. For each n, each $z_1,\ldots,z_n \in \mathcal{Z}^n$, this code achieves lengths (up to 1 bit) $-\log\Pr(z_1,\ldots,z_n \mid \mathcal{P}_\sigma)$. The code corrsponding to Q_σ achieves lengths $-\log Q(z_1,\ldots,z_n)$. We have already shown in Appendix D.4, Equation 82, that, with Q-probability 1, as $n\rightarrow\infty$,

$$\frac{1}{n}\left[-\log\Pr(z_1,\ldots,z_n \mid \mathcal{P}_\sigma) - [-\log Q(z_1,\ldots,z_n)]\right] \quad \rightarrow \quad D(Q_\sigma\|\mathcal{P}_\sigma). \tag{83}$$

Noting that the left hand side is equal to n^{-1}`bit-diff`, Equation 23 follows.

E Proofs

E.1 Preparation

The proof of Theorem 1 uses the following lemma, which is of some independent interest.

Lemma 1 *Let* (Σ,Π,U) *be the game corresponding to an arbitrary* 2 *party,* 2 *measurement settings per party non-locality proof. For any* $a_0, b_0 \in \{1,2\}$*, there exists a* $\pi\in\Pi$ *such that for all* $a,b\in\{1,2\}, (a,b)\neq(a_0,b_0)$ *we have* $Q_{ab} = P_{ab;\pi}$*. Thus, for any three of the four measurement settings, the probability distribution on outcomes can be perfectly explained by a local realist theory.*

Proof We will give a detailed proof for the case that the measurement outcomes are two values $\{R,G\}$; the general case can be proved in a similar way.

Without loss of generality let $(a_0,b_0)=(2,2)$. Now we must prove that the equation $Q_{ab}=P_{ab;\pi}$ holds for the three settings $(a,b)\in\{(1,1),(1,2),(2,1)\}$. Every triple of distributions $P_{ab;\pi}$ for these three settings may be represented by a table of the form

Pr	$X_1=\mathbf{F}$	$X_1=\mathbf{T}$	$X_2=\mathbf{F}$	$X_2=\mathbf{T}$
$Y_1=\mathbf{F}$	p_1	p_2	p_3	p_4
$Y_1=\mathbf{T}$	p_5	p_6	p_7	p_8
$Y_2=\mathbf{F}$	p_9	p_{10}		
$Y_2=\mathbf{T}$	p_{11}	p_{12}		

(84)

with $p_1, p_2, \ldots, p_{12} \geq 0$. Given any table of this form, we say that the LR distribution P_π *corresponds* to the table if $P_{00;\pi}(\mathbf{F}, \mathbf{F}) = p_0/(p_0 + p_1 + p_5 + p_6), P_{10;\pi}(\mathbf{F}, \mathbf{F}) = p_3/(p_3 + p_4 + p_7 + p_8)$ etc. Note that by this implicit normalization, we do not require the four entries within each sub-table to sum to one; in this way, each $p_1, p_2, \ldots, p_{12} \geq 0$ with at least one non-zero element per sub-table corresponds to a set of three conditional distributions.

Our experimental setup implies that the realized measurement setting on A's side should not influence the probability that B observes $Y_1 = y_1$. This is expressed by the equality $p_1 + p_2 = p_3 + p_4$. In total there are four of such no-signaling restrictions:

$$\begin{cases} p_1 + p_2 & = & p_3 + p_4 \\ p_5 + p_6 & = & p_7 + p_8 \\ p_1 + p_5 & = & p_9 + p_{11} \\ p_2 + p_6 & = & p_{10} + p_{12}. \end{cases} \tag{85}$$

We call a table with $p_1, \ldots, p_{12} \geq 0$ and satisfying Equation 85 and with at least one element per sub-table strictly > 0, a Γ-*table.* We already showed that each triple of conditional LR distributions may be represented as a Γ-table. In exactly the same way one shows that each triple of conditional quantum experimentalist distributions $Q_{ab}, (a, b) \neq (2, 2)$ can be represented as a Γ-table. It therefore suffices if we can show that *every* Γ-table corresponds to some LR theory P_π. We show this by considering the 16 possible deterministic theories $T_{x_1x_2y_1y_2}$. Here $T_{x_1x_2y_1y_2}$ is defined as the theory with $P_\pi(X_1 = x_1, X_2 = x_2, Y_1 = y_1, Y_2 = y_2) = \pi_{x_1x_2y_1y_2} = 1$. Each deterministic theory $\pi_{x_1x_2y_1y_2}$ corresponds to a specific Γ-table denoted by $\Gamma_{x_1x_2y_1y_2}$. For example, the theory $T_{\mathbf{FFTF}}$ gives the $\Gamma_{\mathbf{FFTF}}$:

Pr	$X_1 = \mathbf{F}$	$X_1 = \mathbf{T}$	$X_2 = \mathbf{F}$	$X_2 = \mathbf{T}$
$Y_1 = \mathbf{F}$	0	0	0	0
$Y_1 = \mathbf{T}$	1	0	1	0
$Y_2 = \mathbf{F}$	1	0		
$Y_2 = \mathbf{T}$	0	0		

. (86)

We will prove that the set of Γ-tables is in fact the affine hull of the 16 tables $\Gamma_{x_1x_2y_1y_2}$ corresponding to deterministic theories. This shows that any Γ-table can be reproduced by a mixture of deterministic theories. Since every LR theory $\pi \in \Pi$ can be written as such a mixture, this proves the lemma.

Given a Γ-table, we focus on its smallest nonzero entry $\Gamma_{ab} = \varepsilon > 0$. By the restrictions of Equation 85 there exists a deterministic theory T_k such that $\Gamma - \varepsilon\Gamma_k$ has no negative entries. For example, suppose that the smallest element in Γ corresponds to $P_\pi(X_1 = \mathbf{F}, Y_1 = \mathbf{T})$ (denoted as p_5 in the first table above). By the restrictions of Equation 85, either the table $\Gamma - p_5\Gamma_{\mathbf{FFTF}}$ (where $\Gamma_{\mathbf{FFTF}}$ is shown above) or one of the three tables $\Gamma - p_5\Gamma_{\mathbf{FFTT}}, \Gamma - p_5\Gamma_{\mathbf{FTTF}}, \Gamma - p_5\Gamma_{\mathbf{FTTT}}$ only has nonnegative entries.

Let $\Gamma' := \Gamma - \varepsilon\Gamma_k$ where k is chosen such that Γ' has no negative entries. Clearly, either Γ' only has 0-entries or Γ' is a Γ-table with number of nonzero entries one less than that of Γ. Hence by applying the above procedure at most 16 times, we obtain a decomposition $\Gamma = \varepsilon_1\Gamma_{k_1} + \cdots + \varepsilon_{16}\Gamma_{k_{16}}$, which shows that Γ lies in the affine hull of the Γ-tables corresponding to deterministic theories. □

E.2 Proof of Theorem 1

Theorem 1: *Let Q be a given (not necessarily $2 \times 2 \times 2$) nonlocality proof and Π the corresponding set of local realist theories.*

1. *Let $U(\sigma, \pi) := D(Q_\sigma \| P_{\sigma;\pi})$, then:*
   (a) *For a $2 \times 2 \times 2$ proof, we have that*

$$U(\sigma, \pi) \quad = \quad \sum_{a,b \in \{1,2\}} \sigma_{ab} D(Q_{ab}(\cdot) \| P_{ab;\pi}(\cdot)) \tag{87}$$

   Hence, the KL divergence $D(Q_\sigma \| P_{\sigma;\pi})$ may alternatively be viewed as the average KL divergence between the distributions of (X, Y), where the average is over the settings (A, B). For a generalized nonlocality proof, the analogous generalization of Equation 87 holds.

   (b) *For fixed σ, $U(\sigma, \pi)$ is convex and lower semicontinuous on Π, and continuous and differentiable on the interior of Π.*

   (c) *If Q is absolutely continuous with respect to some fixed π, then $U(\sigma, \pi)$ is linear in σ.*

2. Let

$$U(\sigma) \quad := \quad \inf_{\pi \in \Pi} U(\sigma, \pi), \tag{88}$$

then

(a) For all $\sigma \in \Sigma$, the infimum in Equation 88 is achieved for some π^*.

(b) The function $U(\sigma)$ is nonnegative, bounded, concave and continuous on σ.

(c) If Q is not a proper nonlocality proof, then for all $\sigma \in \Sigma, U(\sigma) = 0$. If Q is a proper nonlocality proof, then $U(\sigma) > 0$ for all *σ in the interior of Σ.*

(d) For a 2 party, 2 measurement settings per party nonlocality proof, we further have that, even if Q is proper, then still $U(\sigma) = 0$ for all σ on the boundary of Σ.

3. Suppose that σ is in the interior of Σ, then:

(a) Let Q be a $2 \times 2 \times 2$ nonlocality proof. Suppose that Q is non-trivial in the sense that, for some a, b, Q_{ab} is not a point mass (i.e. $0 < Q_{ab}(x, y) < 1$ for some x, y). Then $\pi^* \in \Pi$ achieves the infimum in Equation 88 if and only if the following 16 (in)equalities hold:

$$\sum_{a,b \in \{1,2\}} \sigma_{ab} \frac{Q_{ab}(x_a, y_b)}{P_{ab;\pi^*}(x_a, y_b)} \quad = \quad 1 \tag{89}$$

for all $(x_1, x_2, y_1, y_2) \in \{\mathbf{F}, \mathbf{T}\}^4$ such that $\pi^*_{x_1,x_2,y_1,y_2} > 0$, and

$$\sum_{a,b \in \{1,2\}} \sigma_{ab} \frac{Q_{ab}(x_a, y_b)}{P_{ab;\pi^*}(x_a, y_b)} \quad \leq \quad 1 \tag{90}$$

for all $(x_1, x_2, y_1, y_2) \in \{\mathbf{F}, \mathbf{T}\}^4$ such that $\pi^*_{x_1,x_2,y_1,y_2} = 0$.
For generalized nonlocality proofs, $\pi^* \in \Pi$ achieves Equation 88 if and only if the corresponding analogues of Equations 89 and 90 both hold.

(b) Suppose that π^* and π° both achieve the infimum in Equation 88. Then, for all $x, y \in \{\mathbf{F}, \mathbf{T}\}$, $a, b \in \{1, 2\}$ with $Q_{ab}(x, y) > 0$, we have $P_{ab;\pi^*}(x, y) = P_{ab;\pi^\circ}(x, y) > 0$. In words, π^* and π° coincide in every measurement setting for every measurement outcome that has positive probability according to Q_σ, and Q is absolutely continuous with respect to π^* and π°.

Proof We only give proofs for the $2 \times 2 \times 2$ case; extension to the general case is entirely straightforward. We define

$$U((a, b), \pi) \quad := \quad D(Q_{ab}(\cdot) \| P_{ab;\pi}(\cdot)) \tag{91}$$

$$= \sum_{\substack{x,y \in \{\mathbf{F},\mathbf{T}\} \\ Q_{ab}(x,y) > 0}} Q_{ab}(x, y)[\log Q_{ab}(x, y) - \log P_{ab;\pi}(x, y)]. \tag{92}$$

Note that $U(\sigma, \pi)$ can be written as $U(\sigma, \pi) = \sum_{a,b \in \{1,2\}} \sigma_{ab} U((a, b), \pi)$.

Part 1 Equation 87 follows directly from the additivity property of KL divergence, Equation 54. Convexity is immediate by Jensen's inequality applied to the logarithm in Equation 91 and the fact that $P_{ab;\pi}(x, y)$ is linear in $\pi_{x_1 x_2 y_1 y_2}$ for each $(x_1, x_2, y_1, y_2) \in \{\mathbf{F}, \mathbf{T}\}^4$. If π lies in the interior of Π, then $P_{ab;\pi}(x, y) > 0$ for $a, b \in \{1, 2\}$ so that $U(\sigma, \pi)$ is finite. Continuity and differentiability are then immediate by continuity and differentiability of $\log x$ for $x > 0$. Lower semicontinuity of $U(\sigma, \pi)$ on Π is implied by the fact that, on general spaces, $D(Q\|P)$ is jointly lower semi-continuous in Q and P in the weak topology, as proved by Posner [27, Theorem 2]. Part 1(c) is immediate.

Part 2 We have already shown that for fixed σ, $U(\sigma, \pi)$ is lower semicontinuous on Π. Lower semicontinuous functions achieve their infimum on a compact domain (see for example [11, page 84]), so that for each σ, Equation 88 is achieved for some π^*. This proves (a). To prove (b), note that nonnegativity of $U(\sigma)$ is immediate by nonnegativity of the KL divergence. Boundedness of $U(\sigma)$ follows by considering the uniform distribution π°, with, for all x_1, x_2, y_1, y_2, $\pi^\circ_{x_1 x_2 y_1 y_2} = 1/16$. π° is in Π, so that

$$U(\sigma) \quad \leq \quad U(\sigma, \pi^\circ) \tag{93}$$

$$= \sum_{a,b \in \{1,2\}} \sigma_{ab} \Bigg[\sum_{\substack{x,y \ \in \{\mathbf{F},\mathbf{T}\} \\ Q_{ab}(x,y) > 0}} Q_{ab}(x, y)[\log Q_{ab}(x, y) + 2] \Bigg] \tag{94}$$

$$\leq \; - \sum_{a,b\{1,2\}} \sigma_{ab} H(Q_{ab}) + 8, \tag{95}$$

where $H(Q_{ab})$ is the Shannon-entropy of the distribution Q_{ab}. Boundedness of $U(\sigma)$ now follows from the fact that $H(Q) \geq 0$ for every distribution Q, which is a standard result (see, e.g. [8]).

Let σ be in the interior of Σ and let $\pi^* \in \Pi$ achieve $\inf_{\pi\in\Pi} U(\sigma,\pi)$. Since $U(\sigma)$ is bounded, π^* is absolutely continuous with respect to σ (otherwise $U(\sigma) = \infty$, a contradiction). Thus, $U(\sigma)$ satisfies

$$U(\sigma) \quad = \inf_{\pi\in\Pi \text{ is absolutely continuous with respect to } Q} U(\sigma,\pi). \tag{96}$$

We already proved that if Q_σ is absolutely continuous with respect to π^*, then $U(\sigma,\pi^*)$ is linear in σ. Thus, by Equation 96, $U(\sigma)$ is an infimum of linear functions, which (by a standard result of convex analysis, see e.g. [30]) is concave. A concave and bounded function with a convex domain must be continuous on the interior of this domain (see, e.g., [30]). It remains to show that $U(\sigma)$ is continuous at boundary points of Σ. Showing this is straightforward by taking limits (but tedious). We omit the details.

Now for part (c). If Q is not a proper nonlocality proof, then by definition there exists a $\pi_0 \in \Pi$ such that, for $a, b \in \{1, 2\}$, we have $Q_{ab} = P_{ab;\pi_0}$ and hence $U(\sigma,\pi_0) = 0$ for all $\sigma \in \Sigma$.

Now suppose Q is a proper nonlocality proof. Let σ be in the interior of Σ. $\inf_\pi U(\sigma,\pi)$ is achieved for some π^*. Suppose, by means of contradiction, that $U(\sigma,\pi^*) = 0$. Since $\sigma_{ab} > 0$ for $a, b \in \{1, 2\}$, we must have $Q_{ab} = P_{ab;\pi^*}$ for $a, b \in \{1, 2\}$. But then Q is not a proper nonlocality proof; contradiction. For part (d), if σ is on the boundary of Σ, then for some a, b, $\sigma_{ab} = 0$. It then follows from Lemma 1 and the fact that, for all P, $D(P\|P) = 0$ that $U(\sigma,\pi^*) = 0$.

Part 3 Part (a) The condition that Q_{ab} is not a point mass for some a, b, implies that all π^* that achieve the infimum must have $\pi^*_{x_1x_2y_1y_2} < 1$ for all x_1, x_2, y_1, y_2, (otherwise $U(\sigma,\pi^*) = \infty$, which is a contradiction). Thus, we assume that $\pi^* \in \Pi_0$, with Π_0 the set of πs that satisfy this "< 1" restriction.

For $\rho \in [0,\infty)^{16}$, let

$$\overline{\rho}_{x_1x_2y_1y_2}(\rho) \quad := \quad \frac{\rho_{x_1x_2y_1y_2}}{\sum_{x_1',x_2',y_1',y_2'\in\{\mathbf{F},\mathbf{T}\}} \rho_{x_1'x_2'y_1'y_2'}}. \tag{97}$$

In this way, each vector ρ with at least one non-zero component uniquely defines a local theory $\overline{\rho} \in \Pi$, and

$$\left\{ \overline{\rho} : \rho \in [0,\infty)^{16} \text{ and } \sum_{x_1,x_2,y_1,y_2\in\{\mathbf{F},\mathbf{T}\}} \rho_{x_1x_2y_1y_2} > 0 \right\} \quad = \quad \Pi_0. \tag{98}$$

Let ρ^* be such that $\overline{\rho}^*$ achieves the infimum in Equation 88. Then Q is absolutely continuous with respect to $\overline{\rho}^*$. One can now show that for each $(x_1, x_2, y_1, y_2) \in \{\mathbf{F},\mathbf{T}\}^4$, the partial derivative $\partial\, U(\sigma,\overline{\rho})/\partial\rho_{x_1,x_2,y_1,y_2}$ evaluated at $\rho = \rho^*$ exists (even if $\rho^*_{x_1,x_2,y_1,y_2} = 0$). Since $\overline{\rho}^*$ achieves the infimum, it follows that, for each $(x_1, x_2, y_1, y_2) \in \{\mathbf{F},\mathbf{T}\}^4$, we must have that $(\partial/\partial\rho_{x_1,x_2,y_1,y_2})U(\sigma,\overline{\rho})$ evaluated at ρ^* is not less than 0, or, equivalently,

$$\left\{ \frac{\partial U(\sigma,\overline{\rho})}{\partial\rho_{x_1,x_2,y_1,y_2}} \right\}_{\rho=\rho^*} \cdot \left(\sum_{x_1,x_2,y_1,y_2} \rho_{x_1,x_2,y_1,y_2} \right) \quad \geq \quad 0 \tag{99}$$

with equality if $\rho^*_{x_1,x_2,y_1,y_2} > 0$. Straightforward evaluation of Equation 99 gives Equations 89 and 90. This shows that each π^* achieving Equation 88 satisfies Equations 89 and 90. On the other hand, each π^* corresponding to a ρ^* with $\overline{\rho}^* = \pi^*$ such that Equation 99 holds for each $(x_1, x_2, y_1, y_2) \in \{\mathbf{F},\mathbf{T}\}^4$ must achieve a local minimum of $U(\sigma,\pi)$ (viewed as a function of π), Since $U(\sigma,\pi)$ is convex, π^* must achieve the infimum of Equation 88.

For part (b), suppose, by way of contradiction, that for at least one $(x_1, y_1) \in \{\mathbf{F},\mathbf{T}\}^2$, $a_0, b_0 \in \{1, 2\}$ with $Q_{a_0b_0}(x_1, y_1) > 0$, we have $P_{a_0b_0;\pi^*}(x_1, y_1) \neq P_{a_0b_0;\pi^\circ}(x_1, y_1)$. For each $x, y \in \{\mathbf{F},\mathbf{T}\}, a, b \in \{1, 2\}$, we can write

$$\begin{aligned} P_{ab;\pi^*}(x,y) &= \pi^*_{k_1} + \pi^*_{k_2} + \pi^*_{k_3} + \pi^*_{k_4}, \\ P_{ab;\pi^\circ}(x,y) &= \pi^\circ_{k_1} + \pi^\circ_{k_2} + \pi^\circ_{k_3} + \pi^\circ_{k_4}, \end{aligned} \tag{100}$$

for some $k_1, \ldots, k_4$ depending on x, y, a, b. Here each k_j is of the form $x_1x_2y_1y_2$ with $x_i, y_i \in \{\mathbf{F},\mathbf{T}\}$. Now consider $\pi^+ := (1/2)\pi^* + (1/2)\pi^\circ$. Clearly $\pi^+ \in \Pi$. By Jensen's inequality applied to the logarithm and using Equation 100, we have for $a, b \in \{1, 2\}$: $Q_{ab}(x,y)[\log Q_{ab}(x,y) - \log P_{ab;\pi^+}(x,y)] \leq Q_{ab}(x,y)[\log Q_{ab}(x,y) - \frac{1}{2}\log P_{ab;\pi^*}(x,y) - \frac{1}{2}\log P_{ab;\pi^\circ}(x,y)]$, where the inequality is strict if $x = x_1, y =$

$y_1, a = a_0$ and $b = b_0$. But then for $a, b \in \{1, 2\}$: $U((a,b), \pi^+) \leq \frac{1}{2}U((a,b), \pi^*) + \frac{1}{2}U((a,b), \pi^\circ)$, which for $(a,b) = (a_0, b_0)$ must be strict. By assumption, $\sigma_{a_0 b_0} > 0$ But that implies $U(\sigma, \pi^+) < U(\sigma, \pi^*) = \inf_{\pi \in \Pi} U(\sigma, \pi)$ and we have arrived at the desired contradiction. □

E.3 Proofs of Game-Theoretic Theorems

E.3.1 Game-Theoretic Preliminaries

Proposition 1 gives a few standard game-theoretic results (partially copied from [11]). We will use these results at several stages in later proofs.

Proposition 1 *Let A and B be arbitrary sets and let $L : A \times B \to \mathbb{R} \cup \{-\infty, \infty\}$ be an arbitrary function on $A \times B$. We have*

1. $\inf_{\beta \in B} \sup_{\alpha \in A} L(\alpha, \beta) \geq \sup_{\alpha \in A} \inf_{\beta \in B} L(\alpha, \beta)$.
2. *Suppose the following conditions hold:*
   - (a) *The game (A, B, L) has a value $V \in \mathbb{R} \cup \{-\infty, \infty\}$, that is $\inf_{\beta \in B} \sup_{\alpha \in A} L(\alpha, \beta) = V = \sup_{\alpha \in A} \inf_{\beta \in B} L(\alpha, \beta)$.*
   - (b) *There exists α^* that achieves $\sup_{\alpha \in A} \inf_{\beta \in B} L(\alpha, \beta)$.*
   - (c) *There exists β^* that achieves $\inf_{\beta \in B} \sup_{\alpha \in A} L(\alpha, \beta)$.*

   Then (α^*, β^*) is a saddle point and $L(\alpha^*, \beta^*) = V$.
3. *Suppose there exists a pair (α^*, β^*) such that*
   - (a) *β^* achieves $\inf_{\beta \in B} L(\alpha^*, \beta)$ and*
   - (b) *β^* is an* equalizer strategy*, that is, there exists a $K \in \mathbb{R} \cup \{-\infty, \infty\}$ with for all $\alpha \in A$, $L(\alpha, \beta^*) = K$.*

   Then the game (A, B, L) has value K, i.e. $\inf_{\beta \in B} \sup_{\alpha \in A} L(\alpha, \beta) = \sup_{\alpha \in A} \inf_{\beta \in B} L(\alpha, \beta) = K$, and (α^*, β^*) is a saddle point.

Proof (1) For all $\alpha' \in A$,

$$\inf_{\beta \in B} \sup_{\alpha \in A} L(\alpha, \beta) \quad \geq \quad \inf_{\beta \in B} L(\alpha', \beta). \tag{101}$$

Therefore, $\inf_{\beta \in B} \sup_{\alpha \in A} L(\alpha, \beta) \geq \sup_{\alpha' \in A} \inf_{\beta \in B} L(\alpha', \beta)$.

(2) Under our assumptions,

$$\begin{aligned} L(\alpha^*, \beta^*) \leq \sup_{\alpha \in A} L(\alpha, \beta^*) \quad &= \quad \inf_{\beta \in B} \sup_{\alpha \in A} L(\alpha, \beta) && (102) \\ &= \quad V && (103) \\ &= \quad \sup_{\alpha \in A} \inf_{\beta \in B} L(\alpha, \beta) && (104) \\ &= \quad \inf_{\beta \in B} L(\alpha^*, \beta) \leq L(\alpha^*, \beta^*), && (105) \end{aligned}$$

so $L(\alpha^*, \beta^*) = V = \inf_{\beta \in B} L(\alpha^*, \beta)$ and $L(\alpha^*, \beta^*) = V = \sup_{\alpha \in A} L(\alpha, \beta^*)$.

(3) To show that the game has a value, by (1) it is sufficient to show that

$$\inf_{\beta \in B} \sup_{\alpha \in A} L(\alpha, \beta) \quad \leq \quad \sup_{\alpha \in A} \inf_{\beta \in B} L(\alpha, \beta). \tag{106}$$

But this is indeed the case: $\inf_{\beta \in B} \sup_{\alpha \in A} L(\alpha, \beta) \leq \sup_{\alpha \in A} L(\alpha, \beta^*) = L(\alpha^*, \beta^*) = K = \inf_{\beta \in B} L(\alpha^*, \beta) \leq \sup_{\alpha \in A} \inf_{\beta \in B} L(\alpha, \beta)$, where the first equalities follow because β^* is an equalizer strategy. Thus, the game has a value equal to K. Since $\sup_{\alpha \in A} L(\alpha, \beta^*) = K$, β^* achieves $\inf_{\beta \in B} \sup_{\alpha \in A} L(\alpha, \beta)$. Since $\inf_{\beta \in B} L(\alpha^*, \beta) = K$, α^* achieves $\sup_{\alpha \in A} \inf_{\beta \in B} L(\alpha, \beta)$. Therefore, (α^*, β^*) is a saddle point. □

E.3.2 Proof of Theorem 2, the Saddle Point Theorem for Correlated Settings and Generalized Non-Locality Proofs

Theorem 2: *For every (generalized) non-locality proof, the correlated game (Π, Σ, U) corresponding to it has a finite value, i.e. there exists $0 \leq V < \infty$ with*

$$\begin{aligned} V &= \inf_{\Pi} \sup_{\Sigma} U(\sigma, \pi) \\ &= \sup_{\Sigma} \inf_{\Pi} U(\sigma, \pi). \end{aligned} \tag{107}$$

The infimum on the first line is achieved for some $\pi^* \in \Pi$; the supremum on the second line is achieved for some σ^* in Σ, so that (π^*, σ^*) is a saddle point.

Proof We use the following well-known minimax theorem due to Ferguson. The form in which we state it is a straightforward combination of Ferguson's [11] Theorem 1, page 78 and Theorem 2.1, page 85, specialized to the Euclidean topology.

Theorem 4 (Ferguson 1967) *Let (A, B, L) be a statistical game where A is a finite set, B is a convex compact subset of $\mathbb{R}^k$ for some $k > 0$ and L is such that for all $\alpha \in A$,*

1. *$L(\alpha, \beta)$ is a convex function of $\beta \in B$.*
2. *$L(\alpha, \beta)$ is lower semicontinuous in $\beta \in B$.*

Let $\mathcal{A}$ be the set of distributions on A and define, for $P \in \mathcal{A}$, $L(P, \beta) = E_P L(\alpha, \beta) = \sum_{\alpha \in A} P_\alpha L(\alpha, \beta)$. Then the game $(\mathcal{A}, B, L)$ has a value V, i.e.

$$\sup_{P \in \mathcal{A}} \inf_{\beta \in B} L(P, \beta) = \inf_{\beta \in B} \sup_{P \in \mathcal{A}} L(P, \beta), \tag{108}$$

and a minimax $\beta^* \in B$ achieving $\inf_{\beta \in B} \sup_{\alpha \in \mathcal{A}} L(\alpha, \beta)$ exists.

By Theorem 1, part (1), for all $\sigma \in \Sigma$, $U(\sigma, \pi) = D(Q_\sigma \| P_{\sigma;\pi})$ is lower semicontinuous in π. Let us now focus on the case of a $2 \times 2 \times 2$ game. We can apply Theorem 4 with $A = \{11, 12, 21, 22\}$, $\mathcal{A} = \Sigma$ and $B = \Pi$. It follows that the game (Σ, Π, U) has a value V, and $\inf_\pi \sup_\sigma U(\sigma, \pi) = V$ is achieved for some $\pi^* \in \Pi$. By Theorem 1, part (2), $0 \leq V < \infty$, and, since $U(\sigma)$ is continuous in σ, there exists some σ^* achieving $\sup_\sigma \inf_\pi U(\sigma, \pi)$.

The proof for generalized nonlocality proofs is completely analogous; we omit details. □

E.3.3 Proof of Theorem 3, Saddle Points and Equalizer Strategies for $2 \times 2 \times 2$ Nonlocality Proofs

Theorem 3: *Fix any proper non-locality proof based on 2 parties with 2 measurement settings per party and let (Σ, Π, U) and $(\Sigma^{\text{UC}}, \Pi, U)$ be the corresponding correlated and uncorrelated games, then:*

1. *The correlated game has a saddle point with value $V > 0$. Moreover,*

$$\sup_{\Sigma^{\text{UC}}} \inf_{\Pi} U(\sigma, \pi) \leq \sup_{\Sigma} \inf_{\Pi} U(\sigma, \pi) = V \tag{109}$$

$$\inf_{\Pi} \sup_{\Sigma^{\text{UC}}} U(\sigma, \pi) = \inf_{\Pi} \sup_{\Sigma} U(\sigma, \pi) = V. \tag{110}$$

2. *Let*

$$\Pi^* := \{\pi : \pi \text{ achieves } \inf_{\Pi} \sup_{\Sigma} U(\sigma, \pi)\}, \tag{111}$$

$$\Pi^{\text{UC}*} := \{\pi : \pi \text{ achieves } \inf_{\Pi} \sup_{\Sigma^{\text{UC}}} U(\sigma, \pi)\}, \tag{112}$$

   then

   (a) Π^* is non-empty.

   (b) $\Pi^* = \Pi^{\text{UC}*}$.

   (c) All $\pi^* \in \Pi^*$ are 'equalizer strategies', i.e. for all $\sigma \in \Sigma, U(\sigma, \pi^*) = V$.

3. *The uncorrelated game has a saddle point if and only if there exists (π^*, σ^*), with $\sigma^* \in \Sigma^{\text{UC}}$, such that*

   (a) π^* achieves $\inf_\pi U(\sigma^*, \pi)$.

(b) π^* is an equalizer strategy.

If such (σ^*, π^*) exists, it is a saddle point.

Proof The correlated game has a value V by Theorem 2 and $V > 0$ by Theorem 1. Inequality 109 is immediate.

Let $U((a,b),\pi)$ be defined as in the proof of Theorem 1 (Equation 91). To prove Equation 110, note that for every $\pi \in \Pi$,

$$\begin{aligned} \sup_{\Sigma^{\text{UC}}} U(\sigma,\pi) &= \sup_{\Sigma} U(\sigma,\pi) && (113)\\ &= \max_{a,b\in\{1,2\}} U((a,b),\pi). && (114) \end{aligned}$$

Thus, Equation 110 and part 2(b) of the theorem follow. Part 2(a) is immediate from Theorem 2. To prove part 2(c), suppose, by way of contradiction, that there exists a $\pi^* \in \Pi^*$ that is not an equalizer strategy. Then the set $\{(a,b) \mid U((a,b),\pi^*) = \max_{a,b\in\{1,2\}} U((a,b),\pi^*)\}$ has less than four elements. By Theorem 2, there exists a $\sigma^* \in \Sigma$ such that (σ^*, π^*) is a saddle point in the correlated game. Since $\sigma^* \in \Sigma$ achieves $\sup_{\Sigma} U(\sigma,\pi^*)$, it follows that for some $a_0, b_0 \in \{1,2\}$, $\sigma^*_{a_0 b_0} = 0$. But then σ^* lies on the boundary of σ. By Theorem 1, part 2(d), this is impossible, and we have arrived at the desired contradiction.

It remains to prove part (3). Part (3), 'if' follows directly from Proposition 1. To prove part (3), 'only if', suppose the uncorrelated game has saddle point (σ^*, π^*). It is clear that π^* achieves $\inf_{\pi} U(\sigma^*, \pi)$. We have already shown above that π^* is an equalizer strategy. □