Unifying Framework for Relaxations of the Causal Assumptions in Bell's Theorem

R. Chaves, R. Kueng, J. B. Brask, and D. Gross

Phys. Rev. Lett. **114**, 140403 — Published 7 April 2015

DOI: 10.1103/PhysRevLett.114.140403

A unifying framework for relaxations of the causal assumptions in Bell's theorem

R. Chaves,[1] R. Kueng,[1] J.B. Brask,[2] and D. Gross[1]

[1] *Institute for Physics, University of Freiburg, Rheinstrasse 10, D-79104 Freiburg, Germany*

[2] *Département de Physique Théorique, Université de Genève, 1211 Genève, Switzerland*

(Dated: March 20, 2015)

Bell's Theorem shows that quantum mechanical correlations can violate the constraints that the causal structure of certain experiments impose on any classical explanation. It is thus natural to ask to which degree the causal assumptions – e.g. locality or measurement independence – have to be relaxed in order to allow for a classical description of such experiments. Here, we develop a conceptual and computational framework for treating this problem. We employ the language of Bayesian networks to systematically construct alternative causal structures and bound the degree of relaxation using quantitative measures that originate from the mathematical theory of causality. The main technical insight is that the resulting problems can often be expressed as computationally tractable linear programs. We demonstrate the versatility of the framework by applying it to a variety of scenarios, ranging from relaxations of the measurement independence, locality and bilocality assumptions, to a novel causal interpretation of CHSH inequality violations.

The paradigmatic Bell experiment [1] involves two distant observers, each with the capability to perform one of two possible experiments on their shares of a joint system. Bell observed that even absent of any detailed information about the physical processes involved, the *causal structure* of the setup alone implies strong constraints on the correlations that can arise from any *classical* description [2]. The physically well-motivated causal assumptions are: (i) *measurement independence*: experimenters can choose which property of a system to measure, independently of how the system has been prepared; (ii) *locality*: the results obtained by one observer cannot be influenced by any action of the other (ideally space-like separated) experimenter. The resulting constraints are Bell's inequalities [1]. Quantum mechanical processes subject to the same causal structure can violate these constraints – a prediction that has been abundantly verified experimentally [3]. This effect is commonly referred to as *quantum nonlocality*.

It is now natural to ask how stable quantum nonlocality is with respect to relaxations of the causal assumptions. Which degree of measurement dependence, e.g., is required to reconcile empirically observed correlations with a classical and local model? Such questions are not only, we feel, of great relevance to foundational questions – they are also of interest to practical applications of nonlocality, e.g. in cryptographic protocols. Indeed, eavesdroppers can (and do [4]) exploit the failure of a given cryptographic device to be constrained by the presumed causal structure to compromise its security. At the same time, it will often be difficult to ascertain that causal assumptions hold *exactly* – which makes it important to develop a systematic quantitative theory.

Several variants of this question have recently attracted considerable attention [5–13]. For example, measurement dependence has been found to be a very strong resource: only about about 1/15 of a bit of correlation between the source and measurements is sufficient to reproduce all correlations obtained by projective measurements on a singlet state [7, 9]. In turn, considering relaxations of the locality assumption, one bit of communication between the distant parties is again sufficient to simulate the correlations of singlet states [5].

In this paper we provide a unifying framework for treating relaxations of the measurement independence and locality assumptions in Bell's theorem. To achieve this, we borrow several concepts from the mathematical theory of *causality*, a relatively young subfield of probability theory and statistics [14, 15]. With the aim of describing the causal relations (rather than mere correlations) between variables that can be extracted from empirical observations, this community has developed a systematic and rigorous theory of causal structures and quantitative measures of causal influence.

Our framework rests on three observations: (i) Alternative causal structures can systematically be represented graphically via Bayesian networks [14]. There, variables are associated with nodes in a graph, and directed edges represent functional dependencies. (ii) These edges can be weighted by quantitative measures of causal influence [14, 16]. (iii) Determining the minimum degree of influence required for a classical explanation of observable distributions can frequently be cast as a computationally tractable linear program.

The versatility of this framework is demonstrated in a variety of applications. We give an operational meaning to the violation of the CHSH inequality [17] as the minimum amount of direct causal influence between the parties required to reproduce the observed correlations. Considering the Collins-Gisin scenario [18], we show that quantum correlations are incompatible with a classical description, even if we allow one of the parties to communicate its outcomes. We also show that the results in [7, 9] regarding measurement-independence relaxations can be improved by considering different Bell scenarios. Finally, we study the bilocality assumption [19] and show that although it defines a non-convex set, its relaxation can also be cast as a linear program, naturally quantifying the degree of nonbilocality.

Bayesian networks and measures for the relaxation of causal assumptions— The causal relationships between n jointly distributed discrete random variables $(X_1, \ldots, X_n)$ are specified by means of a *directed acyclic graph* (DAG). To this end, each variable is associated with one node of the graph. One then says that the X_i's form a *Bayesian network* with respect to the graph, if every variable can be expressed as a deterministic function $X_i = f_i(\mathrm{PA}_i, N_i)$ of its graph-theoretic parents PA_i and an unobserved noise term N_i, such that the N_i's are jointly independent. This is the case if and only if the probability $p(\mathbf{x}) = p(x_1, \ldots, x_n)$ is of the form

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i|\mathrm{pa}_i). \quad (1)$$

This identity encodes the causal relationships implied by the DAG [14].

As a paradigmatic example of a DAG, consider a bipartite Bell scenario (Fig. 1a). In this scenario, two separated observers, Alice and Bob, each perform measurements according to some inputs, here represented by random variables X and Y respectively, and obtain outcomes, represented by A and B. The causal model involves an explicit shared hidden variable Λ which mediates the correlations between A and B. From (1) it follows that $p(x, y, \lambda) = p(x)p(y)p(\lambda)$, which reflects the measurement independence assumption. It also follows that $a = f_A(x, \lambda, n_A)$, $b = f_B(y, \lambda, n_B)$. We incur no loss of generality by absorbing the local noise terms N_A, N_B into Λ and will thus assume from now on that $a = f_A(x, \lambda), b = f_B(y, \lambda)$ for suitable functions f_A, f_B. This encodes the locality assumption. Together, these relations imply the well-known local hidden variable (LHV) model of Bell's theorem:

$$p(a, b|x, y) = \sum_{\lambda} p(a|x, \lambda)p(b|y, \lambda)p(\lambda). \quad (2)$$

Causal mechanisms relaxing locality (Fig. 1b–d) and measurement independence (Fig. 1e) can be easily expressed using Bayesian networks. The networks themselves, however, do not directly quantify the degree of relaxation. Thus, one needs to devise ways of checking and quantifying such causal dependencies. To define a sensible measure of causal influence we introduce a core concept from the causality literature – *interventions* [14].

An intervention is the act of forcing a variable, say X_i, to take on some given value x_i' and is denoted by $do(x_i')$. The effect is to erase the original mechanism $f_i(pa_i, n_i)$ and place X_i under the influence of a new mechanism that sets it to the value x_i' while keeping all other functions f_j for $j \neq i$ unperturbed. The intervention $do(x_i')$ changes the decomposition (1), given by [20]

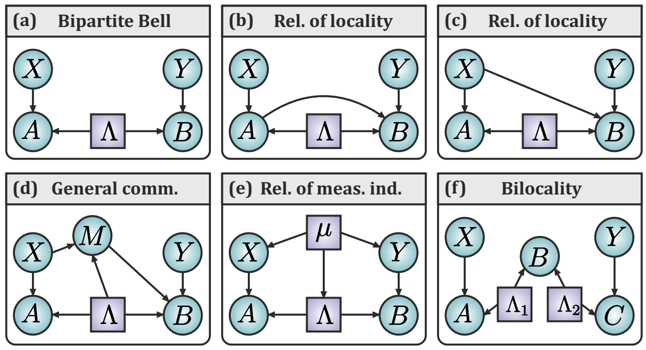

FIG. 1. **(a)** LHV model for the bipartite Bell scenario. **(b)** A relaxation of locality, where A may have direct causal influence on B. **(c)** Another relaxation in which X may have direct causal influence on B. **(d)** The most general communication scenario from Alice to Bob. **(e)** A relaxation of measurement independence. **(f)** The bilocality scenario for which the two sources Λ_1 and Λ_2 are assumed to be independent.

$$p(\mathbf{x}|do(x_i')) = \begin{cases} \prod_{j\neq i}^{n} p(x_j|\mathrm{pa}_j) & \text{if } x_i = x_i', \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

Considering locality relaxations, we can now define a measure $\mathcal{C}_{A\to B}$ for the *direct causal influence* of A into B for the model in Fig. 1b:

$$\mathcal{C}_{A\to B} = \sup_{b,y,a,a'} \sum_{\lambda} p(\lambda)|p(b|do(a), y, \lambda) - p(b|do(a'), y, \lambda)|. \quad (4)$$

It is the maximum shift (averaged over the unobservable Λ) in the probability of B caused by interventions in A. Similarly, one can define $\mathcal{C}_{X\to B}$ for the DAG in Fig. 1c and in other situations. This measure is strictly larger than zero for any underlying causal influence, as opposed to variations of it, such as the widely used *average causal effect* that can be null even in the presence of causal influences [16]. We are also interested in relaxations of measurement independence. Considering the case of a bipartite scenario (illustrated in Fig. 1e and that can be easily extended to multipartite versions), we can define the measure

$$\mathcal{M}_{X,Y:\lambda} = \sum_{x,y,\lambda} |p(x, y, \lambda) - p(x, y)p(\lambda)|. \quad (5)$$

This can be understood as a measure of how much the inputs are correlated with the source, i.e. how much the underlying causal model fails to comply with measurement independence. In the following we focus on the case where $p(x, y) = p(x)p(y)$, as usual in a typical Bell scenario.

The linear programming framework—Given some observed probabilities and a particular measure of relaxation, our aim is to compute the minimum value of the measure compatible with the observations. As sketched below and fully detailed in the Supplemental Material [21], this leads to a tractable linear program.

For simplicity we consider the usual Bell scenario of Fig. 1a. The most general observable quantity is the joint distribution $p(a,b,x,y) = p(a,b|x,y)p(x)p(y)$. Since we control the "inputs" X and Y, their distribution carries no information and we may thus restrict attention to $p(a,b|x,y)$. This conditional probability is, in turn, a linear function of the distribution of Λ. To make this explicit, represent $p(a,b|x,y)$ as a vector $\mathbf{p}$ with components $\mathbf{p}_j$ labeled by the multi-index $j = (a,b,x,y)$. Similarly, identify the distribution of Λ with a finite vector [21] with components $\mathbf{q}_\lambda = p(\Lambda = \lambda)$. Then from the discussion above, we have that $\mathbf{p} = T\mathbf{q}$ where T is a matrix with elements $T_{j,\lambda} = \delta_{a,f_A(x,\lambda)}\delta_{b,f_B(y,\lambda)}$. Conditional expectations that include the application of a *do*-operation are obtained via a modified T matrix. E.g., $\mathbf{q}'_j = p(a,b|x,y,do(a')) = T'q$ for $T'_{j,\lambda} = \delta_{a,a'}\delta_{b,f_B(y,\lambda)}$. The measures $\mathcal{C}$ and $\mathcal{M}$ are easily seen to be convex functions of the conditional probabilities $p(a,b|x,y)$ and their variants arising from the application of *do*'s – and thus convex functions of $\mathbf{q}$. Hence their minimization subject to the linear constraint $T\mathbf{q} = \mathbf{p}$ for an empirically observed distribution $\mathbf{p}$ is a convex optimization problem. This remains true if only some linear function $V\mathbf{p} = VT\mathbf{q}$ (e.g. a Bell inequality) of the distribution $\mathbf{p}$ is constrained. The problem is not manifestly a (computationally tractable) linear program (LP), since neither objective function is linear in $\mathbf{q}$. However, we establish in [21] that it can be cast as such:

Theorem 1. *The minimization of the measures $\mathcal{C}$ and $\mathcal{M}$ over models involving only one independent hidden variable, subject to any linear observation, can be reformulated as a primal linear program (LP). Its solution is equivalent to*

$$\max_{1\le i\le K} \langle \mathbf{v}_i, V\mathbf{p}\rangle, \tag{6}$$

where the $\{\mathbf{v}_i\}_{i=1}^K$ are the vertices of the LP's dual feasible region.

We highlight that (6) is a *closed-form expression in the observations $V\mathbf{p}$*: It is a maximum over finitely many explicit linear functions $V\mathbf{p} \mapsto \langle \mathbf{v}_i, V\mathbf{p}\rangle$. In this way, our result goes significantly beyond previous approaches [8–11], where generally only information about the degree of violation of a specific Bell inequality is utilized. In the following sections, we apply our framework to a variety of applications.

Novel causal interpretation of the CHSH inequality— Intuitively, the more nonlocal a given distribution is, the more direct causal influence between Alice and Bob should be required to simulate it. We make this intuition precise by considering the models in Fig. 1b–c and the CHSH scenario (two inputs, two outputs for Alice and Bob). For any observed distribution $p(a,b|x,y)$, we establish in [21] that

$$(1/2)\min \mathcal{C}_{A\to B} = \min \mathcal{C}_{X\to B} = \max\left[0, \mathrm{CHSH}\right], \tag{7}$$

where the maximum is taken over all the eight symmetries of the CHSH quantity [17]

$$\begin{aligned}\mathrm{CHSH} =\, & p(00|00) + p(00|01) + p(00|10) \\ & - p(00|11) - p^A(0|0) - p^B(0|0),\end{aligned} \tag{8}$$

where the last two terms represent the marginals for Alice and Bob. The CHSH inequality stipulates that for any LHV model, CHSH ≤ 0. Eq. (7) shows that, regardless of the particular distribution, the minimum direct causal influence is exactly quantified by the CHSH violation.

Inspired by the communication scenario of [5] (Fig. 1d) and the operational interpretation of CHSH violation given in [6], we can also quantify the relaxation of the locality assumption as the minimum amount of communication required to simulate a given distribution. We measure the communication by the Shannon entropy $H(m)$ of the message m which is sent. For a binary message, we can use our framework to prove, in complete analogy to (7), that

$$\min H(m) = h(\mathrm{CHSH}) \tag{9}$$

if CHSH > 0 and 0 otherwise. Here $h(v) = -v\log_2 v - (1-v)\log_2(1-v)$ denotes the binary entropy. We note that for maximal quantum violation CHSH $= 1/\sqrt{2} - 1/2$, as produced by a singlet state, a message with $H(m) \approx 0.736$ bits is required. This is less than the ≈ 0.85 bits of communication (after compression) required by the protocol of [5] for reproducing arbitrary correlations of a singlet.

Quantum nonlocality is incompatible with some locality relaxations— Given that violation of CHSH can be directly related to relaxation of locality, one can ask, whether similar interpretations exists for other scenarios. For example, we can consider a setting with three inputs and two outputs for Alice and Bob, and consider the causal model in Fig. 1b. Similar to the usual LHV model (2), the correlations compatible with this model form a polytope. One facet of this polytope is

$$\langle E_{00}\rangle - \langle E_{02}\rangle - \langle E_{11}\rangle + \langle E_{12}\rangle - \langle E_{20}\rangle + \langle E_{21}\rangle \le 4, \tag{10}$$

where $E_{xy} = \langle A_x B_y\rangle = \sum_{a,b}(-1)^{a+b}p(a,b|x,y)$. This inequality can be violated by any quantum state $|\psi\rangle = \sqrt{\epsilon}|00\rangle + \sqrt{(1-\epsilon)}|11\rangle$ with $\epsilon \ne 0, 1$. Consequently, any pure entangled state – no matter how close to separable – generates correlations that cannot be explained even if we allow for a relaxation of the locality assumption, where one of the parties communicates its measurement outcomes to the other.

How much measurement dependence is required to causally explain nonlocal correlations?— The results in

Refs. [7, 9] show that measurement dependence is a very strong resource for simulating nonlocality. In fact, a mutual information as small as $I(X,Y:\lambda)\approx 0.0663$ is already sufficient to simulate all correlations obtained by (any number of) projective measurements on a singlet state [9]. Given the fundamental implication and practical relevance of increasing these requirements, we aim to find larger values for $I(X,Y:\lambda)$ by means of our framework. The result of [9] leaves us with three options, regarding the quantum states: either non-maximally entangled states of two qubits, two-qudit states, or states with more than two parties.

Regarding non-maximally entangled two-qubit states, we were unable to improve the minimal mutual information. Regarding qudits, we have considered relaxations in the CGLMP scenario [29] – a bipartite scenario, where Alice and Bob each have two inputs and d outcomes. The CGLMP inequality is of the form $I_d\leq 2$. We have evaluated the LP for $\min\mathcal{M}$ in the setting of Fig. 1e, for various values of I_d and up to $d=8$. The numerical results strongly suggest that the simple relation

$$\min\mathcal{M}=\max\left[0,(I_d-2)/4\right] \tag{11}$$

holds. Via the Pinsker inequality [30, 31] and the definition of mutual information (see eq. (1) in [31] for further details), (11) provides a lower bound on the minimum mutual information $I(X,Y:\Lambda)\geq\mathcal{M}^2\log_2 \mathrm{e}/2$. This bound implies that for any $I_d\geq 3.214$, the mutual information required exceeds the 0.0663 obtained in Ref. [9]. Using the results in Ref. [32] for the scaling of the optimal quantum violation with d, one sees that this requires $d\geq 16$. However, we note that the bounds provided by the Pinsker inequality are usually far from tight, leaving a lot of room for improvement. Moreover a corresponding upper bound (obtained via the solution to the minimization of $\mathcal{M}$) is larger than the values obtained in [9] as soon as $d\geq 5$. Though this upper bound is not necessarily tight, we highlight the fact that for $d=2$ it gives exactly $I(X,Y:\Lambda)=0.0463$, the value analytically obtained in [9].

Regarding multipartite scenarios, we have considered GHZ correlations [33] in a tripartite scenario where each party has two inputs and two outputs. We numerically observe $0.090\leq I(X,Y,Z:\lambda)\leq 0.207$. This implies that increasing the number of parties can considerably increase the measurement dependence requirements for reproducing quantum correlations.

Bilocality scenario— To illustrate how our formalism can be used in generalized Bell scenarios [19, 34, 35], we briefly explore the entanglement-swapping scenario [36] of Fig. 1f (see details in [21]). The hidden variables in this scenario are independent $p(\lambda_1,\lambda_2)=p(\lambda_1)p(\lambda_2)$, the so-called *bilocality* assumption [19].

As in Ref. [19], we take the inputs x,z and outputs a,c to be dichotomic while b takes four values which we decompose in two bits as $b=(b_0,b_1)$. The distribution of hidden variables can be organized in a 64-dimensional vector $\mathbf{q}$ with components $q_{\alpha_0,\alpha_1,\beta_0,\beta_1,\gamma_0,\gamma_1}$, where α_x specifies the value of a for a given x (and analogously for γ, c and z) and β_i specifies the value of b_i. Thus together the indices label all the deterministic functions for A, B, C given their parents. As shown in [19], bilocality is equivalent to demanding $q^{ac}_{\alpha_0,\alpha_1,\gamma_0,\gamma_1}=q^a_{\alpha_0,\alpha_1}q^c_{\gamma_0,\gamma_1}$, where $q^{ac}_{\alpha_0,\alpha_1,\gamma_0,\gamma_1}=\sum_{\beta_0,\beta_1}q_{\alpha_0,\alpha_1,\beta_0,\beta_1,\gamma_0,\gamma_1}$ is the marginal for AC. Similar to (5) a natural measure $\mathcal{M}_{\mathrm{BL}}$ of nonbilocality quantifies by how much the underlying hidden variable distribution fails to comply with this constraint:

$$\mathcal{M}_{\mathrm{BL}}=\sum_{\alpha_0,\alpha_1,\gamma_0,\gamma_1}|q^{ac}_{\alpha_0,\alpha_1,\gamma_0,\gamma_1}-q^a_{\alpha_0,\alpha_1}q^c_{\gamma_0,\gamma_1}|. \tag{12}$$

Clearly $\mathcal{M}_{\mathrm{BL}}=0$, if and only if bilocality is fulfilled. However, demanding bilocality imposes a quadratic constraint on the hidden variables. This results in a nonconvex set which is extremely difficult to characterize [19, 34, 35]. Nevertheless, our framework is still useful, as using the marginals for a given observed distribution to constrain the problem further, the minimization of $\mathcal{M}_{\mathrm{BL}}$ can be cast in terms of a linear program with a single free parameter, which is further minimised over.

As an illustration we consider the nonbilocal distribution found in Refs. [19]. It is obtained by projective measurements on a pair of identical two-qubit entangled states $\varrho=v|\Psi^-\rangle\langle\Psi^-|+(1-v)\mathbb{I}/4$. This distribution violates the bilocality inequality $\mathcal{B}=\sqrt{|I|}+\sqrt{|J|}\leq 1$ giving a value $\mathcal{B}=\sqrt{2}v$. Using our framework we numerically observe $\mathcal{M}_{\mathrm{BL}}=\max(2v^2-1,0)$. Thus, for this specific distribution (and up to numerical precision), $\mathcal{M}_{\mathrm{BL}}=\mathcal{B}^2-1$, so there is a one-to-one correspondence between the violation of the bilocality inequality and the minimum relaxation of the bilocality constraint required to reproduce the correlations. This assigns an operational meaning to $\mathcal{B}$.

Conclusion— In this work we have revisited nonlocality from a causal inference perspective and provided a linear programming framework for relaxing the measurement independence and locality assumptions in Bell's theorem. Using the framework, we have given a novel causal interpretation of violations of the CHSH inequality, and shown that quantum correlations are still incompatible with classical causal models even if one allows for the communication of measurement outcomes. This implies that quantum nonlocality is even stronger than previously thought. Also, we have shown that the minimal measurement dependence required to simulated nonlocal correlations can be improved by considering different Bell scenarios. Finally we showed how our framework can be extended to treat the non-convex problem arising in the bilocality scenario. In particular, based on numerical evidence for a specific class of nonbilocal distributions, we have conjectured an operational meaning for the bilocality inequality.

In addition to these results, we believe the generality of our framework motivates and, more importantly, provides a basic tool for future research. For instance, it would be interesting to understand how our framework can be generalized in order to derive useful inequalities in the context of randomness expansion [10] Another natural possibility would be to look for a good measure of genuine multipartite nonlocality [37]. Finally, it would be interesting to understand how our treatment of the bilocality problem could be generalized and applied to the characterization of the non-convex compatibility regions of more complex quantum networks [34, 37, 38].

We thank R. Luce and D. Cavalcanti for useful discussions. Research in Freiburg is supported by the Excellence Initiative of the German Federal and State Governments (Grant ZUK 43), the Research Innovation Fund of the University of Freiburg, the US Army Research Office under contracts W911NF-14-1-0098 and W911NF-14-1-0133 (Quantum Characterization, Verification, and Validation), and the DFG. RK in addition acknowledges scholarship funds from the State Graduate Funding Program of Baden-Württemberg. JB was supported by the Swiss National Science Foundation (QSIT director's reserve) and SEFRI (COST action MP1006).

[1] J. S. Bell, Physics **1**, 195 (1964).

[2] Here the third assumption in Bell's theorem, *realism*, enters. It states that one can consistently assign a value to any physical property – independently of whether or not it is measured. In the Bayesian network language, this is expressed by the fact that variables are assumed to be deterministic functions of their parents.

[3] S. J. Freedman and J. F. Clauser, Phys. Rev. Lett. **28**, 938 (1972); A. Aspect, P. Grangier, and G. Roger, Phys. Rev. Lett. **49**, 91 (1982); B. G. Christensen, K. T. McCusker, J. B. Altepeter, B. Calkins, T. Gerrits, A. E. Lita, A. Miller, L. K. Shalm, Y. Zhang, S. W. Nam, N. Brunner, C. C. W. Lim, N. Gisin, and P. G. Kwiat, Phys. Rev. Lett. **111**, 130406 (2013); M. A. Rowe, D. Kielpinski, V. Meyer, C. A. Sackett, W. M. Itano, C. Monroe, and D. J. Wineland, Nature **409**, 791 (2001); M. Giustina, A. Mech, S. Ramelow, B. Wittmann, J. Kofler, J. Beyer, A. Lita, B. Calkins, T. Gerrits, S. W. Nam, *et al.*, *ibid.* **497**, 227 (2013).

[4] L. Lydersen, C. Wiechers, C. Wittmann, D. Elser, J. Skaar, and V. Makarov, Nature photonics **4**, 686 (2010).

[5] B. F. Toner and D. Bacon, Phys. Rev. Lett. **91**, 187904 (2003).

[6] S. Pironio, Phys. Rev. A **68**, 062102 (2003).

[7] J. Barrett and N. Gisin, Phys. Rev. Lett. **106**, 100406 (2011).

[8] M. J. W. Hall, Phys. Rev. A **82**, 062117 (2010).

[9] M. J. W. Hall, Phys. Rev. Lett. **105**, 250404 (2010); Phys. Rev. A **84**, 022102 (2011).

[10] D. E. Koh, M. J. W. Hall, Setiawan, J. E. Pope, C. Marletto, A. Kay, V. Scarani, and A. Ekert, Phys. Rev. Lett. **109**, 160404 (2012).

[11] M. Banik, Phys. Rev. A **88**, 032118 (2013); A. Rai, M. R. Gazi, M. Banik, S. Das, and S. Kunkri, Journal of Physics A: Mathematical and Theoretical **45**, 475302 (2012); B. Paul, K. Mukherjee, and D. Sarkar, Phys. Rev. A **88**, 014104 (2013); L. P. Thinh, L. Sheridan, and V. Scarani, Phys. Rev. A **87**, 062121 (2013).

[12] G. Pütz, D. Rosset, T. J. Barnea, Y.-C. Liang, and N. Gisin, arXiv preprint arXiv:1407.5634 (2014).

[13] K. Maxwell and E. Chitambar, Phys. Rev. A **89**, 042108 (2014).

[14] J. Pearl, *Causality* (Cambridge University Press, 2009).

[15] P. Spirtes, N. Glymour, and R. Scheienes, *Causation, Prediction, and Search, 2nd ed.* (The MIT Press, 2001).

[16] D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Scholkopf, The Annals of Statistics **41**, 2324 (2013).

[17] J. F. Clauser, M. A. Horne, A. Shimony, and R. A. Holt, Phys. Rev. Lett. **23**, 880 (1969).

[18] D. Collins and N. Gisin, Journal of Physics A: Mathematical and General **37**, 1775 (2004).

[19] C. Branciard, N. Gisin, and S. Pironio, Phys. Rev. Lett. **104**, 170401 (2010); C. Branciard, D. Rosset, N. Gisin, and S. Pironio, Phys. Rev. A **85**, 032119 (2012).

[20] We note that the *do*-operation is defined only relative to a causal model as encoded in the DAG (the graph structure enters (3) through the reference to parent nodes pa_j). In particular, $p(y|do(x))$ is in general different from the usual conditional probability $p(y|x)$ – these notions only coincide if the set of parents PA_X and PA_Y are disjoint.

[21] See Supplemental Material, which includes Refs. [22–28], for the proof of Theorem 1 and all technical details regarding the relaxation models displayed in Fig. 1.

[22] S. Boyd and L. Vandenberghe, *Convex optimization* (Cambridge university press, 2009).

[23] A. Barvinok, *A course in convexity* (American Mathematical Society, 2002).

[24] R. Webster, *Convexity* (Oxford University Press, USA, 1995).

[25] A. A. Balke and J. Pearl, *Probabilistic counterfactuals: semantics, computation, and applications*, Tech. Rep. (DTIC Document, 1997).

[26] T. Christof and A. Löbel, "PORTA – POlyhedron Representation Transformation Algorithm," (2009).

[27] T. Fritz and R. Chaves, IEEE Trans. Inform. Theory **59**, 803 (2013).

[28] R. Chaves, Phys. Rev. A **87**, 022102 (2013).

[29] D. Collins, N. Gisin, N. Linden, S. Massar, and S. Popescu, Phys. Rev. Lett. **88**, 040404 (2002).

[30] A. A. Fedotov, P. Harremoës, and F. Topsoe, Information Theory, IEEE Transactions on **49**, 1491 (2003).

[31] M. J. Hall, Entropy **15**, 3698 (2013).

[32] J.-L. Chen, C. Wu, L. C. Kwek, C. H. Oh, and M.-L. Ge, Phys. Rev. A **74**, 032106 (2006).

[33] D. M. Greenberger, M. Horne, and A. Zeilinger, "Bells theorem, quantum theory, and conceptions of the universe," (Kluwer Academic Publishers, Dordrecht, NL).

[34] T. Fritz, New Journal of Physics **14**, 103001 (2012).

[35] R. Chaves and T. Fritz, Phys. Rev. A **85**, 032113 (2012).

[36] M. Zukowski, A. Zeilinger, M. A. Horne, and A. K. Ekert, Phys. Rev. Lett. **71**, 4287 (1993).

[37] R. Gallego, L. E. Wurflinger, A. Acin, and M. Navascues, Phys. Rev. Lett. **109**, 070401 (2012); J.-D. Bancal,

J. Barrett, N. Gisin, and S. Pironio, Phys. Rev. A **88**, 014102 (2013); D. Cavalcanti, M. L. Almeida, V. Scarani, and A. Acin, Nature communications **2**, 184 (2011).

[38] R. Chaves, L. Luft, and D. Gross, New J. Phys. **16**, 043001 (2014); R. Chaves, L. Luft, T. O. Maciel, D. Gross, D. Janzing, and B. Schölkopf, Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence , 112 (2014); R. Chaves, C. Majenz, and D. Gross, Nat. Comm. **6**, 5766 (2015).