# Operational capacity and pseudoclassicality of a quantum channel

Akio Fujiwara* and Hiroshi Nagaoka†

**Abstract**

We explore some basic properties of coding theory of a general quantum communication channel and its *operational* capacity, including (1) adaptive measurement with feedback code, (2) reconsideration of single-letterized capacity formula, and (3) pseudoclassicality of a channel.

Index terms
quantum information theory,
coding theorem,
operational quantum capacity,
pseudoclassical channel,
quantum binary channel

*Department of Mathematics, Osaka University, 1-16 Machikane-yama, Toyonaka, Osaka 560, Japan. E-mail: fujiwara@math.wani.osaka-u.ac.jp

†Graduate School of Information Systems, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo 182, Japan. E-mail: nagaoka@is.uec.ac.jp

# 1 Introduction

In order to consider a communication system which is described by quantum mechanics, we must reformulate information (communication) theory in terms of quantum mechanical language. Suppose an input state $\rho$ is transmitted through a quantum channel to yield the output state $\rho'$. The problem is, of course, to measure how much information is transmitted to the receiver observing the output state. In classical theory, it is measured by the mutual information. Shannon's fundamental result [1] asserts that the supremum of mutual information over input distributions happens to be identical to the (operational) channel capacity, i.e. the maximum rate below which one can transmit information within an arbitrary small error probability. One of the principal themes of the traditional information theory thus consists in establishing coding theorems in various contexts, through which many informational contents are equipped with operational meaning in certain asymptotic frameworks. A legitimate argument of quantum channel coding theory may be summarized as follows.

Let $\mathcal{H}$ be a separable Hilbert space which corresponds to the physical system of interest. A *quantum state*, the quantum counterpart of a probability measure, is represented by a density operator $\rho$ on $\mathcal{H}$ which satisfies $\rho = \rho^* \geq 0$ and $\mathrm{Tr}\,\rho = 1$. A *measurement* $\{M(B)\}_{B\in\mathcal{F}}$ on a measurable space $(\mathcal{X}, \mathcal{F})$ is an operator-valued set function which satisfy the axioms [2]:

(1) $M(B) = M(B)^* \geq 0$ for all $B \in \mathcal{F}$, with $M(\phi) = 0$, $M(\mathcal{X}) = I$,

(2) For all at most countable disjoint sequence $B_j$ in $\mathcal{F}$, $M(\bigcup_j B_j) = \sum_j M(B_j)$ holds, where the series is weakly convergent.

When a measurement $M$ is applied to a quantum state $\rho$, the probability of finding the outcome in a measurable set $B$ is $\mathrm{Tr}\,\rho M(B)$. For mathematical simplicity, we restrict ourselves to finite dimensional Hilbert spaces and to measurements which take values on finite sets in this paper. In this case, $\mathcal{F} = 2^{\mathcal{X}}$ and a measurement is described by a set of nonnegative Hermitian operators $\{M(x)\ ;\ x \in \mathcal{X}\}$ satisfying $\sum_{x\in\mathcal{X}} M(x) = I$. Further, when a measurement $M$ is applied to a state $\rho$, the outcome of the measurement forms an $\mathcal{X}$-valued random variable which obeys the probability distribution $p(x) = \mathrm{Tr}\,\rho M(x)$.

Letting $\mathcal{S}(\mathcal{H}_j)$ be the set of states on $\mathcal{H}_j$, a *quantum channel* between an input system $\mathcal{H}_1$ and an output system $\mathcal{H}_2$ is described by an affine map $\Gamma : \mathcal{S}_1 \to \mathcal{S}(\mathcal{H}_2)$ satisfying $\Gamma(\lambda\rho_1 + (1-\lambda)\rho_2) = \lambda\Gamma(\rho_1) + (1-\lambda)\Gamma(\rho_2)$ for all $\rho_1, \rho_2 \in \mathcal{S}_1$ and $0 \leq \lambda \leq 1$, where $\mathcal{S}_1$ is a certain compact convex subset of $\mathcal{S}(\mathcal{H}_1)$. Although a more restrictive (physical) definition of a channel is often adopted as the dual map of a certain completely positive map [3][4], only the affinity assumption on a convex subset $\mathcal{S}_1 \subset \mathcal{S}(\mathcal{H}_1)$ is sufficient in this paper. It should also be noted that all the analysis about the quantum capacity here can be worked out entirely on the image of $\Gamma$, i.e. $\mathcal{S}_2 \stackrel{\mathrm{def}}{=} \Gamma\mathcal{S}_1 \subset \mathcal{S}(\mathcal{H}_2)$. So in a purely mathematical viewpoint, it is sufficient to consider only for $\Gamma = \mathrm{id}(: \mathcal{S}_2 \to \mathcal{S}_2)$. However we abstain from doing so on the grounds that the use of general $\Gamma$ is contextual with the actual communication system, which may help clarify the similarities and differences between the classical channels and the quantum channels.

To enter on an asymptotic framework, we consider the $n$th extension of the system $\mathcal{H}^{\otimes n} = \mathcal{H}\otimes\cdots\otimes\mathcal{H}$ which describes the situation where the sender transmits $n$ states $\{\sigma_j\}_{j=1}^n$ successively. Let $\mathcal{S}_1^{(n)}$ be the subset of $\mathcal{S}(\mathcal{H}_1^{\otimes n})$ each element of which is an affine combination of product states

$\sigma_1 \otimes \cdots \otimes \sigma_n$. Then an affine map $\Gamma^{(n)} : \mathcal{S}_1^{(n)} \to \mathcal{S}(\mathcal{H}_2^{\otimes n})$ is uniquely determined from $\Gamma : \mathcal{S}_1 \to \mathcal{S}(\mathcal{H}_2)$ by the relation $\Gamma^{(n)}(\sigma_1 \otimes \cdots \otimes \sigma_n) = (\Gamma\sigma_1) \otimes \cdots \otimes (\Gamma\sigma_n)$; i.e. $\Gamma^{(n)}$ is the memoryless extension of $\Gamma$. We drop the superscript $(n)$ when no confusion is likely to arise. It may be worth noting that, when $\Gamma$ is the dual of a completely positive map, the domain of $\Gamma^{(n)}$ can be further extended from $\mathcal{S}_1^{(n)}$ to the much wider set $\mathcal{S}(\mathcal{H}_1^{\otimes n})$ in a natural manner. Such an extension will give another interesting setting, but we do not pursue this in the present paper.

A quantum communication system is described as follows. A *quantum codebook* on $\mathcal{H}_1^{\otimes n}$ is a finite set of product states: $\mathcal{C}_n = \{\sigma^{(n)}(1), \ldots, \sigma^{(n)}(L_n)\}$, where $\sigma^{(n)}(k) = \sigma_1(k) \otimes \cdots \otimes \sigma_n(k)$. The transmitter first selects a codeword $\sigma^{(n)} = \sigma_1 \otimes \cdots \otimes \sigma_n$ which corresponds to the message to be transmitted (encoding), and then transmits each signal $\sigma_1, \ldots, \sigma_n$ successively through a memoryless channel $\Gamma$. The receiver then receives signals $\Gamma\sigma_1, \ldots, \Gamma\sigma_n$ and, by means of a certain measuring process, he estimates which signal among $\mathcal{C}_n$ has been actually transmitted (decoding). The decoder is described by a $\mathcal{C}_n$-valued measurement $T^{(n)}$ over $\mathcal{H}_2^{\otimes n}$.

Once a decoder $T^{(n)}$ is fixed arbitrarily, the error probability $P_e(\mathcal{C}_n, T^{(n)})$ averaged over the codewords becomes well-defined in classical sense:

$$P_e(\mathcal{C}_n, T^{(n)}) = \frac{1}{L_n} \sum_{\sigma^{(n)} \in \mathcal{C}_n} \left\{ 1 - \mathrm{Tr}\left[ (\Gamma\sigma^{(n)})\, T^{(n)}(\sigma^{(n)}) \right] \right\}.$$

Now the quantity $R_n = \log L_n / n$ is called the *rate* for the code $\mathcal{C}_n$. The (operational) *capacity* $C(\Gamma)$ of the channel $\Gamma$ is then defined by the supremum of $\limsup_{n\to\infty} R_n$ over all sequences of codings $\{\mathcal{C}_n, T^{(n)}\}_n$ which satisfy $\lim_{n\to\infty} P_e(\mathcal{C}_n, T^{(n)}) = 0$.

Let us denote by $\mathcal{M}^{(n)}$ the totality of measurements which take values on finite sets (not necessarily $\mathcal{C}_n$) over the extended output system $\mathcal{H}_2^{\otimes n}$. Note that there are some elements in $\mathcal{M}^{(n)}$ which are essentially reduced to measurements over smaller systems. For instance, there are such measurements $M^{(n)} (\in \mathcal{M}^{(n)})$ that are composed of $n$ measurements $\{M_j\}_{j=1}^n$ over $\mathcal{H}_2^{\otimes 1} = \mathcal{H}_2$ as

$$M^{(n)}(x) = \sum_{y^n : g(y^n) = x} \bigotimes_{j=1}^{n} M_j(y_j). \tag{1}$$

This equation is read as follows: performing $M^{(n)}$ to the composite system $\mathcal{H}_2^{\otimes n}$ is equivalent to performing $M_1, \ldots, M_n$ to $n$ copies of the component system $\mathcal{H}_2$ followed by a data processing $g : y^n = (y_1, \ldots, y_n) \mapsto x$. Such measurements are, however, very special ones and most elements of $\mathcal{M}^{(n)}$ cannot be reduced to measurements over subsystems $\mathcal{H}_2^{\otimes k}$ $(k < n)$. In order to implement such a measurement physically, we must invoke some kind of quantum correlation called the *quantum entanglement* among $n$ component systems.

Once a measurement $M^{(n)} (\in \mathcal{M}^{(n)})$ is arbitrarily fixed, we have the classical mutual information[1]

$$I^{(n)}(p^{(n)}, M^{(n)}; \Gamma) \stackrel{\mathrm{def}}{=} \sum_{\sigma^{(n)}} p^{(n)}(\sigma^{(n)}) D_{M^{(n)}}\left( \Gamma\sigma^{(n)} \middle\| \Gamma\rho^{(n)} \right). \tag{2}$$

[1] Let $X$ be an $\mathcal{S}_1^{(n)}$-valued random variable which represents the input states and $Y$ a random variable which represents the measurement outcomes. Then $p(Y = y \,|\, X = \sigma^{(n)}) = \mathrm{Tr}\,[(\Gamma\sigma^{(n)}) M^{(n)}(y)]$ and $p(X = \sigma^{(n)}, Y = y) = p^{(n)}(\sigma^{(n)}) \mathrm{Tr}\,[(\Gamma\sigma^{(n)}) M^{(n)}(y)]$, and $I(X;Y) = \sum_x p(x) D(p(y|x) \| p(y))$ becomes (2).

Here $p^{(n)}(\sigma^{(n)}) = p^{(n)}(\sigma_1, \ldots, \sigma_n)$ is an arbitrary joint distribution with finite support over $\mathcal{S}_1^n = \mathcal{S}_1 \times \cdots \times \mathcal{S}_1$. Let us denote the totality of such distributions by $\mathcal{P}^{(n)}$. Further $D_{M^{(n)}}$ is the Kullback-Leibler divergence between the classical probability distributions $\mathrm{Tr}\,[(\Gamma\sigma^{(n)})M^{(n)}(\cdot)]$ and $\mathrm{Tr}\,[(\Gamma\rho^{(n)})M^{(n)}(\cdot)]$, and

$$\rho^{(n)} \stackrel{\mathrm{def}}{=} \sum_{\sigma^{(n)}} p^{(n)}(\sigma^{(n)})\,\sigma^{(n)} = \sum_{\sigma_1,\ldots,\sigma_n} p^{(n)}(\sigma_1, \ldots, \sigma_n)\,\sigma_1 \otimes \cdots \otimes \sigma_n \tag{3}$$

is the mixture state which is related to the marginal distribution of the measurement outcomes.

Let us introduce for a memoryless channel $\Gamma$ the quantities

$$C^{(n)}(\Gamma) \stackrel{\mathrm{def}}{=} \sup_{p^{(n)} \in \mathcal{P}^{(n)},\, M^{(n)} \in \mathcal{M}^{(n)}} I^{(n)}(p^{(n)}, M^{(n)}; \Gamma). \tag{4}$$

These quantities exhibit the superadditivity $C^{(m+n)}(\Gamma) \geq C^{(m)}(\Gamma) + C^{(n)}(\Gamma)$ and their operational meanings are as follows. $C^{(1)}(\Gamma)$ gives the achievable communication rate when the receiver is permitted to use only restricted decoders of the form (1), i.e., when he cannot use any quantum entanglement over composite systems $\mathcal{H}_2^{\otimes k}$ $(k > 1)$. On the other hand, $C^{(n)}(\Gamma) = C^{(1)}(\Gamma^{(n)})$. These observations immediately lead us to the inequality

$$\frac{C^{(n)}(\Gamma)}{n} \leq C(\Gamma) \tag{5}$$

for all $n$. Furthermore, this relation can be strengthened as follows.

**Proposition 1** *For a memoryless channel* $\Gamma$*,*

$$C(\Gamma) = \lim_{n\to\infty} \frac{C^{(n)}(\Gamma)}{n} = \sup_n \frac{C^{(n)}(\Gamma)}{n}. \tag{6}$$

This primitive version of quantum channel coding theorem [5] is proved along almost the same line to the classical one, see Appendix A. However, since the additivity $C^{(m+n)}(\Gamma) = C^{(m)}(\Gamma) + C^{(n)}(\Gamma)$ does not hold in general, the limiting process in (6) cannot be dispensed with, which is in remarkable contrast to the classical channel.

The problem of finding single-letterized expression for $C(\Gamma)$ had been a long-standing open problem. Historically there was a well-known single-letterized upper bound called the Holevo bound [6]:

$$C(\Gamma) \leq \sup_{p \in \mathcal{P}^{(1)}} \tilde{I}(p; \Gamma), \tag{7}$$

where

$$\tilde{I}(p; \Gamma) \stackrel{\mathrm{def}}{=} \sum_\sigma p(\sigma) D(\Gamma\sigma \| \Gamma\rho), \qquad \left(\rho = \sum_\sigma p(\sigma)\sigma\right)$$

is a formal quantum mutual information defined via the quantum relative entropy $D(\sigma\|\rho) \stackrel{\mathrm{def}}{=} \mathrm{Tr}\,\sigma(\log\sigma - \log\rho)$, a quantum analogue of the Kullback-Leibler divergence. Actually, recalling the

monotonicity relation (see [7] [8, Theorem 1.5], for instance) of the relative entropy $D(\sigma\|\rho) \geq D_M(\sigma\|\rho)$ for all measurements $M$, we see that

$$\tilde{I}^{(n)}(p^{(n)};\Gamma) \stackrel{\text{def}}{=} \sum_{\sigma^{(n)}} p^{(n)}(\sigma^{(n)}) D(\Gamma\sigma^{(n)} \parallel \Gamma\rho^{(n)})$$

exhibits the inequality $\tilde{I}^{(n)}(p^{(n)};\Gamma) \geq I^{(n)}(p^{(n)}, M^{(n)};\Gamma)$ for all $M^{(n)}$. Since

$$\tilde{C}^{(n)}(\Gamma) \stackrel{\text{def}}{=} \sup_{p^{(n)}\in\mathcal{P}^{(n)}} \tilde{I}^{(n)}(p^{(n)};\Gamma)$$

enjoys the additivity $\tilde{C}^{(n)}(\Gamma) = n\tilde{C}^{(1)}(\Gamma)$, we have

$$\tilde{C}^{(1)}(\Gamma) = \lim_{n\to\infty} \frac{\tilde{C}^{(n)}(\Gamma)}{n} \geq \lim_{n\to\infty} \frac{C^{(n)}(\Gamma)}{n} = C(\Gamma).$$

Recently, after the breakthrough by Hausladen *et al.* [9], a definitive result was reported by Holevo [10] and by Schumacher and Westmoreland [11]. They proved the converse inequality of (7), to obtain the following

**Theorem 2** (Quantum channel coding theorem) *For a memoryless channel* $\Gamma$*,*

$$C(\Gamma) = \sup_{p\in\mathcal{P}^{(1)}} \tilde{I}(p;\Gamma). \tag{8}$$

Theorem 2 implies that $\sup_p \tilde{I}(p;\Gamma)$ precisely gives the single-letterized formula for $C(\Gamma)$. This result must lead us to a deeper stage of quantum channel coding theory.

The purpose of this paper is to rearrange and develop some basic characteristics of the operational quantum capacity $C(\Gamma)$ through detailed analyses of quantities $\tilde{I}(p;\Gamma)$ and $I^{(1)}(p, M;\Gamma)$. In Section 2, we show that even if an adaptive strategy of measurement is employed and a kind of feedback is permitted in encoding, the capacity cannot exceed the quantity $C^{(1)}(\Gamma)$ without an essential use of quantum entanglement by the receiver. In Section 3, Theorem 2 is reconsidered from a viewpoint of jointly typical decoding scheme. In Section 4, we prove that the supremizations of $\tilde{I}(p;\Gamma)$ and $I^{(1)}(p, M;\Gamma)$ can be reduced to maximizations on certain finite dimensional compact sets. In Section 5, we introduce the notion of pseudoclassical channel for which $C(\Gamma) = C^{(1)}(\Gamma)$ holds, and derive the necessary and sufficient condition for a quantum channel to be pseudoclassical. In Section 6, we scrutinize quantum binary channels for two-level quantum systems. A geometrical implication for a quantum binary channel to be pseudoclassical is explicitly presented. The final Section 7 gives concluding remarks.

## 2 Adaptive measurement with feedback code

The quantity $C^{(1)}(\Gamma)$ is the capacity of the classical memoryless channel obtained by choosing an optimal measurement $M^{(1)}$ for a single output system. In the present section we show that even if a measurement is optimized in a wider class — adaptive measurements — and even if a kind of

feedback is permitted in encoding, the capacity cannot exceed $C^{(1)}(\Gamma)$. This result enables us to well-recognize the significance of introducing measurements over the composite system of $n$ outputs.

A $\mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n$-valued measurement $M^{(n)}$ over $\mathcal{H}^{\otimes n}$ where $\{\mathcal{Y}_j\}$ are arbitrary finite sets, is called *adaptive* if it takes the form

$$M^{(n)}(y^n) = \bigotimes_{j=1}^{n} M_j(y_j \,|\, y^{j-1}).$$

Here $j$ denotes the order of events, $M_j(\,\cdot\,|\,y^{j-1})$ is a $\mathcal{Y}_j$-valued measurement over $\mathcal{H}$ possibly depending on the previous data $y^{j-1} = (y_1, \ldots, y_{j-1}) \in \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_{j-1}$, and each $y_k$ is the outcome of the measurement $M_k(\,\cdot\,|\,y^{k-1})$ applied to the $k$th output state $\Gamma\sigma_k$. The notion of a *feedback code* associated with such an adaptive measurement can be introduced as in a classical channel. Letting $\mathcal{W}_n = \{1, 2, \ldots, L_n\}$ be the set of messages to be transmitted, an encoder is represented by an $n$-tuple $(f_1, \ldots, f_n)$ of mappings of the type $f_j : \mathcal{W}_n \times \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_{j-1} \to \mathcal{S}(\mathcal{H}_1)$, and a decoder is a mapping $g_n : \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n \to \mathcal{W}_n$.

When an element $w \in \mathcal{W}_n$ is chosen to be transmitted, the sequence of input states $\sigma^{(n)} = \sigma_1 \otimes \cdots \otimes \sigma_n$ is generated according to $\sigma_j = f_j(w, y^{j-1})$ successively for $j = 1, \ldots, n$. Of course every usual encoder without feedback (: $\mathcal{W}_n \to \mathcal{S}(\mathcal{H}_1)^n$ ) is included as a special case in this encoding scheme. After getting the total outcomes $y^n = (y_1, \ldots, y_n)$, the decoder yields the estimate $\hat{w} = g_n(y^n)$ of the transmitted message $w$. In other words, the decoding procedure is a $\mathcal{W}_n$-valued measurement of the form

$$T^{(n)}(w) = \sum_{y^n : g_n(y^n) = w} M^{(n)}(y^n).$$

For such a coding system $\Phi_n = (M^{(n)} = \otimes_j M_j, \{f_j\}, g_n)$, the average error probability is defined by $P_e(\Phi_n) = \mathrm{Prob}\,\{W_n \neq g_n(Y^n)\}$, where $W_n$ is a random variable uniformly distributed on $\mathcal{W}_n$, and $Y^n$ is the corresponding total measurement outcomes. Consider sequences of codes $\{\Phi_n\}_n$ which satisfy $\lim_{n\to\infty} P_e(\Phi_n) = 0$, and denote by $C^{\otimes}(\Gamma)$ the supremum of $\lim_{n\to\infty} \log L_n/n$ over such sequences.

**Theorem 3**

$$C^{\otimes}(\Gamma) = C^{(1)}(\Gamma).$$

**Proof** $C^{\otimes}(\Gamma) \geq C^{(1)}(\Gamma)$ is trivial. We show the converse inequality. Let $M^{(n)}(y^n) = \otimes_j M_j(y_j \,|\, y^{j-1})$ be an adaptive measurement and $\Phi_n = (M^{(n)}, \{f_j\}, g_n)$ be a feedback code with the message set $\mathcal{W}_n = \{1, \ldots, L_n\}$. Further let $W$ be a random variable which is uniformly distributed on $\mathcal{W}_n$, and $X^n = (X_1, \ldots, X_n)$ and $Y^n = (Y_1, \ldots, Y_n)$ be, respectively, the corresponding input states and measurement outcomes when the message $W$ is chosen to be transmitted. In other words, $X_j = f_j(W, Y^{j-1})$ is a $\mathcal{S}(\mathcal{H}_1)$-valued random variable, and $Y_j$ is a $\mathcal{Y}_j$-valued random variable representing the outcome of the measurement $M_j(\,\cdot\,|Y^{j-1})$ applied to the output state $\Gamma X_j$. Now, by virtue of Fano's inequality and Lemma 4 which shall be proved below, we have

$$\begin{aligned} \log 2 + P_e(\Phi_n) \log L_n &\geq H(W|Y^n) \\ &= H(W) - I(W; Y^n) \\ &\geq \log L_n - nC^{(1)}(\Gamma). \end{aligned}$$

Thus we have

$$(1 - P_e(\Phi_n)) \log L_n \leq \log 2 + nC^{(1)}(\Gamma).$$

Since $\lim_{n\to\infty} P_e(\Phi_n) = 0$ is assumed, it follows that

$$\limsup_{n\to\infty} \frac{\log L_n}{n} \leq C^{(1)}(\Gamma),$$

which proves $C^{\otimes}(\Gamma) \leq C^{(1)}(\Gamma)$. □

**Lemma 4**

$$I(W; Y^n) \leq nC^{(1)}(\Gamma).$$

**Proof** The chain rule of the mutual information asserts that

$$I(W; Y^n) = \sum_{j=1}^{n} I(W; Y_j \,|\, Y^{j-1}).$$

Here we observe

$$\begin{aligned} I(W; Y_j \,|\, Y^{j-1}) &= I(WX_j; Y_j \,|\, Y^{j-1}) \\ &= I(X_j; Y_j \,|\, Y^{j-1}) + I(W; Y_j \,|\, X_j Y^{j-1}) \\ &= I(X_j; Y_j \,|\, Y^{j-1}), \end{aligned}$$

where the first equality follows from the fact that $X_j = f_j(W, Y^{j-1})$, the second equality follows from the chain rule, and the third equality follows from the fact that $W \to X_j Y^{j-1} \to Y_j$ forms a Markov chain in this order:

$$p(y_j \,|\, w x_j y^{j-1}) = \mathrm{Tr}\,[(\Gamma x_j) M_j(y_j \,|\, y^{j-1})] = p(y_j \,|\, x_j y^{j-1}).$$

Furthermore,

$$\begin{aligned} I(X_j; Y_j \,|\, Y^{j-1}) &= \sum_{y^{j-1}} p(y^{j-1}) I(X_j; Y_j \,|\, Y^{j-1} = y^{j-1}) \\ &= \sum_{y^{j-1}} p(y^{j-1}) I^{(1)}(p_{X_j}(\,\cdot\,), M_j(\,\cdot \,|\, y^{j-1})\,; \Gamma) \\ &\leq \sum_{y^{j-1}} p(y^{j-1}) \sup_{p,M} I^{(1)}(p, M\,; \Gamma) \\ &= C^{(1)}(\Gamma). \end{aligned}$$

The lemma then immediately follows. □

Theorem 3 implies that the capacity $C(\Gamma)$ cannot be attained by means of an adaptive measurement with a feedback unless $C(\Gamma) = C^{(1)}(\Gamma)$. Thus, it is essential for the capacity $C(\Gamma)$ to consider measurements over the extended Hilbert space which cannot be realized in an adaptive manner.

# 3 On quantum extension of jointly typical decoding

In classical theory, a simple proof of channel coding theorem was provided by the jointly typical decoding [12]. In this section, we try to clarify the reason why a simple application of jointly typical decoding scheme fails in quantum theory.

Let us introduce the quantity

$$\overleftrightarrow{C}^{(n)}(\Gamma) \stackrel{\rm def}{=} \sup_{p^{(n)}\in\mathcal{P}^{(n)}} \sum_{\sigma^{(n)}} p^{(n)}(\sigma^{(n)}) \left[ \sup_{M^{(n)}\in\mathcal{M}^{(n)}} D_{M^{(n)}}(\Gamma\sigma^{(n)}\|\Gamma\rho^{(n)}) \right]. \tag{9}$$

Note that the position of $\sup_{M^{(n)}}$ has been shifted as compared with (2) and (4): the notation $\overleftrightarrow{C}^{(n)}$ indicates that. It is obvious that

$$C^{(n)}(\Gamma) \le \overleftrightarrow{C}^{(n)}(\Gamma) \le \tilde{C}^{(n)}(\Gamma) = n \sup_p \tilde{I}(p;\Gamma). \tag{10}$$

While the following is a direct consequence of Proposition 1 and Theorem 2, we give an alternative proof without invoking them.

**Theorem 5**

$$\lim_{n\to\infty} \frac{\overleftrightarrow{C}^{(n)}(\Gamma)}{n} = \sup_n \frac{\overleftrightarrow{C}^{(n)}(\Gamma)}{n} = \sup_p \tilde{I}(p;\Gamma). \tag{11}$$

**Proof** The first equality follows from the superadditivity

$$\overleftrightarrow{C}^{(m+n)}(\Gamma) \ge \overleftrightarrow{C}^{(m)}(\Gamma) + \overleftrightarrow{C}^{(n)}(\Gamma),$$

which is easily verified as in the superadditivity of $C^{(n)}$. Observing (10), we only need to show

$$\lim_{n\to\infty} \frac{\overleftrightarrow{C}^{(n)}(\Gamma)}{n} \ge \sup_p \tilde{I}(p;\Gamma). \tag{12}$$

Let $p$ be an arbitrary probability distribution on $\mathcal{S}(\mathcal{H}_1)$ with a finite support and $p^{(n)}(\sigma^{(n)}) = p(\sigma_1)\cdots p(\sigma_n)$ its i.i.d. extension. In this case, $\Gamma\rho^{(n)} = \overset{n}{\otimes} \Gamma\rho^{(1)}$ holds where $\rho^{(n)}$ is the marginal state (3). Hiai and Petz [13] have proved that for an arbitrary state $\sigma_0^{(n)}$ in $\mathcal{S}(\mathcal{H}^{\otimes n})$ and for an arbitrary state $\rho_0$ in $\mathcal{S}(\mathcal{H})$, there exists a measurement $M^{(n)}$ over $\mathcal{H}^{\otimes n}$ which satisfies

$$D_{M^{(n)}}(\sigma_0^{(n)}\|\overset{n}{\otimes}\rho_0) \ge D(\sigma_0^{(n)}\|\overset{n}{\otimes}\rho_0) - K\log(n+1),$$

where $K = \dim\mathcal{H}$. Replacing $\sigma_0^{(n)}$ and $\rho_0$ with $\Gamma\sigma^{(n)}$ and $\Gamma\rho^{(1)}$, respectively, we have

$$\begin{aligned} \sup_{M^{(n)}} D_{M^{(n)}}(\Gamma\sigma^{(n)}\|\Gamma\rho^{(n)}) &\ge D(\Gamma\sigma^{(n)}\|\Gamma\rho^{(n)}) - K\log(n+1) \\ &= \sum_{j=1}^n D(\Gamma\sigma_j\|\Gamma\rho^{(1)}) - K\log(n+1). \end{aligned}$$

This implies that

$$\sum_{\sigma^{(n)}} p^{(n)}(\sigma^{(n)}) \sup_{M^{(n)}} D_{M^{(n)}}(\Gamma\sigma^{(n)} \| \Gamma\rho^{(n)}) \geq n \sum_{\sigma} p(\sigma) D(\Gamma\sigma \| \Gamma\rho^{(1)}) - K \log(n+1). \tag{13}$$

We thus have

$$\overleftrightarrow{C}^{(n)}(\Gamma) \geq n \sup_{p} \tilde{I}(p;\Gamma) - K \log(n+1),$$

and (12) immediately follows. □

Let us apply the above argument to the jointly typical decoding scheme. According to (13), there is a family of measurements $\{M^{(n)}_{\sigma^{(n)}} \, ; \, \sigma^{(n)} \in \mathrm{supp}\,(p^{(n)})\}$ which satisfies

$$n\tilde{I}(p;\Gamma) - K \log(n+1) \leq \sum_{\sigma^{(n)}} p^{(n)}(\sigma^{(n)}) D_{M^{(n)}_{\sigma^{(n)}}}(\Gamma\sigma^{(n)} \| \Gamma\rho^{(n)}) \leq n\tilde{I}(p;\Gamma). \tag{14}$$

Here $p^{(n)}$ is the i.i.d. extension of $p$ and $\mathrm{supp}\,(p^{(n)}) = \{\mathrm{supp}\,(p)\}^n$. Note that the quantity appeared in the middle of inequalities (14) is identical to the Kullback-Leibler divergence $D(Q_1 \| Q_0)$ between the probability distributions

$$Q_1(\sigma^{(n)}, y) \stackrel{\mathrm{def}}{=} p^{(n)}(\sigma^{(n)}) \mathrm{Tr}\left[(\Gamma\sigma^{(n)}) M^{(n)}_{\sigma^{(n)}}(y)\right],$$

$$Q_0(\sigma^{(n)}, y) \stackrel{\mathrm{def}}{=} p^{(n)}(\sigma^{(n)}) \mathrm{Tr}\left[(\Gamma\rho^{(n)}) M^{(n)}_{\sigma^{(n)}}(y)\right].$$

Then applying Stein's lemma to the hypothesis testing for $\{Q_0, Q_1\}$, we see that there exists a family of $\{0,1\}$-valued measurements

$$\left\{ N^{(n)}_{\sigma^{(n)}} = (N^{(n)}_{\sigma^{(n)}}(0), N^{(n)}_{\sigma^{(n)}}(1)) \, ; \, \sigma^{(n)} \in \{\mathrm{supp}\,(p)\}^n \right\}$$

such that as $n \to \infty$,

$$\sum_{\sigma^{(n)}} p^{(n)}(\sigma^{(n)}) \mathrm{Tr}\left[(\Gamma\sigma^{(n)}) N^{(n)}_{\sigma^{(n)}}(1)\right] \to 1, \tag{15}$$

$$\sum_{\sigma^{(n)}} p^{(n)}(\sigma^{(n)}) \mathrm{Tr}\left[(\Gamma\rho^{(n)}) N^{(n)}_{\sigma^{(n)}}(1)\right] \sim e^{-n\tilde{I}(p;\Gamma)}. \tag{16}$$

The error exponent in (16) is due to (14).

Now, suppose that a codebook $\mathcal{C}_n = \{\sigma^{(n)}(1), \ldots, \sigma^{(n)}(L)\} \subset \mathcal{S}_1^{(n)}$ is given and assume that the following decoding procedure is practicable: when a signal is received at the extended output system $\mathcal{H}_2^{\otimes n}$, apply all the measurements $\left\{ N^{(n)}_k \stackrel{\mathrm{def}}{=} N^{(n)}_{\sigma^{(n)}(k)} \, ; \, k = 1, 2, \ldots, L \right\}$ "simultaneously" to the signal, and take $\hat{k}$ as the decoded message if $N^{(n)}_{\hat{k}}$ yields the value 1 and all the other $\{N^{(n)}_k \, ; \, k \neq \hat{k}\}$ yield 0. This procedure is an analogue of the jointly typical set decoding [12] which, together with the random coding technique, provides a proof of the direct part of the classical channel coding

theorem. Actually, according to (15) (16), the above decoding procedure, when applied to the random code generated by $p$, exhibits a similar performance to the jointly typical set decoding. We thus conclude that there exists a code possessing the rate arbitrarily close to $\tilde{I}(p;\Gamma)$ and the error probability arbitrarily close to 0. In other words, $\tilde{I}(p;\Gamma)$ is attainable, so that $C(\Gamma) = \sup_p \tilde{I}(p;\Gamma)$.

Unfortunately, such a decoding procedure is generally impracticable because of the noncommutativity of the measurements $\{N_k^{(n)}\}$. It is also worth noting that the difference between the quantum case and the classical case is understood through the notion of "cloning". That is, suppose the receiver is able to transform the output signal with a state $\tau^{(n)} \in \mathcal{S}(\mathcal{H}_2^{\otimes n})$ into, by some "cloning" device, a signal with such a state $\tilde{\tau}^{(nL)} \in \mathcal{S}((\mathcal{H}_2^{\otimes n})^{\otimes L})$ that its all $L$ marginal states coincide with $\tau^{(n)}$. Then all the measurements $\{N_k^{(n)}\}$ can be simultaneously applied in the form of $N_1^{(n)} \otimes \cdots \otimes N_L^{(n)}$ to the transformed signal. In classical case, there is no prohibition of such a cloning procedure. In quantum case, on the other hand, states cannot be cloned in general. Consequently, the above "proof" of attainability of $\tilde{I}(p;\Gamma)$ is fictitious. This argument suggests that one cannot grasp the gist of Theorem 2 by means of a simple application of jointly typical decoding scheme. Indeed, the proofs in [10] [11] are based on highly elaborated noncommutative extensions of jointly typical decoder.

# 4 Supremizations of $\tilde{I}(p;\Gamma)$ and $I^{(1)}(p, M;\Gamma)$

The aim of this section is to show that the supremizations of $\tilde{I}(p;\Gamma)$ and $I^{(1)}(p, M;\Gamma)$ can be reduced to maximizations on certain finite dimensional compact sets. In the sequel we drop the superscript (1) in $\mathcal{P}^{(1)}, \mathcal{M}^{(1)}$, and $I^{(1)}$ and simply write them as $\mathcal{P}, \mathcal{M}$, and $I$, respectively.

We first explore the supremization of $\tilde{I}(p;\Gamma)$. Let us rewrite as

$$\sup_{p\in\mathcal{P}} \tilde{I}(p;\Gamma) = \sup_{\rho\in\mathcal{S}(\mathcal{H}_1)} \sup_{p\in\mathcal{P}(\rho)} \tilde{I}(p;\Gamma),$$

where

$$\mathcal{P}(\rho) \stackrel{\text{def}}{=} \{p \in \mathcal{P} \ ; \ \textstyle\sum_\sigma p(\sigma)\sigma = \rho\}.$$

Naturally $\mathcal{P}$ is regarded as a convex set, and $\mathcal{P}(\rho)$ forms a convex subset of $\mathcal{P}$ for each $\rho$. In the sequel we denote an element of $\mathcal{P}$ having a support set $\mathrm{supp}\,(p) = \{\sigma_1, \ldots, \sigma_n\}$ as

$$p = \sum_{j=1}^{n} \lambda_j \delta_{\sigma_j},$$

where $\lambda_j > 0$, $\sum_j \lambda_j = 1$, and $\delta_{\sigma_j}$ denotes the simple probability measure concentrated at the point $\sigma_j$. The barycenter $\sum_\sigma p(\sigma)\sigma$ of $p$, which is constrained to be $\rho$ in $\mathcal{P}(\rho)$, is then written as $\sum_j \lambda_j \sigma_j$.

For a while, we restrict ourselves to the case where $\mathcal{H}_1 = \mathcal{H}_2 = \mathcal{H}$ and $\Gamma$ is the identity map on $\mathcal{S} = \mathcal{S}_1 \subset \mathcal{S}(\mathcal{H})$, which causes no loss of generality as was already pointed out in Section 1. The subsequent lemmas and corollaries are derived only from the fact that $\mathcal{S}$ is a compact convex set in a finite-dimensional affine space, regardless of the inclusion $\mathcal{S} \subset \mathcal{S}(\mathcal{H})$, and their derivations are essentially parallel with the proof of Caratheodory's theorem [14][15]. The importance of such an argument in an information theoretical context was emphasized by Davies [16].

For a subset $A$ of a convex set, a point $x \in A$ is called *extreme* if $x$ cannot be represented as a nontrivial mixture of points in $A$, and the totality of extreme points of $A$ is denoted by $\partial_e A$ (the extreme boundary of $A$). We use this notation whether $A$ is convex or not. In passing, the condition of affine (in)dependence appeared in the following can be replaced by linear (in)dependence. However we use the former because its use is conceptually natural in an affine space.

**Lemma 6** *For all $\rho \in \mathcal{S}$,*

$$\partial_e \mathcal{P}(\rho) = \{p \in \mathcal{P}(\rho) \; ; \; \mathrm{supp}\,(p) \text{ is an affinely independent subset of } \mathcal{S}\}.$$

**Proof** Let $p = \sum_{j=1}^n \lambda_j \delta_{\sigma_j}$ $(\forall \lambda_j > 0)$. We first assume that $\{\sigma_1, \ldots, \sigma_n\}$ is affinely dependent:

$$\sum_{j=1}^n \alpha_j \sigma_j = 0, \qquad \sum_{j=1}^n \alpha_j = 0,$$

where the coefficients $\alpha_1, \ldots, \alpha_n$ are not all zero. Define $q_1 = \sum_j (\lambda_j + \varepsilon \alpha_j) \delta_{\sigma_j}$ and $q_2 = \sum_j (\lambda_j - \varepsilon \alpha_j) \delta_{\sigma_j}$ for $\varepsilon$ sufficiently small. Then $q_1, q_2 \in \mathcal{P}(\rho)$, $q_1 \neq q_2$, and $p = \frac{1}{2} q_1 + \frac{1}{2} q_2$, which shows that $p$ is not an extreme point.

We next assume that $p$ is a non-extreme point, say $p = a q_1 + (1-a) q_2$, where $q_1, q_2 \in \mathcal{P}(\rho)$, $q_1 \neq q_2$, $0 < a < 1$. Necessarily $\mathrm{supp}\,(q_i) \subset \mathrm{supp}\,(p)$ and hence $q_i$ is written as

$$q_i = \sum_{j=1}^n \alpha_{ij} \delta_{\sigma_j},$$

where $\alpha_{ij} \geq 0$ and $\sum_{j=1}^n \alpha_{ij} = 1$ for $i = 1, 2$. Since $\sum_{j=1}^n \alpha_{ij} \sigma_j = \rho$ for $i = 1, 2$,

$$\sum_{j=1}^n (\alpha_{1j} - \alpha_{2j}) \sigma_j = 0.$$

The coefficients $\alpha_{1j} - \alpha_{2j}$ have zero sum, and they do not all vanish since $q_1 \neq q_2$. Hence $\{\sigma_1, \ldots, \sigma_n\}$ is affinely dependent. □

**Corollary 7** $|\mathrm{supp}\,(p)| \leq \dim \mathcal{S} + 1$ *for all $\rho \in \mathcal{S}$ and $p \in \partial_e \mathcal{P}(\rho)$.*

We next consider the following convex subset of $\mathcal{P}(\rho)$:

$$\mathcal{P}^{(e)}(\rho) \stackrel{\mathrm{def}}{=} \{p \in \mathcal{P}(\rho) \; ; \; \mathrm{supp}\,(p) \subset \partial_e \mathcal{S}\}.$$

Note that when $\mathcal{S} = \mathcal{S}(\mathcal{H})$, an element of $\partial_e \mathcal{S}$ is nothing but a *pure state*, which takes the form $|\phi\rangle\langle\phi|$ where $\phi$ is a normalized vector in $\mathcal{H}$.

**Lemma 8** $\mathcal{P}(\rho) = \mathrm{co}\,\partial_e \mathcal{P}(\rho)$ *and* $\mathcal{P}^{(e)}(\rho) = \mathrm{co}\,\partial_e \mathcal{P}^{(e)}(\rho)$, *where* co *denotes the convex hull.*

**Proof** We will show a more general assertion that for an arbitrary subset $X$ of $\mathcal{S}$,

$$\mathcal{P}_X(\rho) = \mathrm{co}\,\partial_e\mathcal{P}_X(\rho)$$

holds, where

$$\mathcal{P}_X(\rho) \stackrel{\mathrm{def}}{=} \{p \in \mathcal{P}(\rho)\ ;\ \mathrm{supp}\,(p) \subset X\}.$$

Since $\mathcal{P}_X(\rho)$ is convex, $\mathcal{P}_X(\rho) \supset \mathrm{co}\,\partial_e\mathcal{P}_X(\rho)$ is trivial. To show the inverse inclusion, we note that

$$\mathcal{P}_X(\rho) = \bigcup_{F\in\mathcal{F}_X} \mathcal{P}_F(\rho), \tag{17}$$

where $\mathcal{F}_X$ denotes the totality of finite subsets of $X$. For each $F \in \mathcal{F}_X$, we claim:

$$\mathcal{P}_F(\rho) = \mathrm{co}\,\partial_e\mathcal{P}_F(\rho), \tag{18}$$

$$\partial_e\mathcal{P}_F(\rho) = \mathcal{P}_F(\rho) \cap \partial_e\mathcal{P}_X(\rho)\ (\subset \partial_e\mathcal{P}_X(\rho)). \tag{19}$$

Due to the finiteness of $F$, $\mathcal{P}_F(\rho)$ is naturally regarded as a compact set in a finite dimensional vector space, and hence the first relation (18) is an immediate consequence of Krein-Milman's extreme point theorem [17]. The second relation (19) is easily seen by observing that $\mathcal{P}_F(\rho)$ is a *face* of $\mathcal{P}_X(\rho)$; i.e. a convex combination of points $\{p_j\}$ in $\mathcal{P}_X(\rho)$ belongs to $\mathcal{P}_F(\rho)$ if and only if all the $p_j$'s belong to $\mathcal{P}_F(\rho)$, see [15]. The desired inclusion relation is then derived from (17)-(19) as

$$\mathcal{P}_X(\rho) = \bigcup_{F\in\mathcal{F}_X} \mathrm{co}\,\partial_e\mathcal{P}_F(\rho) \subset \mathrm{co} \bigcup_{F\in\mathcal{F}_X} \partial_e\mathcal{P}_F(\rho) \subset \mathrm{co}\,\partial_e\mathcal{P}_X(\rho).$$

This proves the assertion. □

It should be noted that the relation (18) can be understood without a topological argument. In a similar way to Lemma 6, we have

$$\partial_e\mathcal{P}_X(\rho) = \{p \in \mathcal{P}_X(\rho)\ ;\ \mathrm{supp}\,(p) \text{ is an affinely independent subset of } X\ \}.$$

Now, given a $p \in \mathcal{P}_F(\rho)$ arbitrarily, we can pick out all the affinely independent subsets of $\mathrm{supp}\,(p)$ each of which can have the barycenter $\rho$. By using them, $p$ can be represented in an affine combination of elements of $\partial_e\mathcal{P}_F(\rho)$, which implies (18).

Let us apply these preliminary considerations to the analysis of the supremization in $C(\mathrm{id}) = \sup_{p\in\mathcal{P}} \tilde{I}(p)$, where

$$\tilde{I}(p) \stackrel{\mathrm{def}}{=} \tilde{I}(p;\mathrm{id}) = \sum_\sigma p(\sigma)D(\sigma\|\rho).$$

**Proposition 9** *For an arbitrary $p \in \mathcal{P}(\rho)$, there exists a $q \in \partial_e\mathcal{P}^{(e)}(\rho)$ such that $\tilde{I}(q) \geq \tilde{I}(p)$.*

**Proof** Let $p = \sum_j \lambda_j\delta_{\sigma_j}$ be an arbitrary element of $\mathcal{P}(\rho)$ and choose an $r_j \in \mathcal{P}^{(e)}(\sigma_j)$ arbitrarily for each $j$. Here the nonemptiness of $\mathcal{P}^{(e)}(\sigma)$ for all $\sigma \in \mathcal{S}$ is assured by Krein-Milman's theorem: $\mathcal{S} = \mathrm{co}\,\partial_e\mathcal{S}$. Because of the convexity of relative entropy, we have

$$D(\sigma_j\|\rho) = D(\sum_\tau r_j(\tau)\tau\|\rho) \leq \sum_\tau r_j(\tau)D(\tau\|\rho),$$

and therefore

$$\tilde{I}(p) = \sum_j \lambda_j D(\sigma_j \| \rho) \leq \sum_\tau q'(\tau) D(\tau \| \rho) = \tilde{I}(q'),$$

where $q' \stackrel{\text{def}}{=} \sum_j \lambda_j r_j$. Obviously $q'$ belongs to $\mathcal{P}^{(e)}(\rho)$. Furthermore, owing to the linearity of $\tilde{I}(p)$ in $p$ and to Lemma 8, there always exists a point $q$ in $\partial_e \mathcal{P}^{(e)}(\rho)$ satisfying $\tilde{I}(q) \geq \tilde{I}(q')$, which completes the proof. □

**Corollary 10** *For an arbitrary $p \in \mathcal{P}$, there exists a $q \in \mathcal{Q}$ such that $\tilde{I}(q) \geq \tilde{I}(p)$, where*

$$\mathcal{Q} \stackrel{\text{def}}{=} \bigcup_{\rho \in \mathcal{S}} \partial_e \mathcal{P}^{(e)}(\rho) = \{p \in \mathcal{P} \ ; \ \text{supp}\,(p) \textit{ is an affinely independent subset of } \partial_e \mathcal{S}\}.$$

A straightforward consequence of Corollary 10 is $\sup_{p \in \mathcal{P}} \tilde{I}(p) = \sup_{p \in \mathcal{Q}} \tilde{I}(p)$. We further claim that the supremum can be replaced with the maximum. To see this, it is sufficient to show that there is a set $A$ which satisfies $\mathcal{Q} \subset A \subset \mathcal{P}$ and is properly topologized so that $A$ becomes compact and the function $\tilde{I}$ is continuous on $A$. This is actually carried out by setting

$$A = \mathcal{P}_m \stackrel{\text{def}}{=} \{p \in \mathcal{P} \ ; \ |\text{supp}\,(p)| \leq m\},$$

where $m = \dim \mathcal{S} + 1$. Indeed let us introduce

$$\hat{\mathcal{P}}_m \stackrel{\text{def}}{=} \{(\lambda_1, \ldots, \lambda_m, \sigma_1, \ldots, \sigma_m) \in \mathbf{R}^m \times \mathcal{S}^m \ ; \ \forall \lambda_j \geq 0, \ \sum_{j=1}^m \lambda_j = 1\},$$

which is clearly compact with respect to the natural topology, and define the surjective map

$$\omega : \hat{\mathcal{P}}_m \longrightarrow \mathcal{P}_m : (\lambda_1, \ldots, \lambda_m, \sigma_1, \ldots, \sigma_m) \longmapsto \sum_j \lambda_j \delta_{\sigma_j}.$$

Then the composite function $\tilde{I} \circ \omega$ is continuous on the compact set $\hat{\mathcal{P}}_m$. Endowed with the quotient topology by $\omega$, $\mathcal{P}_m$ is found to be the desired compact set $A$.

Translating the above result into the original situation where $\Gamma$ is an arbitrary affine map from $\mathcal{S}_1$ ($\subset \mathcal{S}(\mathcal{H}_1)$) to $\mathcal{S}(\mathcal{H}_2)$, we reach the following theorem.

**Theorem 11**

$$C(\Gamma) = \max_{p \in \mathcal{Q}} \tilde{I}(p; \Gamma) = \max_{p \in \mathcal{Q}'} \tilde{I}(p; \Gamma)$$

*where*

$$\begin{aligned} \mathcal{Q} &\stackrel{\text{def}}{=} \{p \in \mathcal{P} \ ; \ \Gamma(\text{supp}\,(p)) \textit{ is an affinely independent subset of } \partial_e \Gamma(\mathcal{S}_1)\}, \\ \mathcal{Q}' &\stackrel{\text{def}}{=} \{p \in \mathcal{P}^{(e)} \ ; \ |\text{supp}\,(p)| \leq \dim \Gamma(\mathcal{S}_1) + 1\}. \end{aligned}$$

We next tackle the supremization problem of $I(p, M; \Gamma)$ with respect to $p \in \mathcal{P}$ and $M \in \mathcal{M}$. It should be mentioned here that Davies studied essentially the same problem in [16] and obtained a result which is comparable to our Theorem 17 shown below. In his derivation, the supremization with respect to a measurement $M = (M_j)$ is converted into a supremization with respect to a distribution $p = \sum_j \lambda_j \delta_{\sigma_j}$ by the correspondence $\sigma_j = M_j/\mathrm{Tr}\, M_j \in \mathcal{S}(\mathcal{H}_2)$ and $\lambda_j = \mathrm{Tr}\, M_j/\dim \mathcal{H}_2$. We try to give a more transparent proof, treating a measurement as itself.

With no loss of generality, we can regard $\mathcal{M}$ as the totality of sequences $M = (M_1, M_2, \ldots)$, where $M_j$'s are nonnegative Hermitian operators on $\mathcal{H}_2$ vanishing except for a finite number of $j$'s and satisfying $\sum_j M_j = I$. There is a natural convex structure on $\mathcal{M}$: for $M = (M_j)_{j=1}^{\infty}, N = (N_j)_{j=1}^{\infty} \in \mathcal{M}$ and $a \in [0, 1]$,

$$aM + (1-a)N = (aM_j + (1-a)N_j)_{j=1}^{\infty} \in \mathcal{M}.$$

**Lemma 12**

$$\partial_e \mathcal{M} \subset \{M \in \mathcal{M}\ ;\ (M_j)_{j \in \mathrm{supp}\,(M)} \textit{ is a linearly independent sequence of operators}\},$$

*where* $\mathrm{supp}\,(M) \stackrel{\mathrm{def}}{=} \{j\ ;\ M_j \neq 0\}$.

**Proof** Assume that $(M_j)_{j \in \mathrm{supp}\,(M)}$ is linearly dependent: $\sum_j \alpha_j M_j = 0$ where $\{\alpha_j M_j\}$ are not all zero. Define $N_j^{(1)} = (1 + \varepsilon\alpha_j)M_j$ and $N_j^{(2)} = (1 - \varepsilon\alpha_j)M_j$ for sufficiently small $\varepsilon > 0$. Then $N^{(1)} = (N_j^{(1)})_{j=1}^{\infty}$ and $N^{(2)} = (N_j^{(2)})_{j=1}^{\infty}$ are different elements of $\mathcal{M}$ and $M = \frac{1}{2}N^{(1)} + \frac{1}{2}N^{(2)}$, which shows that $M$ is not an extreme point. □

**Corollary 13** $|\mathrm{supp}\,(M)| \leq (\dim \mathcal{H}_2)^2$ *for all* $M \in \partial_e \mathcal{M}$.

Let us define

$$\mathcal{M}^{(e)} \stackrel{\mathrm{def}}{=} \{(M_j) \in \mathcal{M}\ ;\ \mathrm{rank}\, M_j \leq 1 \text{ for } \forall j\},$$

which is not a convex subset of $\mathcal{M}$ but is an *extremal subset* of $\mathcal{M}$; i.e. a convex combination of points $\{M^{(k)}\}$ in $\mathcal{M}$ belongs to $\mathcal{M}^{(e)}$ only if all the $M^{(k)}$'s belong to $\mathcal{M}^{(e)}$, see [17]. Hence

$$\partial_e \mathcal{M}^{(e)} = \mathcal{M}^{(e)} \cap \partial_e \mathcal{M}.$$

Moreover we have

**Lemma 14** $\partial_e \mathcal{M}^{(e)} = \{M \in \mathcal{M}^{(e)}\ ;\ (M_j)_{j \in \mathrm{supp}\,(M)} \textit{ is linearly independent}\}$.

**Proof** From Lemma 12, LHS $\subset$ RHS is obvious. To show LHS $\supset$ RHS, suppose that $M$ is an arbitrary element of RHS and is written as $M = a^{(1)}N^{(1)} + a^{(2)}N^{(2)}$ by some $N^{(1)}, N^{(2)} \in \mathcal{M}^{(e)}$ and $a^{(1)} + a^{(2)} = 1$, $a^{(i)} > 0$. Since $0 \leq a^{(i)}N_j^{(i)} \leq M_j$ and $\mathrm{rank}\, M_j \leq 1$, there exists a constant $c_j^{(i)}$ such that $N_j^{(i)} = c_j^{(i)} M_j$ for each $i, j$. We have

$$\sum_j (c_j^{(1)} - c_j^{(2)}) M_j = \sum_j N_j^{(1)} - \sum_j N_j^{(2)} = I - I = 0,$$

and hence it follows from the linear independence of $(M_j)_{j\in\mathrm{supp}\,(M)}$ that $(c_j^{(1)}-c_j^{(2)})M_j=0$ for all $j$; i.e. $N^{(1)}=N^{(2)}$. This implies that $M\in\partial_e\mathcal{M}^{(e)}$. □

Since, for every finite set $F$ of positive integers, the set $\mathcal{M}_F \stackrel{\mathrm{def}}{=} \{M\in\mathcal{M}\ ;\ \mathrm{supp}\,(M)\subset F\}$ is a compact face of $\mathcal{M}$ and $\mathcal{M}_F^{(e)} \stackrel{\mathrm{def}}{=} \mathcal{M}^{(e)}\cap\mathcal{M}_F$ is a closed (hence compact) subset of $\mathcal{M}_F$, the following lemma is proved in a similar way to Lemma 8.

**Lemma 15** *$\mathcal{M}=\mathrm{co}\,\partial_e\mathcal{M}$ and $\mathcal{M}^{(e)}\subset\mathrm{co}\,\partial_e\mathcal{M}^{(e)}$.*

Let $\mathcal{M}_n^{(e)}$ be the totality of $\{1,2,\ldots,n\}$-valued measurements $M=(M_1,\ldots,M_n)$ over $\mathcal{H}_2$ satisfying $\mathrm{rank}\,M_j\le 1$ for all $j$.

**Proposition 16** *For arbitrary $p\in\mathcal{P}$ and $M\in\mathcal{M}$, there exists an $N\in\partial_e\mathcal{M}_{K^2}^{(e)}$ such that $I(p,N;\Gamma)\ge I(p,M;\Gamma)$, where $K\stackrel{\mathrm{def}}{=}\dim\mathcal{H}_2$.*

**Proof** Let $p\in\mathcal{P}$ and $M\in\mathcal{M}$ be given arbitrarily. We write $I(M)=I(p,M;\Gamma)$ for short. Due to the monotonicity property of Kullback-Leibler divergence with respect to a coarse graining, we can find an $N'\in\mathcal{M}^{(e)}$ satisfying $I(N')\ge I(M)$: indeed it is sufficient to take $N'$ to be a rank one refinement of $M_j$'s [16]. Furthermore, $I(M)$ is a convex function of $M$ due to the joint convexity of divergence and hence Lemma 15 indicates that there always exists an $N\in\partial_e\mathcal{M}^{(e)}$ satisfying $I(N)\ge I(N')$. Since $|\mathrm{supp}\,(N)|\le K^2$ follows from Corollary 13, $N$ can be taken as an element of $\partial_e\mathcal{M}_{K^2}^{(e)}$, which completes the proof. □

In view of Proposition 16, we can use a similar argument to the derivation of Theorem 11 by comparing the obvious inclusion $\partial_e\mathcal{M}_{K^2}^{(e)}\subset\mathcal{M}_{K^2}^{(e)}\subset\mathcal{M}$ with $\mathcal{Q}\subset A\subset\mathcal{P}$ appeared just after Corollary 10. Moreover when $M$ is arbitrarily fixed, the same property as Corollary 10 holds for $I(p,M;\Gamma)$ in $p$. Hence we have the following theorem.

**Theorem 17**

$$C^{(1)}(\Gamma)=\max_{p\in\mathcal{P}}\max_{M\in\partial_e\mathcal{M}_{K^2}^{(e)}} I(p,M;\Gamma),$$

*where the range of $\max_p$ can be reduced to $\mathcal{Q}$ or $\mathcal{Q}'$ as in Theorem 11.*

# 5 Pseudoclassical channels

We say that a channel $\Gamma$ is *pseudoclassical* if $C(\Gamma)=C^{(1)}(\Gamma)$ holds. There is no need for invoking entangled measurements over composite systems iff $\Gamma$ is pseudoclassical. Therefore, for a pseudoclassical channel $\Gamma$, all the quantities $C^{(n)}(\Gamma)/n$, and $\overleftrightarrow{C}^{(n)}(\Gamma)/n$ coincide with the capacity $C(\Gamma)$ just like a classical channel. In this section we explore conditions for $\Gamma$ to be pseudoclassical and demonstrate a simple example.

A distribution $p\in\mathcal{P}$ is called $\Gamma$-*commutative* if $\Gamma(\mathrm{supp}\,(p))$ comprizes mutually commutative density operators in $\mathcal{S}(\mathcal{H}_2)$. The next lemma is due to Holevo [6]. For the reader's convenience, we will outline a slightly simplified proof.

**Lemma 18** *There exists a measurement $M$ satisfying $\tilde{I}(p;\Gamma) = I(p,M;\Gamma)$ iff $p$ is $\Gamma$-commutative.*

**Proof** It suffices to treat the case $\Gamma = \mathrm{id}$ (hence $\mathcal{H}_1 = \mathcal{H}_2 = \mathcal{H}$) and to show that there exists an $M$ satisfying

$$\sum_j \lambda_j D(\sigma_j \| \rho) = \sum_j \lambda_j D_M(\sigma_j \| \rho) \tag{20}$$

iff

$$[\sigma_i, \sigma_j] = 0 \quad \text{for } \forall i, \forall j, \tag{21}$$

where $\forall \lambda_j > 0, \sum_j \lambda_j = 1$ and $\sum_j \lambda_j \sigma_j = \rho$. The 'if' part is then obvious, for the common spectral measure $M$ of $\{\sigma_j\}$ satisfies (20). To show the converse, assume (20) or the equivalent condition

$$D(\sigma_j \| \rho) = D_M(\sigma_j \| \rho) \quad \text{for } \forall j. \tag{22}$$

According to [2, Proposition II.5.1], there exist a Hilbert space $\mathcal{K}$, a pure state $\eta$ on $\mathcal{K}$ and a simple measurement $E = (E_k)$ over $\mathcal{H} \otimes \mathcal{K}$ (i.e. $E_k E_\ell = \delta_{k\ell} E_k$, $\forall k, \forall \ell$), such that

$$\mathrm{Tr}\, \tau M_k = \mathrm{Tr}\, (\tau \otimes \eta) E_k$$

holds for all $\tau \in \mathcal{S}(\mathcal{H})$ and $k$. The condition (22) is then equivalent to

$$D(\sigma_j \otimes \eta \| \rho \otimes \eta) = D_E(\sigma_j \otimes \eta \| \rho \otimes \eta) \quad \text{for } \forall j. \tag{23}$$

Note that the complex linear span $\mathcal{A}$ of $\{E_k\}$ forms a commutative $*$-algebra of operators on $\mathcal{H} \otimes \mathcal{K}$. Then by a slight extension[2] of Petz's theorem [18][8, Proposition 1.16] concerning the equality condition in the monotonicity of $D$, we conclude that (23) is satisfied iff for all $j$, $[\sigma_j \otimes \eta, \rho \otimes \eta] = 0$ and there exists an operator $X_j \in \mathcal{A}$ satisfying

$$\sigma_j \otimes \eta = (\rho \otimes \eta) X_j.$$

Then we have

$$\begin{aligned} (\sigma_i \otimes \eta)(\sigma_j \otimes \eta) &= (\rho \otimes \eta)^2 X_i X_j \\ &= (\rho \otimes \eta)^2 X_j X_i = (\sigma_j \otimes \eta)(\sigma_i \otimes \eta), \end{aligned}$$

which leads to (21). □

**Theorem 19** *A channel $\Gamma$ is pseudoclassical iff $\tilde{I}(p;\Gamma)$ takes the maximum at a $\Gamma$-commutative distribution $p$.*

**Proof** The 'if' part is immediate from Lemma 18. Assume that $\Gamma$ is pseudoclassical and let $(p, M)$ be a pair satisfying $I(p, M; \Gamma) = C^{(1)}(\Gamma)$, the existence of which is ensured by Theorem 17. It then follows from Lemma 18 that $p$ is $\Gamma$-commutative, for

$$I(p, M; \Gamma) = C^{(1)}(\Gamma) = C(\Gamma) \geq \tilde{I}(p; \Gamma) \geq I(p, M; \Gamma).$$

□

The following sufficient conditions are sometimes useful.

[2] The original result was obtained in the restricted case where two strictly positive states are given as arguments of the relative entropy.

**Corollary 20** *Let*

$$\alpha \stackrel{\rm def}{=} \max_{\sigma\in\mathcal{S}_1} \tilde{H}(\Gamma\sigma), \qquad \beta \stackrel{\rm def}{=} \min_{\sigma\in\mathcal{S}_1} \tilde{H}(\Gamma\sigma),$$

*where $\tilde{H}$ denotes the von Neumann entropy: $\tilde{H}(\tau) = -\mathrm{Tr}\,\tau\log\tau$. If there exist a $\rho \in \mathcal{S}_1$ and a $\Gamma$-commutative $p$ in $\mathcal{P}(\rho)$ such that $\tilde{H}(\Gamma\rho) = \alpha$, and for all $\sigma \in \mathrm{supp}\,(p)$, $\tilde{H}(\Gamma\sigma) = \beta$, then $\Gamma$ is pseudoclassical and $C(\Gamma) = \alpha - \beta$.*

**Proof** Obvious from Theorem 19 and the relation

$$\tilde{I}(p;\Gamma) = \tilde{H}(\Gamma\rho) - \sum_\sigma p(\sigma)\tilde{H}(\Gamma\sigma).$$

□

**Corollary 21** *If the image $\mathcal{S}_2 = \Gamma(\mathcal{S}_1)$ of a channel $\Gamma$ is unitarily invariant; i.e. $U\tau U^* \in \mathcal{S}_2$ for all $\tau \in \mathcal{S}_2$ and all unitaries $U$ on $\mathcal{H}_2$, then $\Gamma$ is pseudoclassical and*

$$C(\Gamma) = \log(\dim\mathcal{H}_2) - \min_{\tau\in\mathcal{S}_2} \tilde{H}(\tau).$$

**Proof** Let $\tau$ be a state in $\mathcal{S}_2$ achieving $\beta \stackrel{\rm def}{=} \min_{\tau\in\mathcal{S}_2} \tilde{H}(\tau)$ and denote its Schatten decomposition as

$$\tau = \sum_{j=1}^{K_2} \lambda_j |\psi_j\rangle\langle\psi_j|,$$

where $K_2 = \dim\mathcal{H}_2$ and $\{\psi_j\ ;\ j = 1,\ldots,K_2\}$ is an orthonormal basis of $\mathcal{H}_2$. For every permutation $\pi$ on $\{1,\ldots,K_2\}$, the state

$$\tau_\pi \stackrel{\rm def}{=} \sum_{j=1}^{K_2} \lambda_{\pi(j)} |\psi_j\rangle\langle\psi_j|$$

belongs to $\mathcal{S}_2$ due to the unitary invariance of $\mathcal{S}_2$, and satisfies $\tilde{H}(\tau_\pi) = \beta$. On the other hand

$$\frac{1}{K_2!}\sum_\pi \tau_\pi = \frac{1}{K_2} I \in \mathcal{S}_2,$$

where the summation is taken over all the permutations, and

$$\tilde{H}(\frac{1}{K_2}I) = \log K_2 = \max_{\tau\in\mathcal{S}(\mathcal{H}_2)} \tilde{H}(\tau) = \max_{\tau\in\mathcal{S}_2} \tilde{H}(\tau).$$

Hence the present assertion follows from Corollary 20. □

An obvious example of Corollary 21 is a *surjective* channel $\Gamma$ which maps $\mathcal{S}_1$ onto $\mathcal{S}(\mathcal{H}_2)$. Since the image $\mathcal{S}_2 = \mathcal{S}(\mathcal{H}_2)$ is obviously unitarily invariant and $\min_{\tau\in\mathcal{S}_2} \tilde{H}(\tau) = 0$, $\Gamma$ is pseudoclassical and $C(\Gamma) = \log(\dim\mathcal{H}_2)$. In particular, if $\Gamma$ is *noiseless* in the sense that $\Gamma\sigma = U^*\sigma U$ for all $\sigma \in \mathcal{S}(\mathcal{H})$, where $\mathcal{H} = \mathcal{H}_1 = \mathcal{H}_2$ and $U$ is a fixed unitary operator on $\mathcal{H}$, then $C(\Gamma) = \log(\dim\mathcal{H})$.

# 6 Quantum Binary Channels

In this section we treat a channel whose input and output are two-level quantum systems. Such a simple channel is in a position corresponding to a binary channel in the classical information theory and is called a *quantum binary channel*.

A two-level quantum system, of which the spin 1/2 system is a representative example, is described by the 2-dimensional Hilbert space $\mathbf{C}^2$ and a state of the system can be represented by a $2 \times 2$ Hermitian matrix of the form

$$\rho_\theta = \frac{1}{2} \begin{bmatrix} 1+\theta_3 & \theta_1 - i\theta_2 \\ \theta_1 + i\theta_2 & 1-\theta_3 \end{bmatrix},$$

where $\theta = {}^t(\theta_1, \theta_2, \theta_3)$, with $t$ denoting the transpose, is a column vector belonging to the unit ball

$$\mathcal{V} = \{\theta \in \mathbf{R}^3 \; ; \; \|\theta\|^2 = \theta_1^2 + \theta_2^2 + \theta_3^2 \leq 1\}.$$

The correspondence $\theta \mapsto \rho_\theta$, which is often called the Stokes parametrization, gives an affine isomorphism from $\mathcal{V}$ onto $\mathcal{S}(\mathbf{C}^2)$. The matrix $\rho_\theta$ has the eigenvalues $(1 \pm \|\theta\|)/2$, and hence we have

$$\tilde{H}(\rho_\theta) = h(\frac{1+\|\theta\|}{2}), \tag{24}$$

where $h$ denotes the classical binary entropy:

$$h(p) = -p \log p - (1-p) \log(1-p).$$

It is further noted that two matrices $\rho_\theta$ and $\rho_{\theta'}$ mutually commute iff $\theta$ and $\theta'$ are linearly dependent. These facts will be useful for the later arguments.

An arbitrary channel $\Gamma$ of the type $\mathcal{S}(\mathbf{C}^2) \to \mathcal{S}(\mathbf{C}^2)$ is represented as $\Gamma(\rho_\theta) = \rho_{A\theta+b}$ by a matrix $A \in \mathbf{R}^{3\times 3}$ and a column vector $b \in \mathbf{R}^{3\times 1}$ satisfying $A\mathcal{V} + b \subset \mathcal{V}$. We denote such a channel as $\Gamma = (A, b)$ and will study its property in the sequel, mainly concentrating our attention on the condition for $\Gamma$ to be pseudoclassical.

We first consider the case $b = 0$. Note that the required assumption $A\mathcal{V} \subset \mathcal{V}$ is equivalent to $\|A\| \leq 1$, where $\|A\|$ is the matrix norm of $A$ defined by

$$\|A\| \stackrel{\text{def}}{=} \max_{\theta \in V} \|A\theta\| = \max_{\theta : \|\theta\|=1} \|A\theta\|$$

or, in other words, the maximum singular value of $A$. The capacity formula in the following theorem tempts us to call $\Gamma = (A, 0)$ a *quantum binary symmetric channel*.

**Theorem 22** *The channel $\Gamma = (A, 0)$ is pseudoclassical and its capacity is given by*

$$C(\Gamma) = \log 2 - h(\frac{1+\|A\|}{2}).$$

**Proof** Let $\hat{\theta}$ be a unit vector satisfying $\|A\hat{\theta}\| = \|A\|$. Then we have

$$\tilde{H}(\Gamma\rho_{\pm\hat{\theta}}) = h(\frac{1+\|A\|}{2}) = \min_{\theta\in V} h(\frac{1+\|A\theta\|}{2}) = \min_{\sigma\in\mathcal{S}(\mathbf{C}^2)} \tilde{H}(\Gamma\sigma).$$

On the other hand, the matrices $\Gamma\rho_{\pm\hat{\theta}} = \rho_{\pm A\hat{\theta}}$ mutually commute and the mixture $\frac{1}{2}(\rho_{A\hat{\theta}} + \rho_{-A\hat{\theta}}) = \rho_0 = \frac{1}{2}I$ achieves

$$\tilde{H}(\rho_0) = \log 2 = \max_p h(p) = \max_{\sigma\in\mathcal{S}(\mathbf{C}^2)} \tilde{H}(\Gamma\sigma).$$

Consequently the theorem follows from Corollary 20. □

In order to state the main result on the case $b \neq 0$, we need some preliminary considerations. Let $\{r_i \geq 0\,;\, i = 1,2,3\}$ be the singular values of $A$ (i.e. the square roots of the eigenvalues of $A\,{}^t\!A$) and $\{v_i\,;\, i = 1,2,3\}$ be the corresponding orthonormal eigenvectors of $A\,{}^t\!A$. Then the boundary of $A\mathcal{V}$ is represented as

$$\mathcal{E}(A) \stackrel{\text{def}}{=} \left\{A\theta\,;\, \theta\in\mathbf{R}^3,\, \|\theta\| = 1\right\} = \left\{\sum_{i=1}^3 x_i v_i\,;\, (x_1,x_2,x_3)\in\mathbf{R}^3,\, \sum_{i=1}^3 \frac{x_i^2}{r_i^2} = 1\right\}.$$

That is, $\mathcal{E}(A)$ forms an ellipsoid (including the collapsed case: $r_1r_2r_3 = 0$) with the principal axes $\{\mathcal{L}_i\,;\, i = 1,2,3\}$, where $\mathcal{L}_i$ is the straight line generated by $v_i$. Let us define

$$S(A,\mathcal{L}_i) \stackrel{\text{def}}{=} \max_{j(\neq i)} r_j^2 = \max\{\|x\|^2\,;\, x\in\mathcal{E}(A) \text{ and } x\perp\mathcal{L}_i\}.$$

This quantity measures the thickness of $\mathcal{E}(A)$ around $\mathcal{L}_i$.

For nonnegative numbers $\beta$ and $r$ such that $\beta + r \leq 1$, let[3]

$$T(\beta,r) \stackrel{\text{def}}{=} r^2 - \beta r + \frac{(\beta - r)\Delta h}{h'((1+\beta-r)/2)}, \tag{25}$$

where

$$\Delta h \stackrel{\text{def}}{=} h\left(\frac{1+\beta+r}{2}\right) - h\left(\frac{1+\beta-r}{2}\right)\ (\leq 0),$$

$$h'(p) = \frac{dh(p)}{dp} = \log\frac{1-p}{p}.$$

**Theorem 23** *Assume $b \neq 0$. The channel $\Gamma = (A,b)$ is pseudoclassical iff the following two conditions are satisfied.*

(i) *$b$ belongs to a principal axis $\mathcal{L}$ of $\mathcal{E}(A)$.*

(ii) *$S(A,\mathcal{L}) \leq T(\|b\|,r)$, where $r$ is the singular value of $A$ corresponding to $\mathcal{L}$; i.e. $A\,{}^t\!Ab = r^2 b$.*

[3] In the following, we always assume the convention of continuous prolongation to singular points. For example, $T(\beta,r)$ is continuously prolonged to $\beta = r$ as $T(r,r) = (\log 2 - h(\frac{1}{2}+r))/2$.

*In the pseudoclassical case, the capacity of* $\Gamma$ *coincides with that of the classical binary channel* $X \to Y$ *with the crossover probabilities:*

$$P\{Y=0|X=1\} = \frac{1-\|b\|-r}{2}, \qquad P\{Y=1|X=0\} = \frac{1+\|b\|-r}{2}. \tag{26}$$

Before proceeding to the proof, let us observe some implications of the theorem. For $\Gamma = (A, b)$ to be pseudoclassical[4] , the condition (i) in the theorem requires that the direction of the shift $b$ in the image $A\mathcal{V} + b$ of the channel must be parallel to one of the principal axes of $\mathcal{E}(A) = \partial(A\mathcal{V})$, while the condition (ii) requires that $\mathcal{E}(A)$ must be sufficiently thin around this direction, see Fig. 1. The following upper and lower bounds will be useful in estimating the threshold $T(\beta, r)$ in the condition (ii).

**Proposition 24**

$$r^2 \le T(\beta, r) \le T_+(\beta, r) \le r^2 + \beta r,$$

*where*

$$T_+(\beta, r) \stackrel{\text{def}}{=} r^2 + \beta r - \frac{(\beta + r)\Delta h}{h'((1+\beta+r)/2)}.$$

**Proof** See Appendix B. □

**Corollary 25** $T(\beta, r) = 0$ *iff* $r = 0$.

In the situation of Theorem 23, the condition $r = 0$ means that the ellipsoid $\mathcal{E}(A)$ is collapsed to an elliptic disc orthogonal to $\mathcal{L}$, and the above corollary claims that when $r = 0$ (and $b \neq 0$) the channel $\Gamma = (A, b)$ is always non-pseudoclassical except for the case where $\mathcal{E}(A)$ is collapsed to one point.

In order to obtain a significant consequence from the upper bounds in Proposition 24, we further need some elementary considerations on ellipses. This will also serve as preliminaries for the proof of Theorem 23. Let us consider the ellipse (possibly collapsed)

$$\frac{(x-\beta)^2}{r^2} + \frac{y^2}{s^2} = 1 \tag{27}$$

in the $(x, y)$ plane, where $r, s$ and $\beta$ are arbitrary nonnegative reals. The $x$ coordinate of a point on the ellipse lies in the interval $[\beta - r, \beta + r]$ and is parametrized by $\mu \in [0, 1]$ as $x = \beta - r + 2r\mu$. The corresponding $y$ coordinate is $\pm 2s\sqrt{\mu(1-\mu)}$, and hence the squared norm of the position vector $(x, y)$ is given by

$$f(\mu) = f(\mu; \beta, r, s) \stackrel{\text{def}}{=} (\beta - r + 2r\mu)^2 + 4s^2\mu(1-\mu).$$

The proof of the following lemma is easy and is omitted.

[4] Theorem 23 needs no modification when we impose additional condition that a channel shall be completely positive, although not every ellipsoid inside the unit ball can be realized as the image of a completely positive channel. The condition for a binary channel $\Gamma = (A, b)$ to be completely positive will be presented elsewhere.

**Lemma 26**
(i) *If* $s^2 \leq r^2$, $f(\mu)$ *is convex and* $\max_{0\leq\mu\leq 1} f(\mu) = f(1) = (\beta + r)^2$.
(ii) *If* $r^2 \leq s^2 \leq r^2 + \beta r$, $f(\mu)$ *is concave, monotone increasing and* $\max_{0\leq\mu\leq 1} f(\mu) = f(1) = (\beta + r)^2$.
(iii) *If* $r^2 + \beta r \leq s^2$, $f(\mu)$ *is concave and* $\max_{0\leq\mu\leq 1} f(\mu) = s^2(s^2 - r^2 + \beta^2)/(s^2 - r^2)$.

It is shown from the above lemma that the ellipse (27) is inside the unit circle iff it satisfies the two conditions $\beta + r \leq 1$ and

$$s^2 \leq S_{\max}(\beta, r) \stackrel{\text{def}}{=} \frac{1 + r^2 - \beta^2 + \sqrt{D}}{2}, \tag{28}$$

where

$$D \stackrel{\text{def}}{=} \left\{1 - (\beta + r)^2\right\}\left\{1 - (\beta - r)^2\right\} \ (\geq 0).$$

Therefore, under the condition (i) of Theorem 23, we see that $A\mathcal{V} + b \subset \mathcal{V}$ is satisfied iff $\|b\| + r \leq 1$ and $S(A, \mathcal{L}) \leq S_{\max}(\|b\|, r)$.

**Proposition 27** *Assume that* $\beta + r \leq 1$. *Then we have* $T(\beta, r) \leq S_{\max}(\beta, r)$, *where the equality holds iff* $(\beta, r) = (1, 0)$ *or* $(0, 1)$.

**Proof** The inequality follows from Proposition 24 and

$$S_{\max}(\beta, r) - (r^2 + \beta r) = \frac{1}{2}\left\{1 - (\beta + r)^2 + \sqrt{D}\right\} \geq 0. \tag{29}$$

Suppose that $T(\beta, r) = S_{\max}(\beta, r)$. Then we necessarily have the equations $r^2 + \beta r = S_{\max}(\beta, r)$ and $T(\beta, r) = T_+(\beta, r)$. Owing to (29), the first equation implies that $\beta + r = 1$, and substituting this into the second equation, we have

$$0 = T_+(\beta, 1 - \beta) - T(\beta, 1 - \beta) = \frac{(1 - \beta)\log(1 - \beta) - \beta \log \beta}{h'(\beta)},$$

which means that $\beta = 0$ or $\beta = 1$. We thus obtain $(\beta, r) = (1, 0)$ or $(0, 1)$, and the 'only if' part on the equality condition has been proved. The 'if' part can be verified directly. □

Consider the situation of Theorem 23 where $b \neq 0$ and $A\,{}^t\!Ab = r^2 b$. We have seen that under the constraint $A\mathcal{V} + b \subset \mathcal{V}$, $S(A, \mathcal{L})$ can take an arbitrary value in $0 \leq S(A, \mathcal{L}) \leq S_{\max}(\|b\|, r)$. On the other hand, the above proposition claims that the threshold $T(\|b\|, r)$ for the pseudoclassicality is strictly less than $S_{\max}(\|b\|, r)$ unless $\|b\| = 1$ and $r = 0$, or in other words, unless the image $A\mathcal{V} + b$ of the channel is collapsed to a point on the unit sphere. Therefore Theorem 23 implies that the quantum binary channel exhibits in general a transition between a pseudoclassical channel and a non-pseudoclassical one by varying the value of parameter $S(A, \mathcal{L})$. It should be noted that the condition $(\beta, r) = (0, 1)$ in Proposition 27 corresponds to no situation in Theorem 23 since $b \neq 0$ is assumed in the theorem. In fact, we know from Theorem 22 that the channel is always pseudoclassical when $b = 0$, being regardless of $r$.

Now, the rest of the present section is devoted to the proof of Theorem 23. In the sequel, a distribution $p = \sum_{j=1}^n \lambda_j \delta_{\sigma_j} \in \mathcal{P}$ is denoted as $p = (\lambda_1, \ldots, \lambda_n \,;\, \theta^{(1)}, \ldots, \theta^{(n)}) = \big(\lambda_j \,;\, \theta^{(j)}\big)_{j=1}^n$ when $\sigma_j = \rho_{\theta^{(j)}}$, and we use the notation

$$\tilde{I}(p\,;\,\Gamma) = \tilde{I}(\lambda_1, \ldots, \lambda_n \,;\, \theta^{(1)}, \ldots, \theta^{(n)} \,;\, \Gamma) = \tilde{I}(\lambda_1, \ldots, \lambda_n \,;\, \xi^{(1)}, \ldots, \xi^{(n)}),$$

where $\xi^{(j)} \stackrel{\rm def}{=} A\theta^{(j)} + b$. Letting $\bar{\xi} \stackrel{\rm def}{=} \sum_{j=1}^n \lambda_j \xi^{(j)}$, we have

$$\tilde{I}(\lambda_1, \ldots, \lambda_n \,;\, \xi^{(1)}, \ldots, \xi^{(n)}) = h\left(\frac{1 + \|\bar{\xi}\|}{2}\right) - \sum_{j=1}^n \lambda_j h\left(\frac{1 + \|\xi^{(j)}\|}{2}\right). \tag{30}$$

To begin with, we show that the condition (i) of Theorem 23 is necessary for the channel to be pseudoclassical. Suppose that $\Gamma = (A, b)$ is pseudoclassical and that a $\Gamma$-commutative distribution $\hat{p} = (\hat{\lambda}_j \,;\, \hat{\theta}^{(j)})_{j=1}^n$ achieves $C(\Gamma) = \tilde{I}(\hat{p}\,;\,\Gamma)$. The $\Gamma$-commutativity implies that all the vectors $\{\hat{\xi}^{(j)} = A\hat{\theta}^{(j)} + b \,;\, j = 1, \ldots, n\}$ lie on a 1-dimensional linear subspace, say $\mathcal{L}$, of $\mathbf{R}^3$, and according to Proposition 9 we can assume that $n = 2$ and $\|\hat{\theta}^{(1)}\| = \|\hat{\theta}^{(2)}\| = 1$ with no loss of generality. Except for the trivial case $A = 0$, it holds that $\hat{\lambda}_j > 0 \;\; (j = 1, 2)$ and $\hat{\xi}^{(1)} \neq \hat{\xi}^{(2)}$.

**Lemma 28** *For each $j = 1, 2$,*

$$\hat{\theta}^{(j)} \in {}^t\!A\mathcal{L} = \{{}^t\!A\xi \,;\, \xi \in \mathcal{L}\}. \tag{31}$$

**Proof** We first verify the claim by assuming that $\|\hat{\xi}^{(j)}\| < 1$. In this case, $\tilde{I} = \tilde{I}(\lambda_1, \lambda_2 \,;\, \xi^{(1)}, \xi^{(2)})$ is differentiable at $\hat{p}$ w.r.t. the variable $\xi^{(j)}$ to yield the derivative

$$\begin{aligned}
\left.\frac{\partial \tilde{I}}{\partial \xi^{(j)}}\right|_{p=\hat{p}} &= {}^t\!\left(\frac{\partial \tilde{I}}{\partial \xi_1^{(j)}}, \frac{\partial \tilde{I}}{\partial \xi_2^{(j)}}, \frac{\partial \tilde{I}}{\partial \xi_3^{(j)}}\right)\Bigg|_{p=\hat{p}} \\
&= \hat{\lambda}_j \left\{ \left.\frac{\partial}{\partial \xi} h\left(\frac{1 + \|\xi\|}{2}\right)\right|_{\xi=\bar{\xi}} - \left.\frac{\partial}{\partial \xi} h\left(\frac{1 + \|\xi\|}{2}\right)\right|_{\xi=\hat{\xi}^{(j)}} \right\},
\end{aligned}$$

where $\bar{\xi} \stackrel{\rm def}{=} \sum_j \hat{\lambda}_j \hat{\xi}^{(j)}$. Let $v$ be a unit vector in $\mathcal{L}$. Then we have

$$\begin{aligned}
\frac{\partial}{\partial \xi} h\left(\frac{1 + \|\xi\|}{2}\right) &= \begin{cases} h'\left(\dfrac{1 + \|\xi\|}{2}\right) \dfrac{\xi}{2\|\xi\|} & \text{if} \quad \xi \neq 0 \\ 0 & \text{if} \quad \xi = 0 \end{cases} \\
&= \frac{1}{2} h'\left(\frac{1 + v \cdot \xi}{2}\right) v,
\end{aligned}$$

where the derivative is supposed to be evaluated at a point $\xi$ in $\mathcal{L}$ and $\cdot$ denotes the standard inner product on $\mathbf{R}^3$. Hence it follows that

$$\left.\frac{\partial \tilde{I}}{\partial \xi^{(j)}}\right|_{p=\hat{p}} = \hat{\lambda}_j \left\{ h'\left(\frac{1 + v \cdot \bar{\xi}}{2}\right) - h'\left(\frac{1 + v \cdot \hat{\xi}^{(j)}}{2}\right) \right\} v.$$

Invoking that $h'$ is strictly monotone decreasing and $\bar{\xi} \neq \hat{\xi}^{(j)}$, we can observe from the above equation that $\partial\tilde{I}/\partial\xi^{(j)}|_{p=\hat{p}}$ is a nonzero element of $\mathcal{L}$. Moreover, since $v$ and $\hat{\xi}^{(1)} - \hat{\xi}^{(2)}$ are both nonzero elements of $\mathcal{L}$ and therefore

$$0 \neq v \cdot (\hat{\xi}^{(1)} - \hat{\xi}^{(2)}) = ({}^tAv) \cdot (\hat{\theta}^{(1)} - \hat{\theta}^{(2)}),$$

we can see that ${}^tAv \neq 0$. Thus the derivative

$$\left.\frac{\partial\tilde{I}}{\partial\theta^{(j)}}\right|_{p=\hat{p}} = {}^tA \left.\frac{\partial\tilde{I}}{\partial\xi^{(j)}}\right|_{p=\hat{p}} \qquad (\in {}^tA\mathcal{L})$$

is also shown to be nonzero. On the other hand, recalling that $\tilde{I}$ takes the maximum at $\hat{p}$ under the constraint $\|\theta^{(j)}\|^2 = 1$, we have

$$\exists c \in \mathbf{R}\,; \quad \left.\frac{\partial\tilde{I}}{\partial\theta^{(j)}}\right|_{p=\hat{p}} = c \left.\frac{\partial\|\theta\|^2}{\partial\theta}\right|_{\theta=\hat{\theta}^{(j)}} = 2c\hat{\theta}^{(j)},$$

where $c$ is the corresponding Lagrange indeterminate coefficient. Consequently, $2c\hat{\theta}^{(j)}$ is a nonzero element of ${}^tA\mathcal{L}$ and so is $\hat{\theta}^{(j)}$. Thus the claim has been verified for the case $\|\hat{\xi}^{(j)}\| < 1$.

When $\|\hat{\xi}^{(j)}\| = 1$, the squared norm $\|A\theta + b\|^2$ takes the maximum 1 at $\theta = \hat{\theta}^{(j)}$ under the constraint $\|\theta\|^2 = 1$, and we have

$$\exists c \in \mathbf{R}\,; \quad \left.\frac{\partial}{\partial\theta}\|A\theta + b\|^2\right|_{\theta=\hat{\theta}^{(j)}} = 2{}^tA\hat{\xi}^{(j)} = 2c\hat{\theta}^{(j)},$$

which verifies the claim. □

The nonzero vector $\hat{\xi}^{(1)} - \hat{\xi}^{(2)} = A(\hat{\theta}^{(1)} - \hat{\theta}^{(2)})$ belongs to $A\,{}^tA\mathcal{L}$ owing to Lemma 28 and belongs to $\mathcal{L}$ owing to the assumption that $\hat{\xi}^{(j)} \in \mathcal{L}$. Hence we have $A\,{}^tA\mathcal{L} = \mathcal{L}$, which means that $\mathcal{L}$ is a principal axis of $\mathcal{E}(A)$. Moreover, since $\hat{\xi}^{(j)} \in \mathcal{L}$ and $A\hat{\theta}^{(j)} \in A\,{}^tA\mathcal{L} = \mathcal{L}$, we obtain $b = \hat{\xi}^{(j)} - A\hat{\theta}^{(j)} \in \mathcal{L}$. It is thus concluded the pseudoclassicality of $\Gamma$ implies the condition (i).

From now on, we assume (i). Then the line segment $(A\mathcal{V} + b) \cap \mathcal{L}$ is of length $2r$ with the midpoint $b$, where $r$ is the singular value of $A$ corresponding to $\mathcal{L}$. We denote by $\eta_+$ and $\eta_-$ the endpoints of the segment, whose norms are $\|\eta_\pm\| = |\|b\| \pm r|$. For $0 \leq \forall\lambda \leq 1$, we have

$$\|\lambda\eta_+ + (1-\lambda)\eta_-\| = |\lambda(\|b\| + r) + (1-\lambda)(\|b\| - r)|\,.$$

Note that $(\mathcal{E}(A) + b) \cap \mathcal{L} = \{\eta_+, \eta_-\}$ when $A$ is nonsingular.

Tracing the preceding argument on the necessity of (i), we can see that if $\Gamma = (A, b)$ is pseudoclassical there exists such an optimal distribution $\hat{p} = (\hat{\lambda}, 1-\hat{\lambda}\,;\, \hat{\theta}_+, \hat{\theta}_-)$ that satisfies $A\hat{\theta}_\pm + b = \eta_\pm$. The corresponding channel capacity is then given by

$$\begin{aligned} C(\Gamma) &= \tilde{I}(\hat{\lambda}, 1-\hat{\lambda}\,;\, \eta_+, \eta_-) \\ &= h\left(\frac{1 + \|\hat{\lambda}\eta_+ + (1-\hat{\lambda})\eta_-\|}{2}\right) - \hat{\lambda}h\left(\frac{1+\|\eta_+\|}{2}\right) - (1-\hat{\lambda})h\left(\frac{1+\|\eta_-\|}{2}\right) \end{aligned}$$

$$= h(\hat{\lambda}\delta + (1-\hat{\lambda})\epsilon) - \hat{\lambda}h(\delta) - (1-\hat{\lambda})h(\epsilon)$$
$$= \max_{0\le\lambda\le 1} \{h(\lambda\delta + (1-\lambda)\epsilon) - \lambda h(\delta) - (1-\lambda)h(\epsilon)\},$$

where $\delta \stackrel{\rm def}{=} (1+\|b\|+r)/2$ and $\epsilon \stackrel{\rm def}{=} (1+\|b\|-r)/2$. This is equal to the capacity of the classical binary channel with the crossover probabilities $\{1-\delta, \epsilon\}$ which are identical with (26).

We next proceed to the proof that on the assumption (i) the pseudoclassicality is equivalent to the condition (ii). The following lemma will play an essential role in the proof.

**Lemma 29** *Given arbitrary nonnegative numbers $\beta, r, s$ satisfying $\beta+r \le 1$ and $s^2 \le S_{\max}(\beta, r)$, the following two conditions are mutually equivalent.*

(i) $s^2 \le T(\beta, r)$.

(ii) $$h\left(\frac{1+\sqrt{f(\mu\,;\,\beta,r,s)}}{2}\right) \ge \mu h\left(\frac{1+\beta+r}{2}\right) + (1-\mu)h\left(\frac{1+\beta-r}{2}\right) \quad \text{for} \quad 0 \le \forall\mu \le 1.$$

**Proof** See Appendix C. □

Let $\xi$ be an arbitrary point on the ellipsoid $\mathcal{E}(A)+b$ and $\pi(\xi)$ be the orthogonal projection of $\xi$ onto $\mathcal{L}$. Since $\pi(\xi)$ lies in the line segment $(A\mathcal{V}+b)\cap\mathcal{L}$, it is represented as $\pi(\xi) = \mu(\xi)\eta_+ + (1-\mu(\xi))\eta_-$ by a constant $\mu(\xi) \in [0,1]$.

**Lemma 30** *The condition* (ii) *in Theorem 23 is equivalent to the following:*

(ii)' $$h\left(\frac{1+\|\xi\|}{2}\right) \ge \mu(\xi)h\left(\frac{1+\|\eta_+\|}{2}\right) + (1-\mu(\xi))h\left(\frac{1+\|\eta_-\|}{2}\right) \quad \text{for} \quad \forall\xi \in \mathcal{E}(A)+b.$$

**Proof** Given an arbitrary point $\xi \in \mathcal{E}(A)+b$, let $\mathcal{K}$ be the plane spanned by $\mathcal{L}$ and $\xi$. Then the intersection $(\mathcal{E}(A)+b)\cap\mathcal{K}$ forms an ellipse of the form

$$(\mathcal{E}(A)+b)\cap\mathcal{K} = \left\{xv + yw\,;\ (x,y)\in\mathbf{R}^2,\ \ \frac{(x-\|b\|)^2}{r^2} + \frac{y^2}{s(\xi)^2} = 1\right\},$$

where $v \stackrel{\rm def}{=} \eta_+/\|\eta_+\|$, $w$ is a unit vector in $\mathcal{K}$ orthogonal to $v$, and $s(\xi)$ is a nonnegative constant. Noting that

$$\|\xi\|^2 = f(\mu(\xi)\,;\,\|b\|, r, s(\xi))$$

and

$$\max\{s(\xi)^2\,;\,\xi\in\mathcal{E}(A)+b\} = S(A,\mathcal{L}),$$

we can see that the lemma follows from Lemma 29. □

Let us prove that the pseudoclassicality of $\Gamma = (A,b)$ is equivalent to (ii)' above. We first assume (ii)' and suppose that a distribution $(\lambda_1,\ldots,\lambda_n\,;\,\xi^{(1)},\ldots,\xi^{(n)})$ on $\mathcal{E}(A)+b$ is arbitrarily given. Letting $\bar{\xi} \stackrel{\rm def}{=} \sum_j \lambda_j\xi^{(j)}$ and $\bar{\mu} \stackrel{\rm def}{=} \sum_j \lambda_j\mu(\xi^{(j)})$, we have

$$\tilde{I}(\lambda_1,\ldots,\lambda_n\,;\,\xi^{(1)},\ldots,\xi^{(n)})$$

$$
\begin{aligned}
&= h\left(\frac{1+\|\bar{\xi}\|}{2}\right) - \sum_{j=1}^{n} \lambda_j h\left(\frac{1+\|\xi^{(j)}\|}{2}\right) \\
&\le h\left(\frac{1+\|\bar{\xi}\|}{2}\right) - \bar{\mu} h\left(\frac{1+\|\eta_+\|}{2}\right) - (1-\bar{\mu}) h\left(\frac{1+\|\eta_-\|}{2}\right) \\
&\le \tilde{I}(\bar{\mu}, 1-\bar{\mu}\,;\,\eta_+, \eta_-),
\end{aligned}
$$

where the first inequality follows from (ii)' and the second inequality follows from

$$
\|\bar{\xi}\| \ge \|\pi(\bar{\xi})\| = \|\sum_j \lambda_j \pi(\xi^{(j)})\| = \|\bar{\mu}\eta_+ + (1-\bar{\mu})\eta_-\|.
$$

Thus the maximum of $\tilde{I}$ is attained by a $\Gamma$-commutative distribution and hence $\Gamma$ is pseudoclassical.

Conversely, assume that $\Gamma$ is pseudoclassical. In this case the capacity is given in the form

$$
C(\Gamma) = \tilde{I}(\lambda, 1-\lambda\,;\,\eta_+, \eta_-)
$$

by some constant $0 < \lambda < 1$ as mentioned above. Given an arbitrary point $\xi \in \mathcal{E}(A) + b$, we can always find a triplet $(\nu_1, \nu_2, \nu_3)$ of $\nu_1 \ge 0, \nu_2 \ge 0$ and $\nu_3 > 0$ such that $\nu_1 + \nu_2 + \nu_3 = 1$ and $\lambda = \nu_1 + \nu_3\mu(\xi)$. Consider the point $\xi' \stackrel{\text{def}}{=} 2\pi(\xi) - \xi$, which is in the symmetrical position to $\xi$ w.r.t. the axis $\mathcal{L}$ and satisfies $\xi' \in \mathcal{E}(A) + b$ and $\|\xi'\| = \|\xi\|$. Then we have

$$
\begin{aligned}
\nu_1\eta_+ + \nu_2\eta_- + \frac{\nu_3}{2}\xi + \frac{\nu_3}{2}\xi' &= \nu_1\eta_+ + \nu_2\eta_- + \nu_3\left\{\mu(\xi)\eta_+ + (1-\mu(\xi))\eta_-\right\} \\
&= \lambda\eta_+ + (1-\lambda)\eta_-,
\end{aligned}
$$

and therefore

$$
\begin{aligned}
&\tilde{I}(\nu_1, \nu_2, \frac{\nu_3}{2}, \frac{\nu_3}{2}\,;\,\eta_+, \eta_-, \xi, \xi') \\
&\quad = h\left(\frac{1+\|\lambda\eta_+ + (1-\lambda)\eta_-\|}{2}\right) - \nu_1 h\left(\frac{1+\|\eta_+\|}{2}\right) - \nu_2 h\left(\frac{1+\|\eta_-\|}{2}\right) - \nu_3 h\left(\frac{1+\|\xi\|}{2}\right) \\
&\quad = \tilde{I}(\lambda, 1-\lambda\,;\,\eta_+, \eta_-) - \nu_3\left\{h\left(\frac{1+\|\xi\|}{2}\right) - \mu(\xi) h\left(\frac{1+\|\eta_+\|}{2}\right) - (1-\mu(\xi)) h\left(\frac{1+\|\eta_-\|}{2}\right)\right\}.
\end{aligned}
$$

Since $\tilde{I}(\nu_1, \nu_2, \nu_3/2, \nu_3/2\,;\,\eta_+, \eta_-, \xi, \xi')$ cannot exceed $C(\Gamma) = \tilde{I}(\lambda, 1-\lambda\,;\,\eta_+, \eta_-)$, the condition (ii)' must be satisfied. This completes the proof of Theorem 23.

# 7 Concluding remarks

We explored some basic characteristics of quantum channel capacity $C(\Gamma)$. In this course, the importance and the difficulty of asymptotics in quantum statistics was clarified. There are of course many open questions left. For example, development of efficient algorithms for computing $C(\Gamma)$ and/or finding practical error correcting codes within the framework of this paper are important in a practical viewpoint.

But among others, the big open question is: what will happen when we further assume that $\Gamma$ is the dual map of a completely positive map? In this case, we can consider the extended channel of the form $\Gamma^{(n)} : \mathcal{S}(\mathcal{H}_1^{\otimes n}) \longmapsto \mathcal{S}(\mathcal{H}_2^{\otimes n})$. In other words, we can adopt states which belong to $\mathcal{S}(\mathcal{H}_1^{\otimes n}) \setminus \mathcal{S}_1^{(n)}$ as codewords. The analysis of the corresponding capacity is left untouched.

# Appendices

## A Proof of Proposition 1

The second equality in (6) follows from the superadditivity of $C^{(n)}(\Gamma)$, and $C(\Gamma) \geq \lim_n C^{(n)}(\Gamma)/n$ is an immediate consequence of (5). We show the converse inequality.

By using Fano's inequality and assuming the uniform distribution over the codebook $\mathcal{C}_n$, the average error probability $P_e(\mathcal{C}_n, T^{(n)})$ is evaluated as

$$\begin{aligned}
\log 2 + P_e(\mathcal{C}_n, T^{(n)}) \log |\mathcal{C}_n| &\geq H(\hat{\sigma}^{(n)}|\hat{\tau}^{(n)}) \\
&= H(\hat{\sigma}^{(n)}) - I(\hat{\sigma}^{(n)}; \hat{\tau}^{(n)}) \\
&\geq \log |\mathcal{C}_n| - \sup_{p^{(n)}, T^{(n)}} I^{(n)}(p^{(n)}, T^{(n)}; \Gamma) \\
&\geq \log |\mathcal{C}_n| - \sup_{p^{(n)}, M^{(n)}} I^{(n)}(p^{(n)}, M^{(n)}; \Gamma),
\end{aligned}$$

where $H(\,\cdot\,|\,\cdot\,)$ and $H(\,\cdot\,)$ denote the classical conditional entropy and the Shannon entropy, respectively, $\hat{\sigma}^{(n)}$ denotes the $\mathcal{C}_n$-valued random variable which is uniformly distributed over $\mathcal{C}_n$, and $\hat{\tau}^{(n)}$ denotes the $\mathcal{C}_n$-valued random variable which corresponds to the decoded words when the decoder $T^{(n)}$ is applied to the output state $\Gamma\hat{\sigma}^{(n)}$. This inequality leads to

$$\left(1 - P_e(\mathcal{C}_n, T^{(n)})\right) \frac{\log |\mathcal{C}_n|}{n} \leq \frac{C^{(n)}(\Gamma)}{n} + \frac{\log 2}{n}.$$

Then in order to assure $P_e(\mathcal{C}_n, T^{(n)}) \to 0$ as $n \to \infty$, $\limsup_{n\to\infty} \log |\mathcal{C}_n|/n$ must be less than or equal to $\lim_n C^{(n)}(\Gamma)/n$. □

## B Proof of Proposition 24

For $0 \leq x < 1$, let

$$K(x) \stackrel{\text{def}}{=} h\left(\frac{1+\sqrt{x}}{2}\right) \quad (\geq 0), \tag{32}$$

$$L(x) \stackrel{\text{def}}{=} \log\left(\frac{1-\sqrt{x}}{1+\sqrt{x}}\right) \quad (\leq 0). \tag{33}$$

Direct calculations yield

$$K'(x) = \frac{L(x)}{4\sqrt{x}} \tag{34}$$

$$L'(x) = -\frac{1}{\sqrt{x}(1-x)} \tag{35}$$

$$K''(x) = -\frac{1}{8x\sqrt{x}}\left\{L(x) + \frac{2\sqrt{x}}{1-x}\right\} \tag{36}$$

$$K'(x) + 2xK''(x) = -\frac{1}{2(1-x)}. \tag{37}$$

**Lemma 31** *For* $0 \le \forall x < 1$,

$$-\frac{2\sqrt{x}}{1-x} \le L(x) \le -\frac{6\sqrt{x}}{3-x}.$$

**Proof** Let $a \ge 1$ and define

$$P(x) \stackrel{\text{def}}{=} L(x) + \frac{2a\sqrt{x}}{a-x}.$$

Then $P(0) = 0$ and for $0 < \forall x < 1$

$$\begin{aligned} P'(x) &= \frac{-\sqrt{x}}{(1-x)(a-x)^2}\{(a+1)x + a(a-3)\} \\ &= \begin{cases} \dfrac{2\sqrt{x}}{(1-x)^2} > 0 & \text{if} \quad a = 1 \\ \dfrac{-4x\sqrt{x}}{(1-x)(3-x)^2} < 0 & \text{if} \quad a = 3, \end{cases} \end{aligned}$$

which proves the lemma. □

**Lemma 32** *K(x) is monotone decreasing and concave.*

**Proof** Immediate from (34) (36) and Lemma 31. □

**Proof of Proposition 24** Since the last inequality in the proposition is obvious, we have only to show

$$r^2 \le T(\beta, r) \le T_+(\beta, r). \tag{38}$$

Let

$$x_+ \stackrel{\text{def}}{=} (\beta + r)^2 \ \ge \ x_- \stackrel{\text{def}}{=} (\beta - r)^2,$$

$$\Delta x \stackrel{\text{def}}{=} x_+ - x_- = 4r\beta, \quad \Delta K \stackrel{\text{def}}{=} K(x_+) - K(x_-) = \Delta h.$$

Then we have

$$\begin{aligned} K'(x_-) &= h'(\frac{1+|\beta - r|}{2})/4|\beta - r| = h'(\frac{1+\beta - r}{2})/4(\beta - r) \\ K'(x_+) &= h'(\frac{1+\beta + r}{2})/4(\beta + r) \end{aligned}$$

and

$$T(\beta, r) - r^2 = \frac{\Delta K - K'(x_-)\Delta x}{4K'(x_-)} \tag{39}$$

$$T_+(\beta, r) - r^2 = -\frac{\Delta K - K'(x_+)\Delta x}{4K'(x_+)}. \tag{40}$$

Noting that $\Delta K/\Delta x \leq K'(x_-) \leq 0$ holds according to Lemma 32, the first inequality in (38) is derived from (39). To prove the second inequality, we combine (39) and (40) to obtain

$$T_+(\beta, r) - T(\beta, r) = \frac{M(x_+, x_-)}{4K'(x_+)K'(x_-)}, \tag{41}$$

where

$$M(x, y) \stackrel{\text{def}}{=} 2(x - y)K'(x)K'(y) - \{K(x) - K(y)\}\{K'(x) + K'(y)\}.$$

Now it suffices to show that

$$M(x, y) \geq 0 \quad \text{for} \quad 0 < \forall y \leq \forall x < 1. \tag{42}$$

To this end, let us differentiate $M(x, y)$ into

$$\frac{\partial}{\partial x} M(x, y) = K'(x)K'(y) + K''(x)K(y) + 2(x - y)K''(x)K'(y) - K'(x)^2 - K(x)K''(x) \tag{43}$$

and

$$\begin{aligned} \frac{\partial^2}{\partial y \partial x} M(x, y) &= \{K'(x) + 2xK''(x)\}K''(y) - \{K'(y) + 2yK''(y)\}K''(x) \\ &= \frac{1}{2(1 - x)(1 - y)}\{(1 - x)K''(x) - (1 - y)K''(y)\}, \end{aligned} \tag{44}$$

where the last equality follows from (37). Let

$$N(x) \stackrel{\text{def}}{=} (1 - x)K''(x) = -\frac{(1 - x)L(x) + 2\sqrt{x}}{8x^{3/2}}. \tag{45}$$

Then we have

$$N'(x) = \frac{(3 - x)L(x) + 6\sqrt{x}}{16x^{5/2}},$$

which turns out negative from Lemma 31. Hence $N(x)$ is monotone decreasing, and it follows from (44) that

$$\frac{\partial^2}{\partial y \partial x} M(x, y) \begin{cases} \geq 0 & \text{if} \quad x \leq y \\ \leq 0 & \text{if} \quad x \geq y. \end{cases}$$

This fact, combined with $M(x, x) = \frac{\partial}{\partial x} M(x, x) = 0$, leads to $\frac{\partial}{\partial x} M(x, y) \geq 0$ and (42). □

## C Proof of Lemma 29

Let

$$g(\mu) = g(\mu\,;\,\beta,r,s) \stackrel{\text{def}}{=} h\left(\frac{1+\sqrt{f(\mu\,;\,\beta,r,s)}}{2}\right) - \mu h\left(\frac{1+\beta+r}{2}\right) - (1-\mu)h\left(\frac{1+\beta-r}{2}\right)$$
$$= K(f(\mu)) - \mu\Delta h - h\left(\frac{1+\beta-r}{2}\right). \tag{46}$$

It is easy to show that

$$g'(0) = 4K'((\beta-r)^2)\{s^2 - T(\beta,r)\} \tag{47}$$

and

$$g'(1) = 4K'((\beta+r)^2)\{T_+(\beta,r) - s^2\}. \tag{48}$$

Assume (ii); i.e.,

$$g(\mu) \geq 0 \quad \text{for} \quad 0 \leq \forall\mu \leq 1. \tag{49}$$

Then it follows from $g(0) = 0$ that $g'(0) \geq 0$, which is equivalent to (i) due to (47) since $K'(x) < 0$ for $0 \leq \forall x < 1$. Thus the implication (ii) $\Rightarrow$ (i) has been verified.

Assume (i) in turn. Then we immediately have $s^2 \leq T_+(\beta,r)$ according to Proposition 24. Hence it follows from (47) and (48) that

$$g'(0) \geq 0 \quad \text{and} \quad g'(1) \leq 0. \tag{50}$$

Let us show that the second derivative $g''(\mu)$ satisfies either

$$g''(\mu) \leq 0 \quad \text{for} \quad 0 \leq \forall\mu \leq 1 \quad (\text{i.e.} \quad g \text{ is concave}) \tag{51}$$

or

$$\exists\mu_0 \in [0,1]\,; \quad g''(\mu) \begin{cases} \geq 0 & \text{for} \quad 0 \leq \forall\mu \leq \mu_0 \\ \leq 0 & \text{for} \quad \mu_0 \leq \forall\mu \leq 1. \end{cases} \tag{52}$$

Obviously, both (51) and (52) imply (49) under the condition (50) and $g(0) = g(1) = 0$. Assume that $s^2 \leq r^2$ first. Then $f(\mu)$ is convex according to Lemma 26. Recalling Lemma 32, $K(f(\mu))$ is shown to be concave, and so is $g(\mu)$ from (46). Next we treat the case where $s^2 \geq r^2$. Using the equations (34) (36) and

$$\{f'(\mu)\}^2 = 16\left\{(r^2-s^2)f(\mu) + s^2(s^2-r^2+\beta^2)\right\},$$
$$f''(\mu) = 8(r^2-s^2),$$

we obtain

$$g''(\mu) = K''(f(\mu))\{f'(\mu)\}^2 + K'(f(\mu))f''(\mu)$$
$$= \frac{2}{\{f(\mu)\}^{3/2}}Q(f(\mu)),$$

where

$$Q(x) \stackrel{\text{def}}{=} 2(s^2 - r^2)\frac{x\sqrt{x}}{1-x} - s^2(s^2 - r^2 + \beta^2)\left\{L(x) + \frac{2\sqrt{x}}{1-x}\right\}.$$

The derivative of $Q(x)$ is written as

$$Q'(x) = \frac{\sqrt{x}}{(1-x)^2}\{(r^2 - s^2)(x-3) - 2s^2(s^2 - r^2 + \beta^2)\}.$$

Invoking the assumption $s^2 \geq r^2$, we can see from this equation that $Q'(x)$ for $0 \leq x < 1$ exhibits one of the following: (a) always $\geq 0$, (b) always $\leq 0$, (c) $\geq 0$ when $x \leq x_0$ and $\leq 0$ when $x \geq x_0$ for some constant $x_0$. Since $Q(0) = 0$, similar trichotomy is also valid for $Q(x)$. On the other hand, it follows from the assumption (i) and Proposition 24 that $s^2 \leq r^2 + \beta r$, which implies that $f(\mu)$ is monotone increasing according to Lemma 26. Therefore $g''(\mu)$ satisfies either (51) or (52) or $g''(\mu) \geq 0$ for $0 \leq \forall\mu \leq 1$. Under the condition (50), the last case is reduced to $g''(\mu) = 0$ for $0 \leq \forall\mu \leq 1$, which is a special case of (51) (52) . □

# References


[1] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.*, vol. 27, pp. 379-423, pp. 623-656, 1948.

[2] A. S. Holevo, *Probabilistic and Statistical Aspects of Quantum Theory.* Amsterdam: North-Holland, 1982.

[3] W. F. Stinespring, "Positive functions on $C^*$-algebras," *Proc. Am. Math. Soc.*, vol. 6, pp. 211-216, 1955.

[4] M. Ohya, "On compound state and mutual information in quantum information theory," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 770-774, 1983.

[5] A. S. Holevo, "Capacity of a quantum communication channel," *Problemy Peredachi Informatsii*, vol. 15, pp. 3-11, 1979 [*Problems of Information Transmission*, vol. 15, pp. 247-253, 1979].

[6] A. S. Holevo, "Bounds for the quantity of information transmitted by a quantum communication channel," *Problemy Peredachi Informatsii*, vol. 9, pp. 3-11, 1973 [*Problems of Information Transmission*, vol. 9, pp. 177-183, 1973].

[7] G. Lindblad, "Completely positive maps and entropy inequalities," *Commun. Math. Phys.*, vol. 40, pp. 147-151, 1975.

[8] M. Ohya and D. Petz, *Quantum Entropy and its Use.* Berlin: Springer, 1993.

[9] P. Hausladen, R. Jozsa, B. Schumacher, M. Westmoreland, and W. K. Wootters, "Classical information capacity of a quantum channel," *Phys. Rev. A*, vol. 54, pp. 1869-1876, 1996.

[10] A. S. Holevo, "The capacity of quantum channel for general signal states," LANL archive quant-ph/9611023; to appear in *IEEE Trans. Inform. Theory.*

[11] B. Schumacher and M. D. Westmoreland, "Sending classical information via noisy quantum channels," *Phys. Rev. A*, vol. 56, pp. 131-138, 1997.

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

[13] F. Hiai and D. Petz, "The proper formula for relative entropy and its asymptotics in quantum probability," *Commun. Math. Phys.*, vol. 143, pp. 99-114, 1991.

[14] E. M. Alfsen, *Compact convex sets and boundary integrals.* Berlin: Springer, 1971.

[15] B. Grünbaum, *Convex Polytopes.* London: Wiley, 1967.

[16] E. B. Davies, "Information and quantum measurement," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 596-599, 1978.

[17] N. Dunford and J. T. Schwartz, *Linear operators, Part I: General theory.* New York: Wiley, 1958.

[18] D. Petz, "Properties of quantum entropy," in *Quantum Probability and Applications II*, edited by L. Accardi and W. von Waldenfels, Lec. Notes in Math., vol. 1136, Berlin: Springer, 1985, pp. 428-441.