New Journal of Physics
The open access journal at the forefront of physics
Deutsche Physikalische Gesellschaft DPG IOP Institute of Physics

PAPER • OPEN ACCESS

The chain rule implies Tsirelson's bound: an approach from generalized mutual information

To cite this article: Eyuri Wakakuwa and Mio Murao 2012 *New J. Phys.* **14** 113037

New Journal of Physics
The open access journal for physics

The chain rule implies Tsirelson's bound: an approach from generalized mutual information

Eyuri Wakakuwa[1,3] **and Mio Murao**[1,2]
[1] Department of Physics, Graduate School of Science, The University of Tokyo, Tokyo 113-0033, Japan
[2] Institute for Nano Quantum Information Electronics, The University of Tokyo, Tokyo 113-0033, Japan
E-mail: wakakuwa@eve.phys.s.u-tokyo.ac.jp

New Journal of Physics **14** (2012) 113037 (24pp)
Received 11 July 2012
Published 27 November 2012
Online at http://www.njp.org/
doi:10.1088/1367-2630/14/11/113037

Abstract. In order to analyze an information theoretical derivation of Tsirelson's bound based on information causality, we introduce a generalized mutual information (GMI), defined as the optimal coding rate of a channel with classical inputs and general probabilistic outputs. In the case where the outputs are quantum, the GMI coincides with the quantum mutual information. In general, the GMI does not necessarily satisfy the chain rule. We prove that Tsirelson's bound can be derived by imposing the chain rule on the GMI. We formulate a principle, which we call the *no-supersignaling condition*, which states that the assistance of nonlocal correlations does not increase the capability of classical communication. We prove that this condition is equivalent to the no-signaling condition. As a result, we show that Tsirelson's bound is implied by the nonpositivity of the quantitative difference between information causality and no-supersignaling.

[3] Author to whom any correspondence should be addressed.

Contents

1. Introduction

One of the most counterintuitive phenomena that quantum mechanics predicts is nonlocality. The statistics of the outcomes of measurements carried out on an entangled state at two space-like separated points can exhibit strong correlations that cannot be described within the framework of local realism. This can be formulated in terms of the violation of Bell inequalities [1]. On the other hand, it is also known that quantum correlations still satisfy the no-signaling condition, i.e. they cannot be used for superluminal communication, which is prohibited by special relativity. The amount that quantum mechanics can violate the Clauser–Horne–Shimony–Holt inequality [2] is limited by Tsirelson's bound [8]. In a seminal paper [3], Popescu and Rohrlich showed that Tsirelson's bound is strictly lower than the limit imposed by the no-signaling condition alone. This result raises the question of why the strength of nonlocality is limited to Tsirelson's bound in the quantum world. If we could find an operational principle rather than a mathematical one to answer this question, it would help us better understand why quantum mechanics is the way it is [5–7].

From an information theoretical point of view, it is natural to ask if superstrong nonlocality, i.e. nonlocal correlations exceeding Tsirelson's bound, can be used to increase the capability of classical communication [4]. Suppose that Alice is trying to send classical information to distant Bob with the assistance of nonlocal correlations shared in advance. The no-signaling condition implies that, if no classical communication from Alice to Bob is performed, Bob's information gain is zero bits. In other words, zero bits of classical communication can produce not more than zero bits of classical information gain for the receiver. On the other hand, the no-signaling condition does not eliminate the possibility that $m > 0$ bits of classical communication produce more than m bits of classical information gain for the receiver. Whether such an implausible situation can occur would depend on the strength of nonlocal correlations. In particular, one might expect that Tsirelson's bound could be derived from the impossibility of such a situation.

Motivated by the foregoing considerations, *information causality* has been proposed as an answer to the question [4]. Information causality is the condition that *in bipartite nonlocality-assisted random access coding protocols, the receiver's total information gain cannot be greater than the amount of classical communication allowed in the protocol*. This condition is never violated in classical or quantum theory, whereas it is violated in all 'supernonlocal' theories, i.e. theories that predict supernonlocal correlations [4]. It implies that Tsirelson's bound is derived from this purely information theoretical principle. Thus information causality is regarded as one of the basic informational principles at the foundation of quantum mechanics.

In [4], it was proved that information causality is never violated in any no-signaling theory in which we can define *mutual information* satisfying five particular properties. This implies that in supernonlocal theories, we cannot define a function like the mutual information that satisfies all five. On the other hand, both the classical and quantum mutual information satisfy all of the five properties. It is therefore natural to ask another question: which of the five properties is lost in supernonlocal theories? We address this question to better understand the informational features of supernonlocal theories in comparison with quantum theory.

In order to answer this question, we need to define a generalization of the quantum mutual information that is applicable to general probabilistic theories. Several investigations have been made along this line. In [16, 17], a generalized entropy H is defined, and then mutual information is defined in terms of this by $I(A:B) := H(A) + H(B) - H(A, B)$. Using this mutual information, it is proved that the data processing inequality is not satisfied in supernonlocal theories. Similar results are obtained in [18, 19]. However, the definitions of the entropies in their approaches are mathematical, and do not have clear operational meanings. Note that in classical and quantum information theory, the operational meaning of entropy and mutual information is given by the source coding and channel coding theorems. In [17], a coding theorem analogous to Schumacher's quantum coding theorem [10] is investigated using generalized entropy. However, their consideration is only applicable under several restrictions. As discussed in [16], we need to seek generalizations based on the analysis of data compression or channel capacity. Such an approach is also studied in [9].

Motivated by these discussions, we introduce an operational definition of generalized mutual information (GMI) that is applicable to any general probabilistic theory. This is a generalization of the quantum mutual information between a classical system and a quantum system. Unlike the previous entropic approaches, we directly address the mutual information. The generalization is based on the channel coding theorem. Thus the GMI inherently has an operational meaning as a transmission rate of classical information. Our definition does not require mathematical notions such as state space or fine-grained measurement. The GMI is defined between a classical system and a general probabilistic system—it is not applicable to two general probabilistic systems, but it is sufficient for analyzing the situation describing information causality. The GMI satisfies four of the five properties of the mutual information, the exception being the chain rule. We will show that violation of Tsirelson's bound implies violation of the chain rule of the GMI.

Using the GMI, we further investigate the derivation of Tsirelson's bound in terms of information causality. We formulate a principle, which we call the *no-supersignaling condition*, stating that the assistance of nonlocal correlations does not increase the capability of classical communication. We prove that this condition is equivalent to the no-signaling condition, and thus it is different from information causality. This result is similar to the result obtained in [17], but now becomes operationally supported. It implies that Tsirelson's bound is not derived from the condition that 'm bits of classical communication cannot produce more than m bits of information gain'. We show that Tsirelson's bound is derived from the nonpositivity of the quantitative difference between information causality and no-supersignaling. Our results indicate that the chain rule of the GMI imposes a strong restriction on the underlying physical theory. As an example of this fact, we show that we can derive a bound on the state space of 1 gbit from the chain rule.

This paper is organized as follows. In section 2, we introduce a minimal framework for general probabilistic theories. In section 3, we give a brief review of information causality. In section 4, we define the GMI, and show that Tsirelson's bound is derived from the chain rule. In section 5, we prove that the GMI is a generalization of the quantum mutual information. In section 6, we formulate the no-supersignaling condition, and prove that the condition is equivalent to the no-signaling condition. In section 7, we clarify the relation among no-supersignaling, information causality and Tsirelson's bound. In section 8, we show that we can limit the state space of 1 gbit by assuming the chain rule. We conclude with a summary and discussion in section 9.

2. General probabilistic theories

In this section, we introduce a minimal framework for general probabilistic theories based on [17, 20].

We associate a set of allowed *states* $\mathcal{S}_S$ with each physical system S. We assume that any probabilistic mixture of states is also a state, i.e. if $\phi_1 \in \mathcal{S}_S$ and $\phi_2 \in \mathcal{S}_S$, then $\phi_{\text{mix}} = p\phi_1 + (1-p)\phi_2 \in \mathcal{S}_S$, where $p\phi_1 + (1-p)\phi_2$ denotes the state that is a mixture of ϕ_1 with probability p and ϕ_2 with probability $1-p$.

We also associate a set of allowed *measurements* $\mathcal{M}_S$ with each system S. A set of outcomes $\mathcal{R}_e$ is associated with each measurement $e \in \mathcal{M}_S$. The state determines the probability of obtaining an outcome $r \in \mathcal{R}_e$ when a measurement $e \in \mathcal{M}_S$ is performed on the system S. Thus we associate each outcome $r \in \mathcal{R}_e$ with a functional $e_r\colon \mathcal{S} \to [0, 1]$, such that $e_r(\phi)$ is the probability of obtaining outcome r when a measurement e is performed on a system in the state ϕ. Such a functional is called an *effect*. In order that the statistics of measurements on mixed states fits into our intuition, we require the linearity of each effect, i.e. $e_r(\phi_{\text{mix}}) = pe_r(\phi_1) + (1-p)e_r(\phi_2)$.

It may be possible to perform *transformations* on a system. A transformation on the system S is described by a map $\mathcal{E}\colon \mathcal{S}_S \to \mathcal{S}_{S'}$, where S' denotes the output system. We assume the linearity of transformations, i.e. $\mathcal{E}(\phi_{\text{mix}}) = p\mathcal{E}(\phi_1) + (1-p)\mathcal{E}(\phi_2)$. A measurement $e \in \mathcal{M}_S$ is represented by a transformation $\mathcal{E}_{\text{M}}\colon \mathcal{S}_S \to \mathcal{S}_{T_S}$, where T_S represents a classical system corresponding to the register of the measurement outcome. We assume that the composition of two allowed transformations is also an allowed transformation and that any allowed transformation followed by an allowed measurement is an allowed measurement.

We assume that a composition of two systems is also a system. If we have two systems A and B, we can consider a composite system AB which has its own set of allowed states $\mathcal{S}_{AB}$ and that of allowed measurements $\mathcal{M}_{AB}$. Suppose that measurements $e_A \in \mathcal{M}_A$ and $e_B \in \mathcal{M}_B$ are carried out on the systems A and B, respectively. Such a measurement is called a *product measurement* and is included in $\mathcal{M}_{AB}$. We assume that a global state $\psi \in \mathcal{S}_{AB}$ determines a

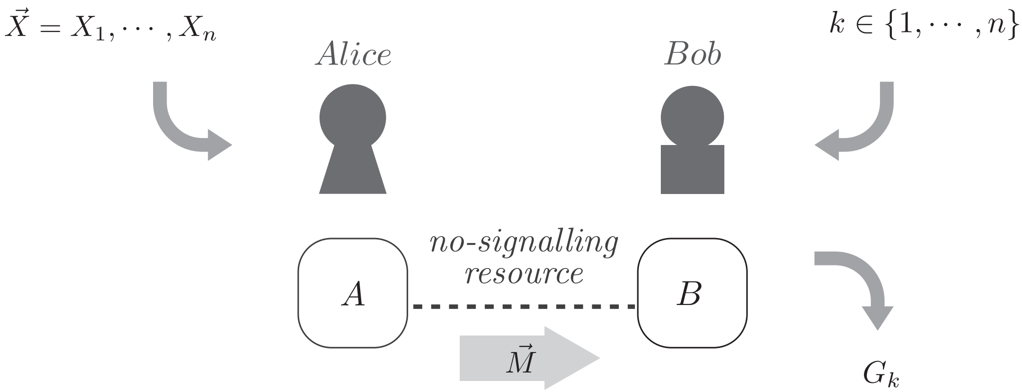

Figure 1. Nonlocality-assisted random access coding. The task is for Bob to correctly guess X_k, where k is a random number unknown to Alice.

joint probability for each pair of effects $(e_{A,r}, e_{B,r'})$. We may also assume that the global state is uniquely specified if the joint probabilities for all pairs of effects $(e_{A,r}, e_{B,r'})$ are specified. Such an assumption is called the *global state assumption*. However, it is known that there exist general probabilistic theories which do not fit into this assumption, such as quantum theory in a real Hilbert space. The arguments presented in the following sections of this paper are developed under the global state assumption, although the main results are valid without this assumption. The generalization for theories without this assumption is given in appendix B.

3. Review of information causality

Information causality, introduced in [4], is the principle that *the total amount of classical information gain that the receiver can obtain in a bipartite nonlocality-assisted random access coding protocol cannot be greater than the amount of classical communication that is allowed in the protocol*. Suppose that a string of n random and independent bits $\vec{X} = X_1, \ldots, X_n$ is given to Alice, and a random number $k \in \{1, \ldots, n\}$ is given to distant Bob. The task is for Bob to correctly guess X_k under the condition that they can use a resource of shared correlations and an m bit one-way classical communication from Alice to Bob (see figure 1). To accomplish this task, Alice first makes a measurement on her part of the resource (denoted by A in the figure), depending on $\vec{X}$. She then constructs an m bit message $\vec{M}$ from $\vec{X}$ and the measurement outcome, and sends it to Bob. Bob, after receiving $\vec{M}$, makes a measurement on his part of the resource (denoted by B in the figure), depending on $\vec{M}$ and k. From the outcome of the measurement he computes his guess G_k for X_k. The efficiency of the protocol is quantified by

$$J := \sum_{k=1}^{n} I_C(X_k : G_k), \tag{1}$$

where $I_C(X_k : G_k)$ is the classical (Shannon) mutual information between X_k and G_k. Information causality is the condition that, whatever strategy they take and whatever resource of shared correlation allowed in the theory they use,

$$J \leqslant m \tag{2}$$

must hold for all $m \geqslant 0$. The derivation of Tsirelson's bound in terms of information causality consists of the following two theorems that are proved in [4].

Figure 2. The channel defining the mutual information between the system X and the system S. It has a classical system as the input system and a general probabilistic system as the output system.

Theorem 3.1. *If we can define a function $I(A:B)$ satisfying the following five properties in the general probabilistic theory, $J \leqslant m$ holds for all $m \geqslant 0$. The properties are*

- Symmetry: $I(A:B) = I(B:A)$ *for any systems A and B.*
- Non-negativity: $I(A:B) \geqslant 0$ *for any systems A and B.*
- Consistency: *If both systems A and B are in classical states, $I(A:B)$ coincides with the classical mutual information.*
- Data processing inequality: *Under any local transformation that maps states of system B into states of another system B' without post-selection, $I(A:B) \geqslant I(A:B')$.*
- Chain rule: *For any systems A, B and C, the conditional mutual information defined by $I(A:B|C) := I(A:B,C) - I(A:C)$ is symmetric in A and B.*

Theorem 3.2. *If there exists a nonlocal correlation exceeding Tsirelson's bound, we can construct a nonlocality-assisted communication protocol by which $J > m$ is achieved.*

Theorem 3.1 guarantees that both classical and quantum theory satisfy information causality. Theorem 3.2 implies that information causality is violated in all supernonlocal theories. These two theorems imply that, in any supernonlocal theory, we cannot define a function of the mutual information that satisfies all five properties.

4. Generalized mutual information

Suppose that there are a classical system X and a system S that is described by a general probabilistic theory. The states of X are labeled by a finite alphabet $\mathcal{X}$. For each state x of X, the corresponding state of S denoted by ϕ_x is determined. The state of the composite system XS is determined by a probability distribution $p(x) = \Pr(X = x)$, which represents the probability that the system X is in the state x, and the corresponding state ϕ_x of S. Thus the state of the composite system XS is identified with an ensemble $\{p(x), \phi_x\}_{x\in\mathcal{X}}$. To define generalized mutual information $I_G(X:S)$ between the system X and the system S in the state $\{p(x), \phi_x\}_{x\in\mathcal{X}}$, we analyze the classical information capacity of a channel that outputs the system S in the state ϕ_x according to the input $X = x$ (figure 2). As usually considered in information theory, the sender Alice, who has access to X, tries to send classical information to the receiver Bob, who has access to S, by using the channel many times. Suppose that they use l identical and independent copies of this channel. Let $X_1, \ldots, X_l$ be the inputs of the l channels and $S_1, \ldots, S_l$ be the corresponding output systems.

Alice's encoding scheme is determined by a codebook. Let $w \in \{1, \ldots, N\}$ be a message that Alice tries to communicate, and the codeword $x^l(w) = x_1(w) \cdots x_l(w)$ be the corresponding input sequence to the channels. The codebook $\mathcal{C}$ is defined as the list of the codewords for all messages by

$$\mathcal{C} := \begin{bmatrix} x_1(1) & \cdots & x_l(1) \\ \vdots & \ddots & \vdots \\ x_1(N) & \cdots & x_l(N) \end{bmatrix}. \tag{3}$$

The letter frequency $f(x)$ for the codebook is defined by

$$f(x) := \frac{|\{(k, w)|x_k(w) = x, 1 \leqslant k \leqslant l, 1 \leqslant w \leqslant N\}|}{lN} \quad (x \in \mathcal{X}). \tag{4}$$

For a given probability distribution $\{p(x)\}_{x\in\mathcal{X}}$, the tolerance τ of the code is defined by

$$\tau := \max_{x\in\mathcal{X}} |p(x) - f(x)|. \tag{5}$$

By making a decoding measurement on the output systems $S_1, \ldots, S_l$, Bob tries to guess what the original message w is. Let $\mathcal{D}$ denote the decoding measurement. Note that, in general, the decoding measurement is not one in which Bob makes a measurement on each of $S_1, \ldots, S_l$ individually, but one in which the whole of the composite system $S_1 \cdots S_l$ is subjected to a measurement. Let W, $\hat{W}$ be Alice's original message and Bob's decoding outcome, respectively. The average error probability P_e is defined by

$$P_\mathrm{e} := \frac{1}{N}\sum_{u=1}^{N} \Pr(\hat{W} \neq u|W = u). \tag{6}$$

The pair of the codebook $\mathcal{C}$ and the decoding measurement $\mathcal{D}$ is called an (N, l) code. The ratio $\log N/l$ is called the rate of the code, and represents how many bits of classical information are transmitted per use of the channel.

Definition 4.1. A rate R is said to be achievable with $p(x)$ if there exists a sequence of $(2^{lR}, l)$ codes $(\mathcal{C}^{(l)}, \mathcal{D}^{(l)})$ such that

(i) $P_\mathrm{e}^{(l)} \to 0$ when $l \to \infty$,
(ii) $\tau^{(l)} \to 0$ when $l \to \infty$.

Definition 4.2. The mutual information between a classical system X and a general probabilistic system S, denoted by $I_\mathrm{G}(X : S)$, is the function which satisfies the condition that

(i) A rate R is achievable with $p(x)$ if $R < I_\mathrm{G}(X : S)$,
(ii) A rate R is achievable with $p(x)$ only if $R \leqslant I_\mathrm{G}(X : S)$.

We also define $I_\mathrm{G}(S : X)$ by $I_\mathrm{G}(S : X) := I_\mathrm{G}(X : S)$.

Theorem 4.1. *$I_\mathrm{G}(X : S)$ exists and satisfies $I_\mathrm{G}(X : S) \leqslant H(X)$. Here, $H(X)$ is the Shannon entropy of the system X defined by $H(X) := -\sum_{x\in\mathcal{X}} p(x) \log p(x)$.*

Proof. First we prove the existence of $R^* := \sup\{R|R$ is achievable with $p(x)\}$. Consider a $(2^{lR}, l)$ code and suppose that Alice's message $W = 1, \ldots, 2^{lR}$ is uniformly distributed. Let

I', H' be the mutual information and the entropy when the input sequence is the codeword corresponding to the uniformly distributed message W. By Fano's inequality, we have

$$H'(W|\hat{W}) \leqslant P_{\mathrm{e}}^{(l)} l R + 1, \tag{7}$$

where $P_{\mathrm{e}}^{(l)} = P(W \neq \hat{W})$. Thus

$$\begin{aligned} lR = H'(W) &= I'(W : \hat{W}) + H'(W|\hat{W}) \\ &\leqslant I'(X^l : \hat{W}) + P_{\mathrm{e}}^{(l)} l R + 1 \\ &\leqslant H'(X^l) + P_{\mathrm{e}}^{(l)} l R + 1. \end{aligned} \tag{8}$$

Here, we use the data processing inequality in the first inequality. By introducing a classical variable K that indicates k with the probability distribution $P(K = k) = 1/l$, we also have

$$H'(X^l) \leqslant \sum_{k=1}^{l} H'(X_k) = l H'(X|K) \leqslant l H'(X), \tag{9}$$

where X is a random variable defined by $\Pr(X = x_k(w)) = 2^{-lR}/l$. From (8) and (9), we obtain

$$P_{\mathrm{e}}^{(l)} \geqslant 1 - \frac{H'(X)}{R} - \frac{1}{lR}. \tag{10}$$

If R is achievable with $p(x)$, there exists a sequence of $(2^{lR}, l)$ codes satisfying $P_{\mathrm{e}}^{(l)} \to 0$ and $H'(X) \to H(X)$ when $l \to \infty$. Thus $R \leqslant H(X)$. Hence R^* exists and satisfies $R^* \leqslant H(X)$.

Next we prove that any rate $R < R^*$ is also achievable with $p(x)$. Let $\{(\mathcal{C}^{*(l)}, \mathcal{D}^{*(l)})\}_l$ be a sequence of $(2^{lR^*}, l)$ codes that satisfies $P_{\mathrm{e}}^{*(l)} \to 0$ and $\tau^{*(l)} \to 0$. For arbitrary $0 \leqslant \lambda < 1$, define another codebook $\mathcal{C}^{(l)}$ by using $\mathcal{C}^{*(\lambda l)}$ for the first λl codeletters and by choosing the last $(1-\lambda)l$ codeletters arbitrarily so that the total tolerance is sufficiently small. Also define the corresponding decoding measurement $\mathcal{D}^{(l)}$ as the measurement in which the output system $S_1 \cdots S_{\lambda l}$ is subjected to the decoding measurement $\mathcal{D}^{*(l)}$ and the output systems $S_{\lambda l+1}, \ldots, S_l$ are ignored. The code sequence $\{(\mathcal{C}^{(l)}, \mathcal{D}^{(l)})\}_l$ constructed in this way is a sequence of $(2^{l\lambda R^*}, l)$ codes that satisfies $P_{\mathrm{e}}^{(l)} \to 0$ and $\tau^{(l)} \to 0$. Thus $R = \lambda R^*$ is achievable with $p(x)$. Hence we obtain $R^* = I_{\mathrm{G}}(X : S)$. □

Note that $I_{\mathrm{G}}(X : S)$ is a function of the state $\Gamma := \{p(x), \phi_x\}_{x \in \mathcal{X}}$ of the composite system XS. To emphasize this, we sometimes use the notation $I_{\mathrm{G}}(X : S)_\Gamma$. Since $R = 0$ is always achievable, $I_{\mathrm{G}}(X : S)$ is nonnegative. Shannon's noisy channel coding theorem guarantees that $I_{\mathrm{G}}(X : S)$ coincides with the classical mutual information $I_{\mathrm{C}}(X : S)$ if S is a classical system [13]. The GMI satisfies the data processing inequality as follows.

Property 4.2. *Let $\mathcal{E}_{S \to S'}$ be any local transformation that maps states of a general probabilistic system S into states of another general probabilistic system S'. If $\mathcal{E}_{S \to S'}$ contains no post-selection, the GMI does not increase under this transformation, i.e. $I_{\mathrm{G}}(X : S) \geqslant I_{\mathrm{G}}(X : S')$. Similarly, $I_{\mathrm{G}}(X : S) \geqslant I_{\mathrm{G}}(X' : S)$ under any local transformation $\mathcal{E}_{X \to X'}$ that maps states of a classical system X into states of another classical system X' without post-selection.*

Proof. Here we only prove the former part. For the latter part, see appendix A. Consider two channels, channels I and II (see figure 3). Depending on the input $X = x$, channel I emits the

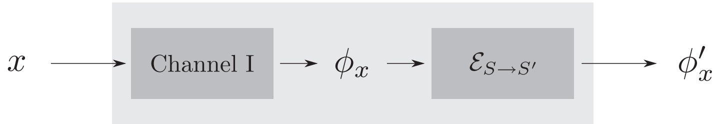

Figure 3. Channel II defined as the combination of channel I and $\mathcal{E}_{S\to S'}$.

system S in the state ϕ_x, and channel II emits the system S' in the state $\phi'_x = \mathcal{E}_{S\to S'}(\phi_x)$. It is only necessary to verify that if a rate R is achievable with $p(x)$ by channel II, R is also achievable with $p(x)$ by channel I. Let $\{(\mathcal{C}'^{(l)}, \mathcal{D}'^{(l)})\}_l$ be a sequence of $(2^{lR}, l)$ codes for channel II with the average error probability $P_{\mathrm{e}}'^{(l)}$ and the tolerance $\tau'^{(l)}$. From the code $(\mathcal{C}'^{(l)}, \mathcal{D}'^{(l)})$, construct a $(2^{lR}, l)$ code $(\mathcal{C}^{(l)}, \mathcal{D}^{(l)})$ for channel I by $\mathcal{C}^{(l)} = \mathcal{C}'^{(l)}$ and $\mathcal{D}^{(l)} = \mathcal{D}'^{(l)} \circ \mathcal{E}_{S\to S'}^{\otimes l}$. Here, $\mathcal{D}'^{(l)} \circ \mathcal{E}_{S\to S'}^{\otimes l}$ represents a process in which first $\mathcal{E}_{S\to S'}$ is applied to each of $S_1, \ldots, S_l$ individually and then the decoding measurement $\mathcal{D}'^{(l)}$ is carried out on the total output system $S'_1 \cdots S'_l$. The average error probability and the tolerance of this code are given by $P_{\mathrm{e}}^{(l)} = P_{\mathrm{e}}'^{(l)}$ and $\tau^{(l)} = \tau'^{(l)}$, respectively. Hence, if $P_{\mathrm{e}}'^{(l)} \to 0$ and $\tau'^{(l)} \to 0$, we also have $P_{\mathrm{e}}^{(l)} \to 0$ and $\tau^{(l)} \to 0$, and thus R is achievable with $p(x)$ by channel I. $\square$

In general probabilistic theories, a measurement on a system S without post-selection is described by a probabilistic map $\mathcal{E}_{\mathrm{M}}$ that maps states of S into states of a classical system T_S. T_S represents the register of the measurement outcomes. As a special case for property 4.2, we have $I_{\mathrm{G}}(X : T_S) \leqslant I_{\mathrm{G}}(X : S)$ under $\mathcal{E}_{\mathrm{M}}$, which is a generalization of Holevo's inequality. Let us define the accessible information $I_{\mathrm{acc}}(X : S)$ by

$$I_{\mathrm{acc}}(X : S) := \max I_{\mathrm{C}}(X : T_S), \tag{11}$$

where the maximization is taken over all possible measurements on S. Then we have $0 \leqslant I_{\mathrm{acc}}(X : S) \leqslant I_{\mathrm{G}}(X : S)$.

To summarize, the GMI satisfies the following properties.

- Symmetry: $I_{\mathrm{G}}(X : S) = I_{\mathrm{G}}(S : X)$.
- Non-negativity: $I_{\mathrm{G}}(X : S) \geqslant 0$
- Consistency: When S is a classical system, $I_{\mathrm{G}}(X : S) = I_{\mathrm{C}}(X : S)$.
- Data processing inequality: $I_{\mathrm{G}}(X : S) \geqslant I_{\mathrm{G}}(X' : S')$ under local stochastic maps $\mathcal{E}_{X\to X'}$ and $\mathcal{E}_{S\to S'}$ that contain no post-selection.

Thus, from theorems 3.1 and 3.2, we conclude that the chain rule of the GMI should be violated in any supernonlocal theory. Conversely, the chain rule implies Tsirelson's bound.

Throughout the rest of this paper, we use the GMI given by definition 4.2.

5. Quantum mutual information

The quantum mutual information between a classical system X and a quantum system S is defined by

$$I_{\mathrm{Q}}(X : S)_{\hat{\rho}} := H(S)_{\bar{\rho}} - \sum_{x\in\mathcal{X}} p(x) H(S)_{\hat{\rho}_x}, \tag{12}$$

where

$$\hat{\rho} = \sum_{x\in\mathcal{X}} p(x)|x\rangle\langle x|^X \otimes \hat{\rho}_x^S, \quad \langle x|x'\rangle = \delta_{xx'}, \tag{13}$$

$$\bar{\rho} = \sum_{x\in\mathcal{X}} p(x)\hat{\rho}_x, \tag{14}$$

and $H(S)$ is the von Neumann entropy. Note that, in quantum theory, a classical system is described by a Hilbert space in which we only consider a set of orthogonal pure states. With a slight generalization of the Holevo–Schumacher–Westmoreland theorem, it is shown that the GMI is a generalization of the quantum mutual information.

Theorem 5.1. *In quantum theory, the GMI coincides with the quantum mutual information, i.e.*

$$I_{\mathrm{G}}(X:S)_{\Gamma_{\hat{\rho}}} = I_{\mathrm{Q}}(X:S)_{\hat{\rho}}, \tag{15}$$

where

$$\hat{\rho} = \sum_{x\in\mathcal{X}} p(x)|x\rangle\langle x|^X \otimes \hat{\rho}_x^S \tag{16}$$

and $\Gamma_{\hat{\rho}} = \{p(x), \hat{\rho}_x\}_{x\in\mathcal{X}}$.

Proof. To prove this, it is only necessary to verify the following two statements:

(i) A rate R is achievable with $p(x)$ if $R < I_{\mathrm{Q}}(X:S)_{\hat{\rho}}$,

(ii) A rate R is achievable with $p(x)$ only if $R \leqslant I_{\mathrm{Q}}(X:S)_{\hat{\rho}}$.

The first statement is proved in [11–15] by using random code generation, and the second statement is proved in the following way. Consider a $(2^{lR}, l)$ code and suppose that Alice's message $W = 1, \ldots, 2^{lR}$ is uniformly distributed. Similarly to (8), we have

$$lR = H'(W) = I'(W:\hat{W}) + H'(W|\hat{W}) \leqslant I'_{\mathrm{Q}}(X^l:S^l) + P_{\mathrm{e}}^{(l)} lR + 1. \tag{17}$$

Here, we use the data processing inequality. We also have

$$\begin{aligned}
I'_{\mathrm{Q}}(X^l:S^l) &= H'(S^l) - H'(S^l|X^l) = H'(S^l) - \sum_{k=1}^{l} H'(S_k|X_k) \\
&\leqslant \sum_{k=1}^{l}(H'(S_k) - H'(S_k|X_k)) = \sum_{k=1}^{l} I'_{\mathrm{Q}}(X_k:S_k) \\
&= lI'_{\mathrm{Q}}(X:S|K) = lI'_{\mathrm{Q}}(X,K:S) - lI'_{\mathrm{Q}}(K:S) \\
&\leqslant lI'_{\mathrm{Q}}(X,K:S) = lI'_{\mathrm{Q}}(X:S).
\end{aligned} \tag{18}$$

In the first line, we use the fact that the state of S_k depends only on X_k. The first inequality is from the subadditivity of the von Neumann entropy. The last equality holds since $K \to X \to S$ forms a Markov chain. From (17) and (18), we obtain

$$P_{\mathrm{e}}^{(l)} \geqslant 1 - \frac{I'_{\mathrm{Q}}(X:S)}{R} - \frac{1}{lR}. \tag{19}$$

If R is achievable with $p(x)$, there exists a sequence of $(2^{lR}, l)$ codes satisfying $P_{\mathrm{e}}^{(l)} \to 0$ and $I'_{\mathrm{Q}}(X:S) \to I_{\mathrm{Q}}(X:S)_{\rho}$ when $l \to \infty$. Thus $R \leqslant I_{\mathrm{Q}}(X:S)_{\rho}$. □

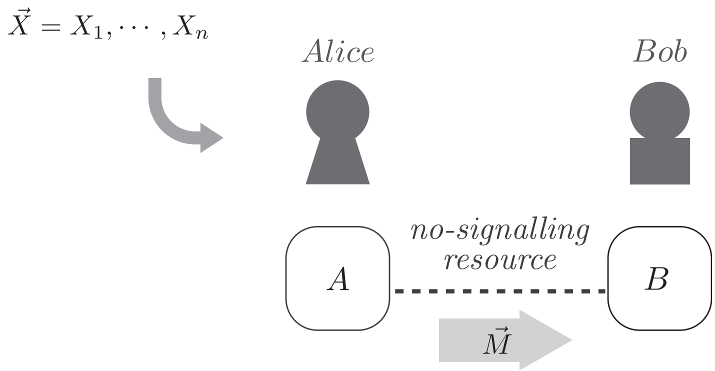

Figure 4. The situation that the no-supersignaling condition refers to. The amount of information about $\vec{X}$ contained in $\vec{M}$ and B is quantified by $I_G(\vec{X}, \vec{M} : B)$.

Figure 5. The channel that we consider to prove lemma 6.1. For each pair of the input $X = x$ and the output $Y = y$, the corresponding state ϕ_{xy} of the output system S is determined.

6. No-supersignaling condition

In this section, to further investigate the derivation of Tsirelson's bound from information causality, we formulate a principle that we call the *no-supersignaling condition* by using the GMI. Suppose that Alice is trying to send to distant Bob information about n independent classical bits $X_1, \ldots, X_n$, under the condition that they can only use an m bit classical communication $\vec{M}$ from Alice to Bob and a supplementary resource of correlations shared in advance (see figure 4). The situation is similar to the setting of information causality described in section 3, but now we do not introduce random access coding. Instead, we evaluate Bob's information gain by $I_G(\vec{X} : \vec{M}, B)$. We say that the no-supersignaling condition is satisfied if

$$I_G(\vec{X} : \vec{M}, B) \leqslant m \tag{20}$$

holds for all $m \geqslant 0$. The condition indicates that *the assistance of correlations cannot increase the capability of classical communication*. It is a direct formulation of the *original* concept of information causality that 'm bits of classical communication cannot produce more than m bits of information gain'. In what follows, we prove that the no-supersignaling condition is equivalent to the no-signaling condition. It indicates that information causality and no-supersignaling are different.

Lemma 6.1. *For any classical systems X, Y and any general probabilistic system S, if $I_{\mathrm{acc}}(X : S) = 0$ then $I_{\mathrm{acc}}(X : S, Y) \leqslant H(Y)$.*

Proof. Consider a channel with an input system X and two output systems S and Y (see figure 5). Let $\mathcal{Z}$ be the set of all measurements on S, and $p(t|x, y, z)$ be the probability of obtaining the outcome t when the measurement $z \in \mathcal{Z}$ is carried out on the system S in the state ϕ_{xy}. To achieve $I_{\mathrm{acc}}(X:S,Y)$, the receiver makes a measurement on S possibly depending on Y. Let $z(y)$ be the optimal choice of the measurement when $Y=y$. The probability of obtaining the outcome t when $X=x$ and $Y=y$ is given by

$$p_1(t|x,y) := p(t|x,y,z(y)). \tag{21}$$

We define

$$p_1(t,x,y) := p(x,y)p_1(t|x,y) = p(x,y)p(t|x,y,z(y)). \tag{22}$$

The condition $I_{\mathrm{acc}}(X:S)=0$ implies that for all $z\in\mathcal{Z}$,

$$\sum_y p(x,y)p(t|x,y,z) = p(x)p_2(t|z), \tag{23}$$

where

$$p_2(t|z) := \sum_{x,y} p(x,y)p(t|x,y,z). \tag{24}$$

Thus, we obtain

$$\begin{aligned} p_1(t,x,y) &= p(x,y)p(t|x,y,z(y)) \\ &\leqslant \sum_{y'} p(x,y')p(t|x,y',z(y)) \\ &= p(x)p_2(t|z(y)). \end{aligned} \tag{25}$$

The accessible information $I_{\mathrm{acc}}(X:S,Y)$ is equal to the mutual information $I_{\mathrm{C}}(X:T,Y)$ calculated for the probability distribution $p_1(t,x,y)$. Therefore

$$\begin{aligned} I_{\mathrm{acc}}(X:S,Y) &= I_{\mathrm{C}}(X:T,Y)_{p_1} \\ &= \sum_{t,x,y} p_1(t,x,y)\log\frac{p_1(t,x,y)}{p(x)p_1(t,y)} \\ &= H(Y) + \sum_{t,x,y} p_1(t,x,y)\log\frac{p_1(t,x,y)p(y)}{p(x)p_1(t,y)} \\ &\leqslant H(Y) + \sum_{t,x,y} p_1(t,x,y)\log\frac{p(x)p(y)p_2(t|z(y))}{p(x)p_1(t,y)} \\ &= H(Y) - \sum_{t,y} p_1(t,y)\log\frac{p_1(t,y)}{p_2(t,y)} \\ &= H(Y) - D(p_1(t,y)\|p_2(t,y)) \\ &\leqslant H(Y). \end{aligned}$$

In the first inequality, we used (25). In the next equality we defined a probability distribution $p_2(t,y) := p_2(t|z(y))p(y)$. The last inequality is from the non-negativity of the relative entropy. □

Theorem 6.1. *The no-supersignaling condition defined in terms of the GMI (20) is equivalent to the no-signaling condition.*

Proof. Consider a $(2^{lR}, l)$ code for the channel presented in figure 5 and let $X = \vec{X}$, $Y = \vec{M}$ and $S = B$. Suppose that Alice's message is uniformly distributed. By Fano's inequality, we have

$$I'(W : \hat{W}) \geqslant lR - 1 - P_{\mathrm{e}}^{(l)} lR. \tag{26}$$

By the data processing inequality, we also have

$$I'(W : \hat{W}) \leqslant I'(X^l : Y^l, T_{S^l}) \leqslant I'_{\mathrm{acc}}(X^l : Y^l, S^l). \tag{27}$$

From the no-signaling condition, we have $I'_{\mathrm{acc}}(X^l : S^l) = 0$. From lemma 6.1, we obtain

$$I'_{\mathrm{acc}}(X^l : Y^l, S^l) \leqslant H'(Y^l), \tag{28}$$

and thus

$$I'(W : \hat{W}) \leqslant H'(Y^l) \leqslant lH'(Y). \tag{29}$$

Hence, we obtain

$$\left(1 - P_{\mathrm{e}}^{(l)}\right) R \leqslant H'(Y) + \frac{1}{l}. \tag{30}$$

If R is achievable with $p(x)$, there exists a sequence of $(2^{lR}, l)$ codes that satisfies $P_{\mathrm{e}}^{(l)} \to 0$ and $H'(Y) \to H(Y)$ when $l \to \infty$. Thus, for any R that is achievable with $p(x)$, we have $R \leqslant H(Y)$. It implies $I_{\mathrm{G}}(X : Y, S) \leqslant H(Y)$ and thus $I_{\mathrm{G}}(\vec{X} : \vec{M}, B) \leqslant m$. Conversely, for $m = 0$, the no-supersignaling condition $I_{\mathrm{G}}(X : B) = 0$ implies the no-signaling condition. □

7. The difference between no-supersignaling and information causality

In this section, we discuss the relation between information causality, no-supersignaling, Tsirelson's bound and the chain rule. Let us define

$$\Delta_{\mathrm{NSS}} := I_{\mathrm{G}}(\vec{X} : \vec{M}, B) - m, \tag{31}$$

$$\Delta_{\mathrm{IC}} := J - m, \tag{32}$$

$$\Delta' := \Delta_{\mathrm{IC}} - \Delta_{\mathrm{NSS}} = J - I_{\mathrm{G}}(\vec{X} : \vec{M}, B). \tag{33}$$

Δ_{NSS} quantifies how much the capability of classical communication is increased with the assistance of nonlocal correlations. No-supersignaling is equivalent to $\Delta_{\mathrm{NSS}} \leqslant 0$, and information causality is equivalent to $\Delta_{\mathrm{IC}} \leqslant 0$. Δ' quantifies the difference between no-supersignaling and information causality.

Theorem 3.2 states that, if Tsirelson's bound is violated, we have $\Delta_{\mathrm{IC}} > 0$. Therefore violation of Tsirelson's bound implies at least either $\Delta_{\mathrm{NSS}} > 0$ or $\Delta' > 0$. Then which does violation of Tsirelson's bound imply, $\Delta_{\mathrm{NSS}} > 0$ or $\Delta' > 0$? As we proved in section 6, $\Delta_{\mathrm{NSS}} \leqslant 0$ is satisfied by all no-signaling theories. Thus violation of Tsirelson's bound only implies $\Delta' > 0$. Therefore, Tsirelson's bound is not derived from the condition that the assistance of nonlocal correlations does not increase the capability of classical communication. Instead, Tsirelson's bound is derived from the nonpositivity of Δ' (see figure 6). Let us further define

$$\Delta_{\mathrm{CR}} := \sum_{k=1}^{n} I_{\mathrm{G}}(X_k : \vec{M}, B, X_1, \ldots, X_{k-1}) - I_{\mathrm{G}}(\vec{X} : \vec{M}, B). \tag{34}$$

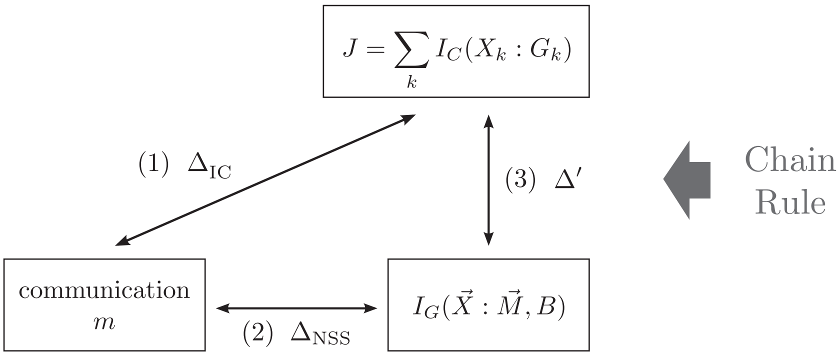

Figure 6. The relation between no-supersignaling and information causality, and the chain rule. Information causality refers to the gap in (1) represented by Δ_{IC}. No-supersignaling refers to the gap in (2) represented by Δ_{NSS}, and is irrelevant to Tsirelson's bound. The gap in (3) represented by Δ' is crucial in the derivation of Tsirelson's bound. Δ' is bounded above by zero if the chain rule is satisfied.

The chain rule is equivalent to $\Delta_{\mathrm{CR}} = 0$. By the data processing inequality, we always have $\Delta_{\mathrm{CR}} \geqslant \Delta'$. Thus the chain rule implies Tsirelson's bound[4] through imposing $\Delta' \leqslant \Delta_{\mathrm{CR}} = 0$.

Let X and Y be two classical systems and S be a general probabilistic system. The chain rule of the GMI is given by

$$I_{\mathrm{G}}(X, Y : S) + I_{\mathrm{G}}(X : Y) = I_{\mathrm{G}}(X : S) + I_{\mathrm{G}}(Y : S, X). \tag{35}$$

Each term in (35) has an operational meaning as an information transmission rate by definition. The relation is satisfied in both classical and quantum theory, but is violated in all supernonlocal theories. Thus we can conclude that this highly nontrivial relation gives a strong restriction on the underlying physical theories. However, the operational meaning of this relation is not clear so far.

8. Restriction on 1 gbit state space

To investigate how the chain rule of the GMI imposes a restriction on physical theories, we consider a gbit—the counterpart of a qubit in general probabilistic theories [15]. Here, we do not make assumptions about a gbit such as the dimension of the state space, or the possibility or impossibility of various measurements and transformations. Instead, we define a gbit as the minimum unit of information in the theory, and require that the classical information capacity of 1 gbit is not more than one bit. Thus we require that

$$I_{\mathrm{G}}(X : S_{1\mathrm{gb}}) \leqslant 1, \tag{36}$$

for any classical system X. When X is a classical system composed of two independent and uniformly random bits X_0 and X_1, we have

$$I_{\mathrm{G}}(X_0, X_1 : S_{1\mathrm{gb}}) \leqslant 1. \tag{37}$$

[4] Another way to show this is to observe that the data processing inequality and the no-supersignaling condition imply that $\Delta_{\mathrm{CR}} \geqslant \Delta_{\mathrm{IC}}$.

By the chain rule, we have

$$I_{\mathrm{G}}(X_0, X_1 : S_{1\mathrm{gb}}) = I_{\mathrm{G}}(X_0 : S_{1\mathrm{gb}}) + I_{\mathrm{G}}(X_1 : S_{1\mathrm{gb}}, X_0). \tag{38}$$

By the data processing inequality, we also have

$$I_{\mathrm{G}}(X_0 : S_{1\mathrm{gb}}) + I_{\mathrm{G}}(X_1 : S_{1\mathrm{gb}}, X_0) \geqslant I_{\mathrm{acc}}(X_0 : S_{1\mathrm{gb}}) + I_{\mathrm{acc}}(X_1 : S_{1\mathrm{gb}}). \tag{39}$$

Thus the chain rule implies

$$I_{\mathrm{acc}}(X_0 : S_{1\mathrm{gb}}) + I_{\mathrm{acc}}(X_1 : S_{1\mathrm{gb}}) \leqslant 1. \tag{40}$$

We consider success probabilities of the decoding measurements on $S_{1\mathrm{gb}}$ for X_0 and X_1. For simplicity, we assume that the optimal measurement carried out on $S_{1\mathrm{gb}}$ to decode X_0 or X_1 has two outcomes $t = 0, 1$. Let $P(t|m, x_0, x_1)$ be the probability of obtaining the outcome t when $X_0 = x_0$, $X_1 = x_1$ and the measurement m is made. The index $m = 0, 1$ corresponds to the optimal measurement for decoding X_0, X_1, respectively. The list of all probabilities $\{P(t|m, x_0, x_1)\}_{t,m,x_0,x_1=0,1}$ can be regarded as representing a 'state'. We compare the state space of a qubit and the state space determined by (40). For further simplicity, we assume that for all x_0 and x_1,

$$P(t = x_0|m = 0, x_0, x_1) = \frac{1+\alpha}{2} \quad (0 \leqslant \alpha \leqslant 1),$$

$$P(t = x_1|m = 1, x_0, x_1) = \frac{1+\beta}{2} \quad (0 \leqslant \beta \leqslant 1).$$

Then we have

$$\begin{aligned} I_{\mathrm{acc}}(X_0 : S_{1\mathrm{gb}}) &= I_{\mathrm{C}}(x_1 : t|m = 0) = 1 - H(x_0|t, m = 0) \\ &= 1 - H(x_0 \oplus t|m = 0) = 1 - h\left(\frac{1+\alpha}{2}\right), \end{aligned} \tag{41}$$

and

$$I_{\mathrm{acc}}(X_1 : S_{1\mathrm{gb}}) = 1 - h\left(\frac{1+\beta}{2}\right). \tag{42}$$

Here, $h(x)$ is the binary entropy defined by $h(x) := -x\log x - (1-x)\log(1-x)$. From (40)–(42), we have

$$h\left(\frac{1+\alpha}{2}\right) + h\left(\frac{1+\beta}{2}\right) \geqslant 1. \tag{43}$$

This inequality gives a restriction on the state space of 1 gbit (see figure 7). It is shown in appendix B that in the case of one qubit, the obtainable region is given by $\alpha^2 + \beta^2 \leqslant 1$.

9. Conclusions and discussions

We have defined a GMI between a classical system and a general probabilistic system. Since the definition is based on the channel coding theorem, the GMI inherently has an operational meaning as an information transmission rate. We showed that the GMI coincides

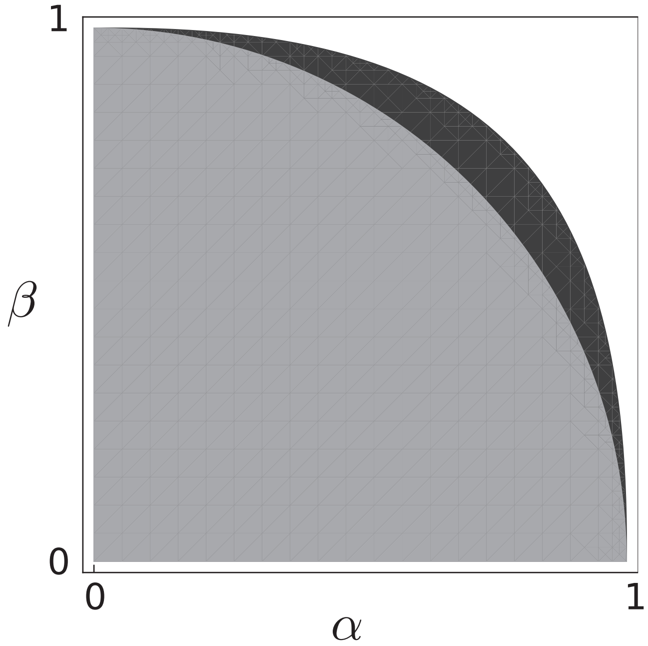

Figure 7. Comparison of the state space of a qubit and the boundary given by the chain rule. The gray region indicates the state space of a qubit given by $\alpha^2 + \beta^2 \leqslant 1$. The black region in addition to the gray region indicates the region defined by (43).

with the quantum mutual information if the output system is quantum. The GMI satisfies non-negativity, symmetry, the data processing inequality and the consistency with the classical mutual information, but does not necessarily satisfy the chain rule.

Using the GMI, we have analyzed the derivation of Tsirelson's bound from information causality defined in terms of the efficiency of nonlocality-assisted random access coding. We showed that the chain rule of the GMI, which is satisfied in both classical and quantum theory, is violated in any theory in which the existence of nonlocal correlations exceeding Tsirelson's bound is allowed. Thus we conclude that the chain rule of the GMI implies Tsirelson's bound.

We formulated a condition, the no-supersignaling condition, which states that the assistance of nonlocal correlations does not increase the capability of classical communication. We proved that this condition is equivalent to the no-signaling condition. We also clarified the relation among no-supersignaling, information causality, Tsirelson's bound and the chain rule.

The derivation of Tsirelson's bound from information causality proposed in [4] is remarkable in that Tsirelson's bound is exactly derived and that to do so we only need the five properties of the mutual information. However, information causality is different from the condition that 'm bits of classical communication cannot produce more than m bits of information gain'. This derivation shows that several laws of the Shannon theory[5], represented by the five properties of the mutual information, taken together impose a strong restriction on the underlying physical theory. If we take the GMI as the definition of the mutual information,

[5] By the Shannon theory we mean the theoretical framework composed of various theorems on the asymptotic coding rate of the sources and the channels.

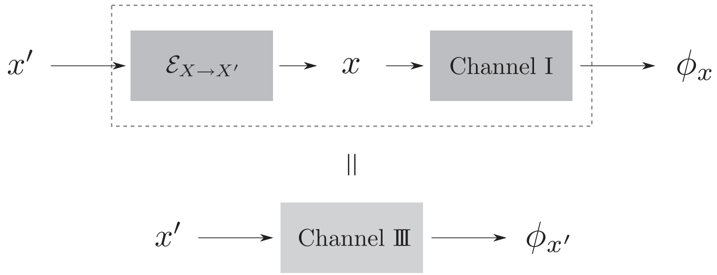

Figure A.1. Channel III defined as the combination of $\mathcal{E}_{X\to X'}$ and channel I. This channel as a whole is equivalent to a channel with the input x' and the output $\phi_{x'}$.

it reduces to the statement that 'a law of Shannon theory, namely the chain rule of the GMI, imposes a strong restriction on the underlying physical theory'.

Although the operational meaning of the GMI is clear, we have not yet succeeded in finding a clear operational meaning of the chain rule. In classical and quantum Shannon theory, the chain rule appears in a number of proofs of coding theorems. Therefore, investigation of the meaning of the chain rule would lead us to a better understanding of the informational foundations of quantum mechanics. On the other hand, our definition of the GMI is not the only way to generalize the quantum mutual information. It would also be fruitful to seek out other operationally motivated definitions of the GMI and compare them.

Acknowledgments

We thank T Sugiyama, P S Turner and S Beigi for useful discussions. We also thank the referees for their useful comments. This work was supported by Project for Developing Innovation Systems of the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. MM acknowledges support from JSPS of KAKENHI (grant no. 23540463).

Appendix A. Data processing inequality

We prove the latter part of theorem 4.2, which states that under any local stochastic map $\mathcal{E}_{X\to X'}$ that contains no post-selection, we have

$$I_{\mathrm{G}}(X:S) \geqslant I_{\mathrm{G}}(X':S). \tag{A.1}$$

The effect of $\mathcal{E}_{X\to X'}$ is determined by a conditional probability distribution $p_{\mathcal{E}}(x'|x)$, where x and x' denote the states of X and X', respectively. Let $\{p(x), \phi_x\}_{x\in\mathcal{X}}$ be the state of XS before applying $\mathcal{E}_{X\to X'}$. We can define probability distributions $p_{\mathcal{E}}(x,x') = p(x)p_{\mathcal{E}}(x'|x)$, $p(x') = \sum_x p_{\mathcal{E}}(x,x')$ and $p_{\mathcal{E}}(x|x') = p_{\mathcal{E}}(x,x')/p(x')$ for $x \in \mathcal{X}$ and $x' \in \mathcal{X}'$. The state of $X'S$ after applying $\mathcal{E}_{X\to X'}$ is $\{p(x'), \phi_{x'}\}_{x'\in\mathcal{X}'}$, where $\phi_{x'}$ is the mixture of ϕ_x with the probability given by $p_{\mathcal{E}}(x|x')$. We assume that $|\mathcal{X}|, |\mathcal{X}'| < \infty$.

To prove (A.1), consider two channels, channels I and III (see figure A.1). Channel I outputs the system S in the state ϕ_x according to the input $X = x$, and channel III outputs the system S in the state $\phi_{x'}$ according to the input $X' = x'$. It is only necessary to show that if a rate R is achievable with $p(x')$ by channel III, R is also achievable with $p(x)$ by channel I. Consider a sequence of $(2^{lR}, l)$ codes $(\mathcal{C}'^{(l)}, \mathcal{D}'^{(l)})$ for channel III that satisfies

(i) $P_{\mathrm{e}}'^{(l)} \to 0$ when $l \to \infty$,
(ii) $\tau'^{(l)} \to 0$ when $l \to \infty$.

Such a sequence exists if R is achievable with $p(x')$ by channel III. From the code $(\mathcal{C}'^{(l)}, \mathcal{D}'^{(l)})$, we randomly construct $(2^{lR}, l)$ codes $(\mathcal{C}^{(l)}, \mathcal{D}^{(l)})$ for channel I in the following way.

- For any w and k $(1 \leqslant w \leqslant 2^{lR}, 1 \leqslant k \leqslant l)$, generate the codeletter $x_k(w)$ randomly and independently according to the probability distribution $P(x_k(w) = x) = p_{\mathcal{E}}(x|x_k'(w))$.
- Regardless of the randomly generated codebook $\mathcal{C}^{(l)}$, use the same decoding measurement $\mathcal{D}^{(l)} = \mathcal{D}'^{(l)}$.

Let $P_{\mathrm{e}}^{\mathcal{C}^{(l)}}$ be the average error probability of the code $(\mathcal{C}^{(l)}, \mathcal{D}^{(l)})$ defined by

$$P_{\mathrm{e}}^{\mathcal{C}^{(l)}} := \frac{1}{2^{lR}} \sum_{u=1}^{2^{lR}} P(\hat{W} \neq u | W = u, \mathcal{C}^{(l)}). \tag{A.2}$$

Averaging $P_{\mathrm{e}}^{\mathcal{C}^{(l)}}$ over all codebooks $\mathcal{C}^{(l)}$ that are randomly generated, we obtain

$$\bar{P}_e^{(l)} := \sum_{\mathcal{C}^{(l)}} P(\mathcal{C}^{(l)})\, P_{\mathrm{e}}^{\mathcal{C}^{(l)}}, \tag{A.3}$$

where $P(\mathcal{C}^{(l)})$ is the probability of obtaining the codebook $\mathcal{C}^{(l)}$ as a result of random code generation. In lemma A.1, we show that $\bar{P}_{\mathrm{e}}^{(l)} \to 0$ in the limit of $l \to \infty$. In lemma A.2, we prove that for a sufficiently large l, the tolerance $\tau^{(l)}$ of the codebook $\mathcal{C}^{(l)}$ is almost equal to 0 with arbitrarily high probability. Finally, we give the proof for (A.1) in theorem A.3.

Lemma A.1.

$$\lim_{l \to \infty} \bar{P}_{\mathrm{e}}^{(l)} = 0. \tag{A.4}$$

Proof. $\bar{P}_{\mathrm{e}}^{(l)}$ defined by (A.3) is calculated as

$$\begin{aligned}
\bar{P}_{\mathrm{e}}^{(l)} &= \sum_{\mathcal{C}^{(l)}} P(\mathcal{C}^{(l)}) \times \frac{1}{2^{lR}} \sum_{u=1}^{2^{lR}} P(\hat{W} \neq u | W = u, \mathcal{C}^{(l)}) \\
&= \frac{1}{2^{lR}} \sum_{u=1}^{2^{lR}} \sum_{\mathcal{C}^{(l)}} P(\mathcal{C}^{(l)}) P(\hat{W} \neq u | W = u, \mathcal{C}^{(l)}) \\
&= \frac{1}{2^{lR}} \sum_{u=1}^{2^{lR}} \bar{P}(\hat{W} \neq u | W = u),
\end{aligned} \tag{A.5}$$

where

$$\bar{P}(\hat{W} \neq u | W = u) := \sum_{\mathcal{C}^{(l)}} P(\mathcal{C}^{(l)}) P(\hat{W} \neq u | W = u, \mathcal{C}^{(l)}). \tag{A.6}$$

The codebook $\mathcal{C}^{(l)}$ is determined by the codeletters $x_k(w)$ $(1 \leqslant w \leqslant 2^{lR}, 1 \leqslant k \leqslant l)$. Due to the way of randomly generating the code, the probability of obtaining the codebook $\mathcal{C}^{(l)}$ such that $x_k(w) = \xi_{wk}$ $(1 \leqslant w \leqslant 2^{lR}, 1 \leqslant k \leqslant l)$ is given by

$$\begin{aligned} P(\mathcal{C}^{(l)}) &= P(\{x_k(w)\}_{w,k} = \{\xi_{wk}\}_{w,k}) \\ &= \prod_{w=1}^{2^{lR}} \prod_{k=1}^{l} P(x_k(w) = \xi_{wk}) \\ &= \prod_{w=1}^{2^{lR}} \prod_{k=1}^{l} p_{\mathcal{E}}(x = \xi_{wk} | x' = x'_k(w)). \end{aligned} \tag{A.7}$$

Let $D(\phi_{x_1} \cdots \phi_{x_l})$ be the result of the decoding measurement $\mathcal{D}^{(l)}$ on the composite system $S_1 \cdots S_l$ in the state $\phi_{x_1} \cdots \phi_{x_l}$. We have

$$\begin{aligned} P(\hat{W} \neq u | W = u, \mathcal{C}^{(l)}) &= P(D(\phi_{x_1(u)} \cdots \phi_{x_l(u)}) \neq u | \{x_k(w)\}_{w,k} = \{\xi_{wk}\}_{w,k}) \\ &= P(D(\phi_{\xi_{u1}} \cdots \phi_{\xi_{ul}}) \neq u), \end{aligned} \tag{A.8}$$

and we obtain

$$\begin{aligned} \bar{P}(\hat{W} \neq u | W = u) &= \sum_{\{\xi_{wk}\}_{w,k}} P(D(\phi_{x_1(u)} \cdots \phi_{x_l(u)}) \neq u | \{x_k(w)\}_{w,k} = \{\xi_{wk}\}_{w,k}) \\ &\quad \times P(\{x_k(w)\}_{w,k} = \{\xi_{wk}\}_{w,k}) \\ &= \sum_{\{\xi_{uk}\}_k} P(D(\phi_{\xi_{u1}} \cdots \phi_{\xi_{ul}}) \neq u) \times P(\{x_k(u)\}_k = \{\xi_{uk}\}_k) \\ &= \sum_{\{\xi_{uk}\}_k} P(D(\phi_{\xi_{u1}} \cdots \phi_{\xi_{ul}}) \neq u) \times \prod_{k=1}^{l} p_{\mathcal{E}}(x = \xi_{uk} | x' = x'_k(u)). \end{aligned} \tag{A.9}$$

On the other hand, the error probability for the message w when channel III is used with the code $(\mathcal{C}'^{(l)}, \mathcal{D}'^{(l)})$ is given by

$$\begin{aligned} P'(\hat{W} \neq u | W = u) &= P(D(\phi_{x'_1(u)} \cdots \phi_{x'_l(u)}) \neq u) \\ &= \sum_{\{x_k\}_k} P(D(\phi_{x_1} \cdots \phi_{x_l}) \neq u) \times \prod_{k=1}^{l} p_{\mathcal{E}}(x = x_k | x' = x'_k(u)). \end{aligned} \tag{A.10}$$

From (A.9) and (A.10), we obtain that

$$\bar{P}(\hat{W} \neq u | W = u) = P'(\hat{W} \neq u | W = u), \tag{A.11}$$

and consequently

$$\bar{P}_{\mathrm{e}}^{(l)} = P'^{(l)}_{\mathrm{e}}. \tag{A.12}$$

Therefore $\bar{P}_{\mathrm{e}}^{(l)} \to 0$ when $l \to \infty$. □

Lemma A.2. $\tau^{(l)} \to 0$ *in probability in the limit of* $l \to \infty$.

Proof. Let $f(x)^{(l)}$ and $f(x')^{(l)}$ be the letter frequency of the codebook $\mathcal{C}^{(l)}$ and $\mathcal{C}'^{(l)}$, respectively. We have

$$\begin{aligned}
\left|f(x)^{(l)} - p(x)\right| &= \left|f(x)^{(l)} - \sum_{x'\in\mathcal{X}'} p_{\mathcal{E}}(x|x')p(x')\right| \\
&\leqslant \left|f(x)^{(l)} - \sum_{x'\in\mathcal{X}'} f(x')^{(l)} p_{\mathcal{E}}(x|x')\right| + \left|\sum_{x'\in\mathcal{X}'} f(x')^{(l)} p_{\mathcal{E}}(x|x') - \sum_{x'\in\mathcal{X}'} p_{\mathcal{E}}(x|x')p(x')\right| \\
&\leqslant \left|f(x)^{(l)} - \sum_{x'\in\mathcal{X}'} f(x')^{(l)} p_{\mathcal{E}}(x|x')\right| + \sum_{x'\in\mathcal{X}'} p_{\mathcal{E}}(x|x')\left|f(x')^{(l)} - p(x')\right|.
\end{aligned}$$

Define

$$f(x,x')^{(l)} := \frac{|\{(k,w)|x_k(w) = x,\, x'_k(w) = x',\, 1 \leqslant k \leqslant l,\, 1 \leqslant w \leqslant 2^{lR}\}|}{l \cdot 2^{lR}}$$

for $x \in \mathcal{X}$, $x' \in \mathcal{X}'$. By using the relation

$$f(x)^{(l)} = \sum_{x'\in\mathcal{X}'} f(x')^{(l)} \frac{f(x,x')^{(l)}}{f(x')^{(l)}}, \tag{A.13}$$

we obtain

$$\begin{aligned}
\Delta(x)^{(l)} :&= \left|f(x)^{(l)} - \sum_{x'\in\mathcal{X}'} f(x')^{(l)} p_{\mathcal{E}}(x|x')\right| \\
&\leqslant \sum_{x'\in\mathcal{X}'} f(x')^{(l)} \left|\frac{f(x,x')^{(l)}}{f(x')^{(l)}} - p_{\mathcal{E}}(x|x')\right|.
\end{aligned} \tag{A.14}$$

Applying the weak law of large numbers for each term in the sum, we have $\Delta(x)^{(l)} \to 0$ $(l \to \infty)$ in probability. We also have

$$\sum_{x'\in\mathcal{X}'} p_{\mathcal{E}}(x|x')\left|f(x')^{(l)} - p(x')\right| \leqslant \tau'^{(l)} \cdot |\mathcal{X}'| \tag{A.15}$$

and thus

$$\lim_{l\to\infty} \sum_{x'\in\mathcal{X}'} p_{\mathcal{E}}(x|x')\left|f(x')^{(l)} - p(x')\right| = 0. \tag{A.16}$$

Therefore, we obtain that

$$\tau^{(l)} = \max_x \left|f(x)^{(l)} - p(x)\right| \to 0 \quad \text{in probability.} \tag{A.17}$$

□

Theorem A.1. *R is achievable with $p(x)$ by channel I.*

Proof. Take arbitrary $\epsilon, \delta, \eta > 0$. From lemmas A.1 and A.2, for a sufficiently large l we have

$$\bar{P}_{\rm e}^{(l)} < \epsilon \tag{A.18}$$

and

$$\Pr\{\tau^{(l)} < \delta\} > 1 - \eta. \tag{A.19}$$

Define $C_\delta^{(l)} := \{\mathcal{C}^{(l)} | \tau^{(l)} < \delta\}$. The average error probability averaged over all codebooks in $C_\delta^{(l)}$ is calculated as

$$\frac{\sum_{\mathcal{C}^{(l)} \in C_\delta^{(l)}} P(\mathcal{C}^{(l)}) P_{\rm e}^{\mathcal{C}^{(l)}}}{\sum_{\mathcal{C}^{(l)} \in C_\delta^{(l)}} P(\mathcal{C}^{(l)})} = \frac{\bar{P}_{\rm e}^{(l)} - \sum_{\mathcal{C}^{(l)} \notin C_\delta^{(l)}} P(\mathcal{C}^{(l)}) P_{\rm e}^{\mathcal{C}^{(l)}}}{\sum_{\mathcal{C}^{(l)} \in C_\delta^{(l)}} P(\mathcal{C}^{(l)})} \leqslant \frac{\bar{P}_e^{(l)}}{\sum_{\mathcal{C}^{(l)} \in C_\delta^{(l)}} P(\mathcal{C}^{(l)})} < \frac{\epsilon}{1-\eta}.$$

Thus there exists at least one codebook $\mathcal{C}^{(l)} \in C_\delta^{(l)}$ such that $P_{\rm e}^{\mathcal{C}^{(l)}} < \epsilon' = \epsilon/(1-\eta)$ and, by definition, $\tau^{(l)} < \delta$. Hence there exists a sequence of $(2^{lR}, l)$ codes for channel I such that $P_{\rm e}^{(l)} \to 0$ and $\tau'^{(l)} \to 0$ when $l \to \infty$, and thus R is achievable with $p(x)$ by channel I. □

Appendix B. Beyond the global state assumption

In this appendix, we generalize the results presented in the main sections to general probabilistic theories which do not satisfy the global state assumption. Suppose that there are l independent copies of a channel that outputs the system S in the state ϕ_x according to the input $X = x$. If the input sequence is $x_1 \cdots x_l$, the state of the output system $S_1 \cdots S_l$ is $\phi_{x_1} \cdots \phi_{x_l}$. However, without the global state assumption, this does not specify the 'global' state of the composite system: it only specifies the state of the composite system for product measurements. Thus it is not sufficient to determine the rate of the channel. To avoid this difficulty, we introduce the notion of 'consistency' of the states. Let $\Phi_{x_1 \cdots x_l}$ be a global state of $S_1 \cdots S_l$. We say that $\Phi_{x_1 \cdots x_l}$ is consistent with $\phi_{x_1} \cdots \phi_{x_l}$ if the two states exhibit the same statistics for any product measurement. $\Phi^{(l)} := \{\Phi_{x_1 \cdots x_l}\}_{x_1 \cdots x_l \in \mathcal{X}^l}$ is said to be consistent with $\{\phi_{x_1} \cdots \phi_{x_l}\}_{x_1 \cdots x_l \in \mathcal{X}^l}$ if $\Phi_{x_1 \cdots x_l}$ is consistent with $\phi_{x_1} \cdots \phi_{x_l}$ for all $x_1 \cdots x_l \in \mathcal{X}^l$. With a slight abuse of terminology, we say that $\Phi := \{\Phi^{(l)}\}_{l=1}^\infty$ is consistent with $\{\phi_x\}_{x \in \mathcal{X}}$ if $\Phi^{(l)}$ is consistent with $\{\phi_{x_1} \cdots \phi_{x_l}\}_{x_1 \cdots x_l \in \mathcal{X}^l}$ for all l. Let $\Gamma_\Phi := \{\Gamma_\Phi^{(l)}\}_{l=1}^\infty$ be the sequence of the channel $\Gamma_\Phi^{(l)}$ that outputs the system $S_1 \cdots S_l$ in the state $\Phi_{x_1 \cdots x_l} \in \Phi^{(l)} \in \Phi$ according to the input $X_1 \cdots X_l = x_1 \cdots x_l$.

Definition B.1. A rate R is said to be achievable with $p(x)$ for Φ if there exists a sequence of $(2^{lR}, l)$ codes $(\mathcal{C}^{(l)}, \mathcal{D}^{(l)})$ for $\Gamma_\Phi^{(l)} \in \Gamma_\Phi$ such that

(i) $P_{\rm e}^{(l)} \to 0$ when $l \to \infty$,
(ii) $\tau^{(l)} \to 0$ when $l \to \infty$.

Definition B.2. A rate R is said to be achievable with $p(x)$ if R is achievable with $p(x)$ for all Φ that is consistent with $\{\phi_x\}_{x \in \mathcal{X}}$.

We define the GMI by definition 4.2 and its existence is proved by theorem 4.1. The data processing inequality (property 4.2) is proved as follows.

Proof. The inequality $I_{\mathrm{G}}(X:S) \geqslant I_{\mathrm{G}}(X:S')$ under local transformation $\mathcal{E}_{S\to S'}$ is proved as follows.

$$\begin{aligned}
I_{\mathrm{G}}(X:S') &= \sup\{R|R \text{ is achievable for all } \Phi' \text{ that is consistent with } \{\mathcal{E}(\phi_x)\}_{x\in\mathcal{X}}\}\\
&\leqslant \sup\{R|R \text{ is achievable for } \mathcal{E}(\Phi) \text{ for all } \Phi \text{ that is consistent with } \{\phi_x\}_{x\in\mathcal{X}}\}\\
&\leqslant \sup\{R|R \text{ is achievable for all } \Phi \text{ that is consistent with } \{\phi_x\}_{x\in\mathcal{X}}\}\\
&= I_{\mathrm{G}}(X:S).
\end{aligned} \tag{B.1}$$

Here, $\mathcal{E}(\Phi) := \{\mathcal{E}^{\otimes l}(\Phi^{(l)})\}_{l=1}^{\infty}$ and $\mathcal{E}^{\otimes l}(\Phi^{(l)}) := \{\mathcal{E}^{\otimes l}(\Phi_{x_1\cdots x_l})\}_{x_1\cdots x_l\in\mathcal{X}^l}$. The first inequality comes from the fact that $\mathcal{E}(\Phi)$ is consistent with $\{\mathcal{E}(\phi_x)\}_{x\in\mathcal{X}}$ if Φ is consistent with $\{\phi_x\}_{x\in\mathcal{X}}$. The second inequality is proved in the same way as the proof presented on page 8.

The inequality $I_{\mathrm{G}}(X:S) \geqslant I_{\mathrm{G}}(X':S)$ under local transformation $\mathcal{E}_{X\to X'}$ is proved as follows.

$$\begin{aligned}
I_{\mathrm{G}}(X':S) &= \sup\{R|R \text{ is achievable for all } \Phi' \text{ that is consistent with } \{\phi_{x'}\}_{x'\in\mathcal{X}'}\}\\
&\leqslant \sup\{R|R \text{ is achievable for } \Phi_{X'} \text{ for all } \Phi \text{ that is consistent with } \{\phi_x\}_{x\in\mathcal{X}}\}\\
&\leqslant \sup\{R|R \text{ is achievable for all } \Phi \text{ that is consistent with } \{\phi_x\}_{x\in\mathcal{X}}\}\\
&= I_{\mathrm{G}}(X:S)
\end{aligned} \tag{B.2}$$

Here, $\Phi_{X'} := \{\Phi_{X'}^{(l)}\}_{l=1}^{\infty}$ and $\Phi_{X'}^{(l)} := \{\Phi_{x'_1\cdots x'_l}\}_{x'_1\cdots x'_l\in\mathcal{X}'^l}$, where $\Phi_{x'_1\cdots x'_l}$ is the mixture of $\Phi_{x_1\cdots x_l} \in \Phi^{(l)} \in \Phi$ with the probability $\prod_{k=1}^{l} p_{\mathcal{E}}(x_k|x'_k)$. The first inequality comes from the fact that $\Phi_{X'}$ is consistent with $\{\phi_{x'}\}_{x'\in\mathcal{X}'}$ if Φ is consistent with $\{\phi_x\}_{x\in\mathcal{X}}$. The second inequality is proved in the same way as the proof in appendix A, where $\phi_{x_1}\cdots\phi_{x_l}$ is replaced by $\Phi_{x_1\cdots x_l}$. □

The equivalence of no-supersignaling and no-signaling (theorem 6.1) is proved as follows.

Proof. Due to the no-signaling condition, there exists Φ that is consistent with $\{\phi_{xy}\}_{x\in\mathcal{X},y\in\mathcal{Y}}$, and satisfies $I'_{\mathrm{acc}}(X^l:S^l)=0$ for all $\Gamma_\Phi^{(l)}\in\Gamma_\Phi$. Here, $\Gamma_\Phi^{(l)}$ is a channel with an input system X^l and two output systems Y^l and S^l. According to the input $X^l=x^l$, the channel outputs $Y^l=y^l$ with the probability $\prod_{k=1}^{l} p(y_k|x_k)$ and the system S^l in the state $\Phi_{x_1y_1\cdots x_ly_l}\in\Phi^{(l)}\in\Phi$. Consider a $(2^{lR},l)$ code for the channel. In the same way as the proof of theorem 6.1, we have $(1-P_{\mathrm{e}}^{(l)})R \leqslant H'(Y)+1/l$. If R is achievable with $p(x)$ for Φ, there exists a sequence of $(2^{lR},l)$ code for $\Gamma_\Phi^{(l)}$ that satisfies $P_{\mathrm{e}}^{(l)}\to 0$ and $H'(Y)\to H(Y)$ when $l\to\infty$. Thus, for any R that is achievable with $p(x)$, we have $R\leqslant H(Y)$. It implies $I_{\mathrm{G}}(X:Y,S)\leqslant H(Y)$ and thus $I_{\mathrm{G}}(\vec{X}:\vec{M},B)\leqslant m$. Conversely, for $m=0$, the no-supersignaling condition $I_{\mathrm{G}}(X:B)=0$ implies the no-signaling condition. □

Appendix C. State space of a qubit

Suppose that two independent and uniformly random bits X_0, X_1 are encoded in the state of a qubit $\hat{\rho}_{x_0x_1}$. Let $\{\hat{M}_t^m\}_{t=0,1}$ be the optimal measurement for decoding X_m $(m=0,1)$, where the mutual information $I_{\mathrm{C}}(X_m:T)$ between X_m and the measurement outcome T is maximized when the measurement m is carried out. We assume that for all x_0 and x_1,

$$P(t=x_0|m=0,x_0,x_1)=\mathrm{tr}\left[\hat{M}_{x_0}^0\hat{\rho}_{x_0x_1}\right]=\frac{1+\alpha}{2}\quad(0\leqslant\alpha\leqslant 1), \tag{C.1}$$

$$P(t = x_1|m = 1, x_0, x_1) = \mathrm{tr}\left[\hat{M}^1_{x_1}\hat{\rho}_{x_0x_1}\right] = \frac{1+\beta}{2} \quad (0 \leqslant \beta \leqslant 1). \tag{C.2}$$

In what follows, we prove that such a set of density operators $\{\hat{\rho}_{x_0x_1}\}_{x_0,x_1=0,1}$ and POVM operators $\{\hat{M}^m_t\}_{m,t=0,1}$ exists if and only if $\alpha^2+\beta^2 \leqslant 1$. Considering the parameterization of a qubit state using the Bloch sphere, the 'if' part is obviously verified. The 'only if' part is proved as follows. Let $\boldsymbol{r}_{x_0x_1}$ be the Bloch vector representation of $\hat{\rho}_{x_0x_1}$ and $\boldsymbol{u}$, $\boldsymbol{v}$ be those of $\hat{M}^0_0$ and $\hat{M}^1_0$, respectively. Formally, we have

$$\hat{\rho}_{x_0x_1} = \frac{1}{2}(I + \boldsymbol{r}_{x_0x_1}\cdot\hat{\boldsymbol{\sigma}}) \quad (\|\boldsymbol{r}_{x_0x_1}\| \leqslant 1), \tag{C.3}$$

$$\hat{M}^0_t = \frac{1}{2}(I + (-1)^t\boldsymbol{u}\cdot\hat{\boldsymbol{\sigma}}), \tag{C.4}$$

and

$$\hat{M}^1_t = \frac{1}{2}(I + (-1)^t\boldsymbol{v}\cdot\hat{\boldsymbol{\sigma}}), \tag{C.5}$$

where $\hat{\boldsymbol{\sigma}} = (\hat{\sigma}_x, \hat{\sigma}_y, \hat{\sigma}_z)$. The optimality of the measurement implies that $\|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$. From the conditions (C.1) and (C.2), we obtain that

$$\begin{aligned} \boldsymbol{u}\cdot\boldsymbol{r}_{00} = \boldsymbol{u}\cdot\boldsymbol{r}_{01} = -\boldsymbol{u}\cdot\boldsymbol{r}_{10} = -\boldsymbol{u}\cdot\boldsymbol{r}_{11} = \alpha,\\ \boldsymbol{v}\cdot\boldsymbol{r}_{00} = -\boldsymbol{v}\cdot\boldsymbol{r}_{01} = \boldsymbol{v}\cdot\boldsymbol{r}_{10} = -\boldsymbol{v}\cdot\boldsymbol{r}_{11} = \beta. \end{aligned} \tag{C.6}$$

Let $\bar{\boldsymbol{r}}_{x_0x_1}$ be the projection vectors of $\boldsymbol{r}_{x_0x_1}$ onto the two-dimensional subspace spanned by $\boldsymbol{u}$ and $\boldsymbol{v}$. Then we have

$$\bar{\boldsymbol{r}}_{00} + \bar{\boldsymbol{r}}_{11} = \bar{\boldsymbol{r}}_{01} + \bar{\boldsymbol{r}}_{10} = \boldsymbol{0} \tag{C.7}$$

and

$$\boldsymbol{u}\cdot(\bar{\boldsymbol{r}}_{00} - \bar{\boldsymbol{r}}_{01}) = \boldsymbol{v}\cdot(\bar{\boldsymbol{r}}_{00} - \bar{\boldsymbol{r}}_{10}) = 0. \tag{C.8}$$

Due to the optimality of the decoding measurements, we also have $\boldsymbol{u} \parallel (\bar{\boldsymbol{r}}_{00} + \bar{\boldsymbol{r}}_{01})$ and $\boldsymbol{v} \parallel (\bar{\boldsymbol{r}}_{00} + \bar{\boldsymbol{r}}_{10})$. Thus we obtain $\boldsymbol{u}\cdot\boldsymbol{v} = 0$. Hence

$$\alpha^2 + \beta^2 = (\boldsymbol{u}\cdot\bar{\boldsymbol{r}}_{x_0x_1})^2 + (\boldsymbol{v}\cdot\bar{\boldsymbol{r}}_{x_0x_1})^2 \leqslant \|\boldsymbol{r}_{x_0x_1}\|^2 \leqslant 1. \tag{C.9}$$

Appendix D. Inclusion relation of the sets of no-signaling correlations

Inclusion relations of the sets of bipartite and multipartite no-signaling correlations are given in (D.1).

$$\mathcal{NS} = \underset{(a)}{\mathcal{NSS}} \supset \underset{(b)}{\mathcal{IC}} \supseteq \underset{(c)}{\mathcal{CR}} \supseteq \underset{(d)}{\mathcal{Q}} \supset \underset{(e)}{\mathcal{C}} \tag{D.1}$$

$\mathcal{NS}$ is the set of all no-signaling correlations. $\mathcal{NSS}$ is the set of all no-signaling correlations that satisfies the no-supersignaling condition. By 'satisfy' we mean that for any communication protocol using that correlation, the condition is never violated. Similarly, $\mathcal{IC}$ and $\mathcal{CR}$ are the sets of all no-signaling correlations that satisfy information causality and the chain rule, respectively. $\mathcal{Q}$ and $\mathcal{C}$ are the sets of quantum and classical correlations, respectively. $\supset$ represents the strict inclusion relation, and $\supseteq$ indicates that we do not know whether the sets are equivalent or strictly included. (a) is proved in section 6. (b) is proved in [4]. (c) follows from the discussion in section 3. (d) is obvious and (e) is proved in [1]. Recently, it was proved from the observation of tripartite nonlocal correlations that at least one of (c) and (d) is a strict inclusion [21, 22].

References

[1] Bell J S 1964 *Physics* **1** 195
[2] Clauser J F, Horne M A, Shimony A and Holt R A 1969 *Phys. Rev. Lett.* **23** 880
[3] Popescu S and Rohrlich D 1994 *Found. Phys.* **24** 379
[4] Pawłowski M, Paterek T, Kaszlikowski D, Scarani V, Winter A and Żukowski M 2009 *Nature* **461** 1101
[5] Brassard G, Buhrman H, Linden N, Méthot A A, Tapp A and Unger F 2006 *Phys. Rev. Lett.* **96** 250401
[6] Linden N, Popescu S, Short A J and Winter A 2007 *Phys. Rev. Lett.* **99** 180502
[7] Brunner N and Skrzypczyk P 2009 *Phys. Rev. Lett.* **102** 160403
[8] Tsirel'son B S 1980 *Lett. Math. Phys.* **4** 93
[9] Beigi S and Gohari A 2011 arXiv:1111.3151
[10] Schumacher B 1995 *Phys. Rev.* A **51** 2738
[11] Hausladen P, Jozsa R, Schumacher B, Westmoreland M and Wooters W K 1996 *Phys. Rev.* A **54** 1869
[12] Schumacher B and Westmoreland M 1997 *Phys. Rev.* A **56** 131
[13] Cover T M and Thomas J A 2006 *Elements of Information Theory* 2nd edn (Hoboken, NJ: Wiley-Interscience) pp 199–210
[14] Holevo A S 1998 *IEEE Trans. Inform. Theory* **44** 269
[15] Barrett J 2007 *Phys. Rev.* A **75** 032304
[16] Barnum H, Barrett J, Clark L O, Leifer M, Spekkens R, Stepanik N, Wilce A and Wilke R 2010 *New J. Phys.* **12** 033024
[17] Short A J and Wehner S 2010 *New J. Phys.* **12** 033023
[18] Dahlsten O C O, Lercher D and Renner R 2012 *New J. Phys.* **14** 063024
[19] Al-Safi S W and Short A J 2011 *Phys. Rev.* A **84** 042323
[20] Barnum H, Barrett J, Leifer M and Wilce A 2007 *Phys. Rev. Lett.* **99** 240501
[21] Gallego R, Würflinger L E, Acín A and Navascués M 2011 *Phys. Rev. Lett.* **107** 210403
[22] Yang T H, Cavalcanti D, Almeida M L, Teo C and Scarani V 2012 *New J. Phys.* **14** 013061