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

Eyuri Wakakuwa¹ and Mio Murao¹,²
¹Department of Physics, Graduate School of Science, The University of Tokyo, Tokyo 113-0033, Japan
²Institute for Nano Quantum Information Electronics, The University of Tokyo, Tokyo 113-0033, Japan
E-mail: wakakuwa@eve.phys.s.u-tokyo.ac.jp

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-supersignalling 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-signalling condition. As a result, we show that Tsirelson’s bound is implied by the nonpositivity of the quantitative difference between information causality and no-supersignalling.
1. Introduction

One of the most counterintuitive phenomena that quantum mechanics predicts is nonlocality. The statistics of the outcomes of measurements performed 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-signalling 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 (CHSH) inequality [2] is limited by Tsirelson’s bound [9]. In a seminal paper [3], Popescu and Rohrlich showed that Tsirelson’s bound is strictly lower than the limit imposed by the no-signalling 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 [6, 7, 8].

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-signalling 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 no more than zero bits of classical information gain for the receiver. On the other hand, the no-signalling condition does not eliminate the possibility that $m > 0$ bits of classical communication produces more than $m$ bits of classical information gain for the receiver. Whether such an implausible situation can occur would be depending 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 foundations of quantum mechanics.

In [4], it is proved that information causality is never violated in any no-signalling theory in which we can define a 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
The chain rule implies Tsirelson’s bound

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 [19, 20], a generalized entropy $H$ is defined, and then a 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 [21, 22]. 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 [20], a coding theorem analogous to Schumacher’s quantum coding theorem [12] is investigated using generalized entropy. However, their consideration is only applicable under several restrictions. As discussed in [19], we need to seek generalizations based on the analysis of data compression or channel capacity. Such an approach is also studied in [11].

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-supersignalling 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-signalling condition, and thus it is different from information causality. This result is similar to the result obtained in [20], 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-supersignalling. Our results indicate that the chain rule of the GMI imposes a strong restriction on the underlying physical theory. As an example
The chain rule implies Tsirelson’s bound

d of this fact, we show that we can derive a bound on the state space of one 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 generalized mutual information, 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-supersignalling condition, and prove that the condition is equivalent to the no-signalling condition. In Section 7, we clarify the relation among no-supersignalling, information causality and Tsirelson’s bound. In Section 8, we show that we can limit the state space of one 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 [20, 23].

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 $\phi_1$ with probability $p$ and $\phi_2$ 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 $e_r : \mathcal{S} \rightarrow [0,1]$, such that $e_r(\phi)$ is the probability of obtaining outcome $r$ when a measurement $e$ is performed on a system in the state $\phi$. 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 $\mathcal{E} : \mathcal{S}_S \rightarrow \mathcal{S}_{S'}$, 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 $\mathcal{E}_M : \mathcal{S}_S \rightarrow \mathcal{S}_{T_S}$, 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 performed on the system $A$ and $B$, respectively. Such a measurement is called a product measurement and is included.
The chain rule implies Tsirelson’s bound

in $\mathcal{M}_{AB}$. We assume that a global state $\psi \in S_{AB}$ determines a 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 exists 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 a $m$ bit one way classical communication from Alice to Bob (see Figure 1). To accomplish this task, Alice first performs a measurement on her part of the resource (denoted by $A$ in the figure), depending on $\vec{X}$. She then constructs a $m$ bit message $\vec{M}$ from $\vec{X}$ and the measurement outcome, and sends it to Bob. Bob, after receiving $\vec{M}$, performs 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

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

(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,

$$J \leq m$$

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

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

- **Symmetry**: $I(A : B) = I(B : A)$ for any systems $A$ and $B$.
- **Nonnegativity**: $I(A : B) \geq 0$ for any systems $A$ and $B$. 
The chain rule implies Tsirelson’s bound

\[ \tilde{X} = x_1, \ldots, x_n \]

\[ \text{Alice} \quad \text{no-signalling resource} \quad \text{Bob} \]

\[ k \in \{1, \ldots, n\} \]

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.

- **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, \( I(A : B) \geq I(A : B') \).

- **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 \( \phi_x \) 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 \( \phi_x \) 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 a 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 \( \phi_x \) 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.
The chain rule implies Tsirelson’s bound

\[
\begin{array}{c}
X \\
x \\
\text{ } \\
\phi_x \\
S
\end{array}
\]

**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.

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 \(C\) is defined as the list of the codewords for all messages by

\[
C := \begin{bmatrix}
    x_1(1) & \cdots & x_l(1) \\
    \vdots & \ddots & \vdots \\
    x_1(N) & \cdots & x_l(N)
\end{bmatrix}
\]  
(3)

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

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

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

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

By performing a decoding measurement on the output systems \(S_1, \ldots, S_l\), Bob tries to guess what the original message \(w\) is. Let \(D\) denote the decoding measurement. Note that, in general, the decoding measurement is not one in which Bob performs 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

\[
P_e := \frac{1}{N} \sum_{u=1}^{N} \Pr(\hat{W} \neq u|W = u).
\]  
(6)

The pair of the codebook \(C\) and the decoding measurement \(D\) 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 is 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 \((C^{(l)}, D^{(l)})\) such that

(i) \(P^{(l)}_e \to 0\) 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 \(I_G(X : S)\), is the function which satisfies the condition that
The chain rule implies Tsirelson’s bound

(i) A rate $R$ is achievable with $p(x)$ if $R < I_G(X : S)$,
(ii) A rate $R$ is achievable with $p(x)$ only if $R \leq I_G(X : S)$.

We also define $I_G(S : X)$ by $I_G(S : X) := I_G(X : S)$.

**Theorem 4.3** $I_G(X : S)$ exists and satisfies $I_G(X : S) \leq H(X)$. Here, $H(X)$ is the Shannon entropy of the system $X$ defined by $H(X) := -\sum_{x \in X} p(x) \log p(x)$.

**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

$$H'(W|\hat{W}) \leq P_e^{(l)} lR + 1$$

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

$$lR = H'(W) = I'(W : \hat{W}) + H'(W|\hat{W})$$

$$\leq I'(X^l : \hat{W}) + P_e^{(l)} lR + 1$$

$$\leq H'(X^l) + P_e^{(l)} lR + 1$$

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

$$H'(X^l) \leq \sum_{k=1}^{l} H'(X_k) = lH'(X|K) \leq lH'(X),$$

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

$$P_e^{(l)} \geq 1 - \frac{H'(X)}{R} - \frac{1}{IR}.\tag{10}$$

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

Next we prove that any rate $R < R^*$ is also achievable with $p(x)$. Let $\{(C^{(l)}; D^{(l)})\}_l$ be a sequence of $(2^{lR}, l)$ codes that satisfies $P_e^{(l)} \to 0$ and $\tau^{(l)} \to 0$. For arbitrary $0 \leq \lambda < 1$, define another codebook $C^{(l)}$ by using $C^{(\lambda l)}$ for the first $\lambda l$ 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 $D^{(l)}$ as the measurement in which the output system $S_1 \cdots S_M$ is subjected to the decoding measurement $D^{(\lambda l)}$ and the output systems $S_{M+1}, \ldots, S_l$ are ignored. The code sequence $\{(C^{(l)}; D^{(l)})\}_l$ constructed in this way is a sequence of $(2^{lAR^*}, l)$ codes that satisfies $P_e^{(l)} \to 0$ and $\tau^{(l)} \to 0$. Thus $R = \lambda R^*$ is achievable with $p(x)$. Hence we obtain $R^* = I_G(X : S)$. \qed
The chain rule implies Tsirelson's bound

Note that \( I_G(X : S) \) 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 \( I_G(X : S) \). Since \( R = 0 \) is always achievable, \( I_G(X : S) \) is nonnegative. Shannon’s noisy channel coding theorem guarantees that \( I_G(X : S) \) coincides with the classical mutual information \( I_C(X : S) \) if \( S \) is a classical system \([15]\). The generalized mutual information satisfies the data processing inequality as follows.

**Property 4.4** Let \( \mathcal{E}_{S \rightarrow S'} \) be any local transformation that maps states of a general probabilistic system \( S \) into states of another general probabilistic system \( S' \). If \( \mathcal{E}_{S \rightarrow S'} \) contains no post-selection, the generalized mutual information does not increase under this transformation, i.e., \( I_G(X : S) \geq I_G(X : S') \). Similarly, \( I_G(X : S) \geq I_G(X' : S) \) under any local transformation \( \mathcal{E}_{X \rightarrow X'} \) 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, channel I and channel II (see Figure 3). Depending on the input \( X = x \), channel I emits the system \( S \) in the state \( \phi_x \), and channel II emits the system \( S' \) in the state \( \phi'_x = \mathcal{E}_{S \rightarrow S'}(\phi_x) \). 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 \( \{(C^{(l)}, D^{(l)})\}_l \) be a sequence of \( (2^{lR}, l) \) codes for channel II with the average error probability \( P_{e}^{(l)} \) and the tolerance \( \tau^{(l)} \). From the code \( (C^{(l)}, D^{(l)}) \), construct a \( (2^{lR}, l) \) code \( (C^{(l)}, D^{(l)}) \) for channel I by \( C^{(l)} = C^{(l)} \) and \( D^{(l)} = D^{(l)} \circ \mathcal{E}_{S \rightarrow S'}^{(l)} \). Here, \( D^{(l)} \circ \mathcal{E}_{S \rightarrow S'}^{(l)} \) represents a process in which first \( \mathcal{E}_{S \rightarrow S'} \) is applied to each of \( S_1, \ldots, S_l \) individually and then the decoding measurement \( D^{(l)} \) is performed on the total output system \( S'_1 \cdots S'_l \). The average error probability and the tolerance of this code are given by \( P_{e}^{(l)} = P_{e}^{(l)} \) and \( \tau^{(l)} = \tau^{(l)}, \) respectively. Hence, if \( P_{e}^{(l)} \rightarrow 0 \) and \( \tau^{(l)} \rightarrow 0 \), we also have \( P_{e}^{(l)} \rightarrow 0 \) and \( \tau^{(l)} \rightarrow 0 \), 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 \( \mathcal{E}_M \) 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 4.4, we have \( I_G(X : T_S) \leq I_G(X : S) \) under \( \mathcal{E}_M \), which is a generalization of Holevo’s inequality. Let us define the accessible information \( I_{acc}(X : S) \) by

\[
I_{acc}(X : S) := \max I_G(X : T_S),
\]

where the maximization is taken over all possible measurements on \( S \). Then we have

\[
0 \leq I_{acc}(X : S) \leq I_G(X : S).
\]

To summarize, the generalized mutual information satisfies the following properties.

- **Symmetry:** \( I_G(X : S) = I_G(S : X) \).
- **Nonnegativity:** \( I_G(X : S) \geq 0 \)
- **Consistency:** When \( S \) is a classical system, \( I_G(X : S) = I_C(X : S) \).
The chain rule implies Tsirelson’s bound

\[ x \xrightarrow{} \text{Channel I} \xrightarrow{} \phi_x \xrightarrow{} \mathcal{E}_{S\rightarrow S'} \xrightarrow{} \phi'_x \]

**Figure 3.** Channel II defined as the combination of channel I and \( \mathcal{E}_{S\rightarrow S'} \).

- **Data Processing Inequality:** \( I_G(X : S) \geq I_G(X' : S') \) under local stochastic maps \( \mathcal{E}_{X\rightarrow X'} \) and \( \mathcal{E}_{S\rightarrow S'} \) that contain no post-selection.

Thus, from Theorem 3.1 and Theorem 3.2, we conclude that the chain rule of the generalized mutual information should be violated in any supernonlocal theory. Conversely, the chain rule implies Tsirelson’s bound.

Throughout the rest of this paper, we use the generalized mutual information (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

\[ I_Q(X : S)_{\hat{\rho}} := H(S)_{\hat{\rho}} - \sum_{x \in \mathcal{X}} p(x) H(S)_{\hat{\rho}_x}, \quad (12) \]

where

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

\[ \tilde{\rho} = \sum_{x \in \mathcal{X}} p(x) \hat{\rho}_x, \quad (14) \]

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.,

\[ I_G(X : S)_{\Gamma_{\hat{\rho}}} = I_Q(X : S)_{\hat{\rho}} \quad (15) \]

where

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

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 \( R < I_Q(X : S)_{\hat{\rho}} \),
The chain rule implies Tsirelson’s bound

(ii) A rate $R$ is achievable with $p(x)$ only if $R \leq I_Q(X : S)_{\tilde{\rho}}$.

The first statement is proved in [13, 14] 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 $W = 1, \cdots, 2^{lR}$ is uniformly distributed. Similarly to (8), we have

$$lR = H'(W) = I'(W : \tilde{W}) + H'(W|\tilde{W}) \leq I'_Q(X^l : S^l) + P_e^{(l)} lR + 1. \quad (17)$$

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

$$I'_Q(X^l : S^l) = H'(S^l) - H'(S^l|X^l) = H'(S^l) - \sum_{k=1}^l H'(S_k|X_k)$$

$$\leq \sum_{k=1}^l (H'(S_k) - H'(S_k|X_k)) = \sum_{k=1}^l I'_Q(K : S_k)$$

$$= \sum_{k=1}^l I'_Q(X : S) = \sum_{k=1}^l I'_Q(X, K : S) - \sum_{k=1}^l I'_Q(K : S)$$

$$\leq \sum_{k=1}^l I'_Q(X, K : S) = \sum_{k=1}^l I'_Q(X : S). \quad (18)$$

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

$$P_e^{(l)} \geq 1 - \frac{I'_Q(X : S)}{R} - \frac{1}{lR}. \quad (19)$$

If $R$ is achievable with $p(x)$, there exists a sequence of $(2^{lR}, l)$ codes satisfying $P_e^{(l)} \to 0$ and $I'_Q(X : S) \to I_Q(X : S)_{\tilde{\rho}}$ when $l \to \infty$. Thus $R \leq I_Q(X : S)_{\tilde{\rho}}$. \qed

6. No-supersignalling 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-supersignalling condition by using the GMI. Suppose that Alice is trying to send to distant Bob information about $n$ independent classical bits $X_1, \cdots, X_n$, under the condition that they can only use a $m$ bit classical communication $\tilde{M}$ 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 $I_G(\tilde{X} : \tilde{M}, B)$. We say that the no-supersignalling condition is satisfied if

$$I_G(\tilde{X} : \tilde{M}, B) \leq m \quad (20)$$

holds for all $m \geq 0$. 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-supersignalling condition is equivalent to the no-signalling condition. It indicates that information causality and no-supersignalling are different.
The chain rule implies Tsirelson’s bound

\[ \tilde{X} = x_1, \ldots, x_n \]

\[ Alice \quad \rightarrow \quad Bob \]

\[ A \quad \text{no-signalling resource} \quad B \]

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

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

**Proof.** Consider a channel with an input system \( X \) and two output systems \( S \) and \( Y \) (see Figure 5). Let \( Z \) 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 \( z \in Z \) is performed on the system \( S \) in the state \( \phi_{xy} \). To achieve \( I_{\text{acc}}(X : S, Y) \), the receiver performs 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

\[ p_1(t|x, y) := p(t|x, y, z(y)) . \]  

(21)

We define

\[ p_1(t, x, y) := p(x, y)p_1(t|x, y) = p(x, y)p(t|x, y, z(y)) . \]  

(22)

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

\[ \sum_y p(x, y)p(t|x, y, z) = p(x)p_2(t|z) , \]  

(23)

where

\[ p_2(t|z) := \sum_{x,y} p(x, y)p(t|x, y, z) . \]  

(24)

Thus we obtain

\[ p_1(t, x, y) = p(x, y)p(t|x, y, z(y)) \]
\[ \leq \sum_{y'} p(x, y')p(t|x, y', z(y)) \]
\[ = p(x)p_2(t|z(y)) . \]  

(25)

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

\[ I_{\text{acc}}(X : S, Y) = I_C(X : T, Y)_{p_1} . \]
The chain rule implies Tsirelson’s bound

\[
\sum_{t,x,y} p_1(t,x,y) \log \frac{p_1(t,x,y)}{p(x)p_1(t,y)} = H(Y) + \sum_{t,x,y} p_1(t,x,y) \log \frac{p_1(t,x,y) p(y)}{p(x)p_1(t,y)} \\
\leq H(Y) + \sum_{t,x,y} p_1(t,x,y) \log \frac{p(x)p(y) p_2(t|z(y))}{p(x)p_1(t,y)} \\
= H(Y) - \sum_{t,y} p_1(t,y) \log \frac{p_1(t,y)}{p_2(t,y)} \\
= H(Y) - D(p_1(t,y)||p_2(t,y)) \\
\leq H(Y).
\]

In the first inequality, we used \eqref{eq: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 nonnegativity of the relative entropy. \( \square \)

**Theorem 6.2** The no-supersignalling condition defined in terms of the GMI \eqref{eq:20} is equivalent to the no-signalling condition.

**Proof.** Consider a \((2^R, l)\) code for the channel presented in Figure 5 and let \( X = \bar{X} \), \( Y = \bar{Y} \) and \( S = B \). Suppose that Alice’s message is uniformly distributed. By Fano’s inequality, we have

\[
I'(W : \hat{W}) \geq lR - 1 - P_e^{(l)} lR. \tag{26}
\]

By the data processing inequality, we also have

\[
I'(W : \hat{W}) \leq I'(X^l : Y^l, T_{s^l}) \leq I'_{acc}(X^l : Y^l, S^l). \tag{27}
\]

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

\[
I'_{acc}(X^l : Y^l, S^l) \leq H'(Y^l), \tag{28}
\]

and thus

\[
I'(W : \hat{W}) \leq H'(Y^l) \leq lH'(Y). \tag{29}
\]

Hence we obtain

\[
(1 - P_e^{(l)}) R \leq H'(Y) + \frac{1}{l}. \tag{30}
\]

If \( R \) is achievable with \( p(x) \), there exists a sequence of \((2^l, l)\) codes that satisfies \( P_e^{(l)} \to 0 \) and \( H'(Y) \to H(Y) \) when \( l \to \infty \). Thus, for any \( R \) that is achievable with \( p(x) \), we have
The chain rule implies Tsirelson’s bound

$R \leq H(Y)$. It implies $I_G(X : Y, S) \leq H(Y)$ and thus $I_G(\vec{X} : \vec{M}, B) \leq m$. Conversely, for $m = 0$, the no-supersignalling condition $I_G(X : B) = 0$ implies the no-signalling condition. □

7. The difference between no-supersignalling and information causality

In this section, we discuss the relation among information causality, no-supersignalling, Tsirelson’s bound and the chain rule. Let us define

$$
\Delta_{\text{NSS}} := I_G(\vec{X} : \vec{M}, B) - m, \tag{31}
$$

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

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

$\Delta_{\text{NSS}}$ quantifies how much the capability of classical communication is increased by the assistance of nonlocal correlations. No-supersignalling is equivalent to $\Delta_{\text{NSS}} \leq 0$, and information causality is equivalent to $\Delta_{\text{IC}} \leq 0$. $\Delta'$ quantifies the difference between no-supersignalling and information causality.

Theorem 3.2 states that, if Tsirelson’s bound is violated, we have $\Delta_{\text{IC}} > 0$. Therefore violation of Tsirelson’s bound implies at least either $\Delta_{\text{NSS}} > 0$ or $\Delta' > 0$. Then which does violation of Tsirelson’s bound imply, $\Delta_{\text{NSS}} > 0$ or $\Delta'> 0$? As we proved in Section 6, $\Delta_{\text{NSS}} \leq 0$ is satisfied by all no-signalling 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 $\Delta'$ (see Figure 6). Let us further define

$$
\Delta_{\text{CR}} := \sum_{k=1}^{n} I_G(X_k : \vec{M}, B, X_1, \cdots, X_{k-1}) - I_G(\vec{X} : \vec{M}, B). \tag{34}
$$

The chain rule is equivalent to $\Delta_{\text{CR}} = 0$. By the data processing inequality, we always have $\Delta_{\text{CR}} \geq \Delta'$. Thus the chain rule implies Tsirelson’s bound through imposing $\Delta' \leq \Delta_{\text{CR}} = 0$.

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

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

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.

Another way to show this is to observe that the data processing inequality and the no-supersignalling condition imply $\Delta_{\text{CR}} \geq \Delta_{\text{IC}}$. \[\]
The chain rule implies Tsirelson’s bound

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

8. Restriction on one 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 [18]. 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 one gbit is not more than one bit. Thus we require

$$I_G(X : S_{1gb}) \leq 1 \quad (36)$$

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

$$I_G(X_0, X_1 : S_{1gb}) \leq 1 \quad . \quad (37)$$

By the chain rule, we have

$$I_G(X_0, X_1 : S_{1gb}) = I_G(X_0 : S_{1gb}) + I_G(X_1 : S_{1gb}, X_0) \quad . \quad (38)$$

By the data processing inequality, we also have

$$I_G(X_0 : S_{1gb}) + I_G(X_1 : S_{1gb}, X_0) \geq I_{acc}(X_0 : S_{1gb}) + I_{acc}(X_1 : S_{1gb}) \quad . \quad (39)$$

Thus the chain rule implies

$$I_{acc}(X_0 : S_{1gb}) + I_{acc}(X_1 : S_{1gb}) \leq 1 \quad . \quad (40)$$

We consider success probabilities of the decoding measurements on $S_{1gb}$ for $X_0$ and $X_1$. For simplicity, we assume that the optimal measurement performed on $S_{1gb}$ 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 performed. The index $m = 0, 1$ corresponds to the optimal measurement for decoding $X_0, X_1$, respectively.
The chain rule implies Tsirelson’s bound

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 \( x_1 \),

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

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

Then we have

\[
I_{acc}(X_0 : S_{1gb}) = I_C(x_1 : t|m = 0) = 1 - H(x_0|t, m = 0)
\]

\[
= 1 - H(x_0 \oplus t|m = 0) = 1 - h \left( \frac{1 + \alpha}{2} \right), \tag{41}
\]

and

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

Here, \( h(x) \) is the binary entropy defined by \( h(x) := -x \log x - (1 - x) \log (1 - x) \). From (40), (41) and (42), we have

\[
h \left( \frac{1 + \alpha}{2} \right) + h \left( \frac{1 + \beta}{2} \right) \geq 1. \tag{43}
\]

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

9. Conclusions and discussions

We have defined a generalized mutual information (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 with the quantum mutual information if
The chain rule implies Tsirelson’s bound

the output system is quantum. The GMI satisfies nonnegativity, 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-supersignalling 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-signalling condition. We also clarified the relation among no-supersignalling, 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 the 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 Shannon theory, 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, 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 lot 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 generalized mutual information 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 generalized mutual information and compare them.

Acknowledgments

We thank Takanori Sugiyama and Salman Beigi for useful discussions. We also thank referees for useful comments. This work was supported by Project for Developing Innovation Systems of Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. MM acknowledges support from JSPS by KAKENHI (Grant No. 23540463).

§ By Shannon theory we mean the theoretical framework composed of various theorems on the asymptotic coding rate of the sources and the channels.
Appendix A. Data processing inequality

We prove the latter part of Theorem 4.4, which states that under any local stochastic map $\mathcal{E}_{X \rightarrow X'}$, that contains no post-selection, we have

$$I_G(X : S) \geq I_G(X' : S).$$  \hfill (A.1)

The effect of $\mathcal{E}_{X \rightarrow X'}$ 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 $\{p(x), \phi_x\}_{x \in X}$ be the state of $XS$ before applying $\mathcal{E}_{X \rightarrow X'}$. 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 $x \in X$ and $x' \in X'$. The state of $X'S$ after applying $\mathcal{E}_{X \rightarrow X'}$ is $\{p(x'), \phi_{x'}\}_{x' \in X'}$, where $\phi_{x'}$ is the mixture of $\phi_x$ with the probability given by $p_{\mathcal{E}}(x|x')$. We assume that $|X|, |X'| < \infty$.

To prove (A.1), consider two channels, channel I and channel III (see Figure A1). Channel I outputs the system $S$ in the state $\phi_x$ 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 $(2^R, l)$ codes $(\mathcal{C}^{(l)}, \mathcal{D}^{(l)})$ for channel III that satisfies

(i) $P_e^{(l)} \rightarrow 0$ when $l \rightarrow \infty$,

(ii) $\tau^{(l)} \rightarrow 0$ when $l \rightarrow \infty$.

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 $(2^R, l)$ codes $(\mathcal{C}^{(l)}, \mathcal{D}^{(l)})$ for channel I in the following way.

- For any $w$ and $k$ ($1 \leq w \leq 2^R, 1 \leq k \leq l$), 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 $P_e^{\mathcal{C}^{(l)}}$ be the average error probability of the code $(\mathcal{C}^{(l)}, \mathcal{D}^{(l)})$ defined by

$$P_e^{\mathcal{C}^{(l)}} := \frac{1}{2^R} \sum_{u=1}^{2^R} P(\hat{W} \neq u|W = u, \mathcal{C}^{(l)}).$$ \hfill (A.2)

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

$$\bar{P}_e^{(l)} := \sum_{\mathcal{C}^{(l)}} P(\mathcal{C}^{(l)}) P_e^{\mathcal{C}^{(l)}},$$ \hfill (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 $\bar{P}_e^{(l)} \rightarrow 0$ in the limit of $l \rightarrow \infty$. In Lemma A.2, we prove that for 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.
The chain rule implies Tsirelson’s bound

\[ x' \xrightarrow{\mathcal{E}_{X \rightarrow X'}} x \xrightarrow{\text{Channel I}} \phi_x \]

\[ x' \xrightarrow{\text{Channel III}} \phi_{x'} \]

**Figure A1.** Channel III defined as the combination of $\mathcal{E}_{X \rightarrow X'}$ and channel I. This channel as a whole is equivalent to a channel with the input $x'$ and the output $\phi_{x'}$.

**Lemma A.1**

\[
\lim_{l \to \infty} \overline{P}_e^{(l)} = 0 . \tag{A.4}
\]

**Proof.** $\overline{P}_e^{(l)}$ defined by (A.3) is calculated to

\[
\begin{align*}
\overline{P}_e^{(l)} &= \sum_{\mathcal{C}^{(l)}} P(\mathcal{C}^{(l)}) \times \frac{1}{2^{2R}} \sum_{u=1}^{2^R} P(\hat{W} \neq u|W = u, \mathcal{C}^{(l)}) \\
&= \frac{1}{2^{2R}} \sum_{u=1}^{2^R} \sum_{\mathcal{C}^{(l)}} P(\mathcal{C}^{(l)}) P(\hat{W} \neq u|W = u, \mathcal{C}^{(l)}) \\
&= \frac{1}{2^{2R}} \sum_{u=1}^{2^R} \overline{P}(\hat{W} \neq u|W = u) , \tag{A.5}
\end{align*}
\]

where

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

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

\[
P(\mathcal{C}^{(l)}) = P(\{x_k(w)\}_{w,k} = \{\xi_{wk}\}_{w,k}) = \prod_{w=1}^{2^{2R}} \prod_{k=1}^{l} P(x_k(w) = \xi_{wk}) = \prod_{w=1}^{2^{2R}} \prod_{k=1}^{l} p_{\mathcal{E}}(x = \xi_{wk}|x' = x'_k(w)) . \tag{A.7}
\]

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

\[
P(\hat{W} \neq u|W = u, \mathcal{C}^{(l)}) = P(D(\phi_{x_1(u)} \cdots \phi_{x_l(u)}) \neq u|\{x_k(w)\}_{w,k} = \{\xi_{wk}\}_{w,k}) = P(D(\phi_{\xi_{w_1}} \cdots \phi_{\xi_{wl}}) \neq u) , \tag{A.8}
\]
and we obtain
\[
\bar{P}(\hat{W} \neq u | W = u) = \sum_{\{x_k\}_k} P(D(\phi_{x_1(u)} \cdots \phi_{x_l(u)}) \neq u | \{x_k(w)\}_w, k = \{\xi_{w,k}\}_w, k) \times P(\{x_k(w)\}_w, k = \{\xi_{w,k}\}_w, k)
\]
\[
= \sum_{\{x_k\}_k} P(D(\phi_{x_1} \cdots \phi_{x_l}) \neq u) \times P(\{x_k(u)\}_k = \{\xi_{w,k}\}_w, k)
\]
\[
= \sum_{\{x_k\}_k} P(D(\phi_{x_1} \cdots \phi_{x_l}) \neq u) \times \prod_{k=1}^l p_{E}(x = \xi_{uk}|x' = x'_k(u)) . \tag{A.9}
\]
On the other hand, the error probability for the message \(w\) when channel III is used with the code \((C^{(l)}, D^{(l)})\) is given by
\[
P'(\hat{W} \neq u | W = u) = P(D(\phi_{x'_1(u)} \cdots \phi_{x'_l(u)}) \neq u)
\]
\[
= \sum_{\{x_k\}_k} P(D(\phi_{x_1} \cdots \phi_{x_l}) \neq u) \times \prod_{k=1}^l p_{E}(x = x_k|x' = x'_k(u)) . \tag{A.10}
\]
From (A.9) and (A.10), we obtain
\[
\bar{P}(\hat{W} \neq u | W = u) = P'(\hat{W} \neq u | W = u) , \tag{A.11}
\]
and consequently
\[
\bar{p}'_{e^{(l)}} = \bar{p}_{e^{(l)}} . \tag{A.12}
\]
Therefore \(\bar{p}_{e^{(l)}} \to 0\) when \(l \to \infty\).

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

\textbf{Proof.} Let \(f(x)^{(l)}\) and \(f(x')^{(l)}\) be the letter frequency of the codebook \(C^{(l)}\) and \(C'^{(l)}\), respectively. We have
\[
|f(x)^{(l)} - p(x)| = \left| f(x)^{(l)} - \sum_{x' \in \mathcal{X}'} p_E(x|x')p(x') \right|
\]
\[
\leq \left| f(x)^{(l)} - \sum_{x' \in \mathcal{X}'} f(x')^{(l)}p_E(x|x') \right| + \left| \sum_{x' \in \mathcal{X}'} f(x')^{(l)}p_E(x|x') - \sum_{x' \in \mathcal{X}'} p_E(x|x')p(x') \right|
\]
\[
\leq \left| f(x)^{(l)} - \sum_{x' \in \mathcal{X}'} f(x')^{(l)}p_E(x|x') \right| + \sum_{x' \in \mathcal{X}'} p_E(x|x') \left| f(x')^{(l)} - p(x') \right| .
\]
Define
\[
f(x, x')^{(l)} := \frac{|\{(k, w)|x_k(w) = x, x'_k(w) = x', 1 \leq k \leq l, 1 \leq w \leq 2^{lR} \}|}{l \cdot 2^{lR}}
\]
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
\[ \Delta(x)^{(l)} := \left| f(x)^{(l)} - \sum_{x' \in \mathcal{X}^l} f(x')^{(l)} p_{e}(x|x') \right| \]
\[ \leq \sum_{x' \in \mathcal{X}^l} f(x')^{(l)} \left| \frac{f(x, x')^{(l)}}{f(x')^{(l)}} - p_{e}(x|x') \right|. \] (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
\[ \sum_{x' \in \mathcal{X}^l} p_{e}(x|x') \left| f(x')^{(l)} - p(x') \right| \leq \tau^{(l)} \cdot |\mathcal{X}^l| \] (A.15)
and thus
\[ \lim_{l \to \infty} \sum_{x' \in \mathcal{X}^l} p_{e}(x|x') \left| f(x')^{(l)} - p(x') \right| = 0. \] (A.16)

Therefore we obtain
\[ \tau^{(l)} = \max_{x} \left| f(x)^{(l)} - p(x) \right| \to 0 \quad \text{in probability}. \] (A.17)

**Theorem A.3** \( R \) is achievable with \( p(x) \) by channel I.

**Proof.** Take arbitrary \( \epsilon, \delta, \eta > 0 \). From Lemma A.1 and Lemma A.2, for sufficiently large \( l \) we have
\[ \bar{P}_{e}^{(l)} < \epsilon \] (A.18)
and
\[ \Pr\{\tau^{(l)} < \delta\} > 1 - \eta. \] (A.19)

Define \( C^{(l)}_{\delta} := \{C^{(l)}|\tau^{(l)} < \delta\} \). The average error probability averaged over all codebooks in \( C^{(l)}_{\delta} \) is calculated to
\[
\frac{\sum_{C^{(l)} \in C^{(l)}_{\delta}} P(C^{(l)}) P_{e}^{C^{(l)}}}{\sum_{C^{(l)} \in C^{(l)}_{\delta}} P(C^{(l)})} = \frac{\bar{P}_{e}^{(l)} - \sum_{C^{(l)} \not\in C^{(l)}_{\delta}} P(C^{(l)}) P_{e}^{C^{(l)}}}{\sum_{C^{(l)} \in C^{(l)}_{\delta}} P(C^{(l)})} \leq \frac{\bar{P}_{e}^{(l)}}{\sum_{C^{(l)} \in C^{(l)}_{\delta}} P(C^{(l)})} < \epsilon/(1 - \eta). 
\]

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

**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
Here, \( \phi_x \) 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 \( \Phi_{x_1 \cdots x_l} \) is consistent with \( \phi_{x_1} \cdots \phi_{x_l} \) if the two states exhibit the same statistics for any product measurement. \( \Phi^{(l)} := \{ \Phi_{x_1 \cdots x_l} \} \) is said to be consistent with \( \{ \phi_{x_1} \cdots \phi_{x_l} \} \) 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 \( \Phi := \{ \Phi^{(l)} \}_{l=1}^{\infty} \) is consistent with \( \{ \phi_{x} \} \) if \( \Phi^{(l)} \) is consistent with \( \{ \phi_{x_1} \cdots \phi_{x_l} \} \) 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 \( \Phi \) if there exists a sequence of \( (2^{Rl}, l) \) codes \( (C^{(l)}, D^{(l)}) \) for \( \Gamma_{\phi}^{(l)} \in \Gamma_{\phi} \) such that

(i) \( P_{e}^{(l)} \to 0 \) 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 \( \Phi \) that is consistent with \( \{ \phi_{x} \}_{x \in \mathcal{X}} \).

We define the generalized mutual information by Definition 4.2 and its existence is proved by Theorem 4.3. The data processing inequality (Property 4.4) is proved as follows.

**Proof.** The inequality \( I_G(X : S) \geq I_G(X : S') \) under local transformation \( \mathcal{E}_{S \rightarrow S'} \) is proved as follows.

\[
I_G(X : S') = \sup\{ R | R \text{ is achievable for all } \Phi' \text{ that is consistent with } \{ \mathcal{E}(\phi_x) \}_{x \in \mathcal{X}} \} \\
\leq \sup\{ R | R \text{ is achievable for } \mathcal{E}(\Phi) \text{ for all } \Phi \text{ that is consistent with } \{ \phi_x \}_{x \in \mathcal{X}} \} \\
\leq \sup\{ R | R \text{ is achievable for all } \Phi \text{ that is consistent with } \{ \phi_x \}_{x \in \mathcal{X}} \} \\
= I_G(X : S). \tag{B.1}
\]

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 \( \Phi \) is consistent with \( \{ \phi_x \}_{x \in \mathcal{X}} \). The second inequality is proved in the same way as the proof presented in page 9.

The inequality \( I_G(X : S) \geq I_G(X' : S) \) under local transformation \( \mathcal{E}_{X \rightarrow X'} \) is proved as follows.

\[
I_G(X' : S) = \sup\{ R | R \text{ is achievable for all } \Phi' \text{ that is consistent with } \{ \phi_{x'} \}_{x' \in \mathcal{X}'} \}
\]
The chain rule implies Tsirelson's bound

\[ \leq \sup\{R|R \text{ is achievable for } \Phi_{X'} \text{ for all } \Phi \text{ that is consistent with } \{\phi_x\}_{x \in \mathcal{X}} \} \]
\[ \leq \sup\{R|R \text{ is achievable for all } \Phi \text{ that is consistent with } \{\phi_x\}_{x \in \mathcal{X}} \} \]
\[ = I_G(X : S) \quad (\text{B.2}) \]

Here, \( \Phi_{X'} := \{\Phi_{X'}^{(l)}\}_{l=1}^{\infty} \) and \( \Phi_{X'}^{(l)} := \{\Phi_{x'_1 \cdots x'_l}^{(l)}\}_{x'_1 \cdots x'_l \in \mathcal{X}^l} \), where \( \Phi_{x'_1 \cdots x'_l}^{(l)} \) is the mixture of \( \Phi_{x_1 \cdots x_l} \in \Phi^{(l)} \in \Phi \) with the probability \( \prod_{k=1}^{l} p(\varepsilon(x_k | x'_k)). \) The first inequality comes from the fact that \( \Phi_{X'} \) is consistent with \( \{\phi_{x'}\}_{x' \in \mathcal{X}'} \) if \( \Phi \) 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-supersignalling and no-signalling (Theorem 6.2) is proved as follows.

**Proof.** Due to the no-signalling condition, there exists \( \Phi \) that is consistent with \( \{\phi_{xy}\}_{x \in \mathcal{X}, y \in \mathcal{Y}} \), and satisfies \( I_{acc}(X^l : S^l) = 0 \) 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 \((2^R,l)\) code for the channel. In the same way as the proof of Theorem 6.2, we have \((1 - P_e^{(l)})R \leq H'(Y) + 1/l. \) If \( R \) is achievable with \( p(x) \) for \( \Phi \), there exists a sequence of \((2^R,l)\) code for \( \Gamma_{\Phi}^{(l)} \) that satisfies \( P_e^{(l)} \to 0 \) and \( H'(Y) \to H(Y) \) when \( l \to \infty. \) Thus, for any \( R \) that is achievable with \( p(x) \), we have \( R \leq H(Y). \) It implies \( I_G(X : Y, S) \leq H(Y) \) and thus \( I_G(X : \tilde{M}, B) \leq m. \) Conversely, for \( m = 0 \), the no-supersignalling condition \( I_G(X : B) = 0 \) implies the no-signalling condition.

**Appendix C. State space of a qubit**

Suppose that two independent and uniformly random bits \( X_0, X_1 \) are encoded into the state of a qubit \( \hat{\rho}_{x_0x_1}. \) Let \( \{\hat{M}_i^m\}_{i=0,1} \) be the optimal measurement for decoding \( X_m \) \( (m = 0, 1) \), where the mutual information \( I_G(X_m : T) \) between \( X_m \) and the measurement outcome \( T \) is maximized when the measurement \( m \) is performed. We assume that for all \( x_0 \) and \( x_1, \)
\[
P(t = x_0 | m = 0, x_0, x_1) = \text{tr}[\hat{M}_0^{x_0} \hat{\rho}_{x_0x_1}] = \frac{1 + \alpha}{2} \quad (0 \leq \alpha \leq 1), \tag{C.1} \]
\[
P(t = x_1 | m = 1, x_0, x_1) = \text{tr}[\hat{M}_1^{x_1} \hat{\rho}_{x_0x_1}] = \frac{1 + \beta}{2} \quad (0 \leq \beta \leq 1). \tag{C.2} \]

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 \( \{\hat{M}_i^m\}_{m,i=0,1} \) exists if and only if \( \alpha^2 + \beta^2 \leq 1. \) Considering the parametrization of a qubit state using the Bloch sphere, the “if” part is obviously verified. The “only if” part is proved as follows. Let \( \mathbf{r}_{x_0x_1} \) be the Bloch vector representation of \( \hat{\rho}_{x_0x_1} \) and \( \mathbf{u}, \mathbf{v} \) be those of \( \hat{M}_0^0 \) and \( \hat{M}_0^1 \), respectively. Formally, we have
\[
\hat{\rho}_{x_0x_1} = \frac{1}{2}(I + \mathbf{r}_{x_0x_1} \cdot \hat{\sigma}) \quad (||\mathbf{r}_{x_0x_1}|| \leq 1), \tag{C.3} \]
The chain rule implies Tsirelson’s bound

\[ \hat{M}_t^0 = \frac{1}{2} (I + (-1)^t \mathbf{u} \cdot \hat{\sigma}), \]  
\[ \hat{M}_t^1 = \frac{1}{2} (I + (-1)^t \mathbf{v} \cdot \hat{\sigma}), \]

where \( \hat{\sigma} = (\hat{\sigma}_x, \hat{\sigma}_y, \hat{\sigma}_z) \). The optimality of the measurement implies that \( \|\mathbf{u}\| = \|\mathbf{v}\| = 1 \).

From the condition (C.1) and (C.2), we obtain

\[ \mathbf{u} \cdot \mathbf{r}_{00} = \mathbf{u} \cdot \mathbf{r}_{01} = -\mathbf{u} \cdot \mathbf{r}_{10} = -\mathbf{u} \cdot \mathbf{r}_{11} = \alpha, \]
\[ \mathbf{v} \cdot \mathbf{r}_{00} = -\mathbf{v} \cdot \mathbf{r}_{01} = \mathbf{v} \cdot \mathbf{r}_{10} = -\mathbf{v} \cdot \mathbf{r}_{11} = \beta. \]

Let \( \tilde{\mathbf{r}}_{x_0x_1} \) be the projection vectors of \( \mathbf{r}_{x_0x_1} \) onto the two dimensional subspace spanned by \( \mathbf{u} \) and \( \mathbf{v} \). Then we have

\[ \tilde{\mathbf{r}}_{00} + \tilde{\mathbf{r}}_{11} = \tilde{\mathbf{r}}_{01} + \tilde{\mathbf{r}}_{10} = 0. \]

and

\[ \mathbf{u} \cdot (\tilde{\mathbf{r}}_{00} - \tilde{\mathbf{r}}_{01}) = \mathbf{v} \cdot (\tilde{\mathbf{r}}_{00} - \tilde{\mathbf{r}}_{10}) = 0. \]

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

\[ \alpha^2 + \beta^2 = (\mathbf{u} \cdot \mathbf{\tilde{r}}_{x_0x_1})^2 + (\mathbf{v} \cdot \mathbf{\tilde{r}}_{x_0x_1})^2 \leq \|\mathbf{r}_{x_0x_1}\|^2 \leq 1. \]

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

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

\[ \mathcal{NS} = \mathcal{NSS} \supseteq \mathcal{IC} \supseteq \mathcal{CR} \supseteq \mathcal{Q} \supseteq \mathcal{C} \]  
\[ (a) \quad (b) \quad (c) \quad (d) \quad (e) \]

\( \mathcal{NS} \) is the set of all no-signalling correlations. \( \mathcal{NSS} \) is the set of all no-signalling correlations that satisfies the no-supersignalling 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-signalling correlations that satisfy information causality and the chain rule, respectively. \( \mathcal{Q} \) and \( \mathcal{C} \) are the sets of quantum and classical correlations, respectively. \( \supseteq \) 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 8. (d) is obvious and (e) is proved in [1]. Recently it is proved from the observation of tripartite nonlocal correlations that at least one of (c) and (d) is a strict inclusion [24, 25].
The chain rule implies Tsirelson’s bound

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] Pawłowski M and Scarani V 2011 arXiv:1112.1142
[6] Brassard G, Buhrman H, Linden N, Méthot A A, Tapp A and Unger F 2006 Phys. Rev. Lett. 96 250401
[7] Linden N, Popescu S, Short A J and Winter A 2007 Phys. Rev. Lett. 99 180502
[8] Brunner N and Skrzypczyk P 2009 Phys. Rev. Lett. 102 160403
[9] Tsirel’son B S 1980, Lett. Math. Phys. 4 93
[10] Ambainis A, Nayak A, Ta-Shma A and Vazirani U 2002 J. ACM 49 496
[11] Beigi S and Gohari A 2011 arXiv:1111.3151
[12] Schumacher B 1995 Phys. Rev. A 51 2738
[13] Hausladen P, Jozsa R, Schumacher B, Westmoreland M and Wooters W K 1996 Phys. Rev. A 54 1869
[14] Schumacher B and Westmoreland M 1997 Phys. Rev. A 56 131
[15] Cover T M and Thomas J A 2006 Elements of Information Theory 2nd edn (Hoboken: Wiley-Interscience) pp 199-210
[16] Holevo A S 1998 IEEE Trans. Inf. Theory 44 269
[17] Ver Steeg G and Wehner S 2009 arXiv:0811.3771
[18] Barrett J 2007 Phys. Rev. A 75 032304
[19] Barnum H, Barrett J, Clark L O, Leifer M, Spekkens R, Stepanik N, Wilce A and Wilke R 2010 New J. Phys. 12 033024
[20] Short A J and Wehner S 2010 New J. Phys. 12 033023
[21] Dahlsten O C O, Lercher D and Renner R 2012 New J. Phys. 14 063024
[22] Al-Safi S W and Short A J 2011 Phys. Rev. A 84 042323
[23] Barnum H, Barrett J, Leifer M and Wilce A 2007 Phys. Rev. Lett. 99 240501
[24] Gallego R, Würflinger L E, Acín A and Navascués M 2011 Phys. Rev. Lett. 107 210403
[25] Yang T H, Cavalcanti D, Almeida M L, Teo C and Scarani V 2012 New J. Phys. 14 013061